Chapter 2
Multiple Linear Regression

This chapter introduces the multiple linear regression model, the response plot for checking goodness of fit, the residual plot for checking lack of fit, the ANOVA $F$ test, the partial $F$ test, the $t$ tests, and least squares. The problems use software $R$, $SAS$, $Minitab$, and $Arc$.

2.1 The MLR Model

Definition 2.1. The response variable is the variable that you want to predict. The predictor variables are the variables used to predict the response variable.

Notation. In this text the response variable will usually be denoted by $Y$ and the $p$ predictor variables will often be denoted by $x_1, \ldots, x_p$. The response variable is also called the dependent variable while the predictor variables are also called independent variables, explanatory variables, carriers, or covariates. Often the predictor variables will be collected in a vector $\mathbf{x}$. Then $\mathbf{x}^T$ is the transpose of $\mathbf{x}$.

Definition 2.2. Regression is the study of the conditional distribution $Y|\mathbf{x}$ of the response variable $Y$ given the vector of predictors $\mathbf{x} = (x_1, \ldots, x_p)^T$.

Definition 2.3. A quantitative variable takes on numerical values while a qualitative variable takes on categorical values.

Example 2.1. Archeologists and crime scene investigators sometimes want to predict the height of a person from partial skeletal remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (e.g., ancient Egyptians or modern US citizens). The response variable $Y$ is height and the predictor variables might be $x_1 \equiv 1$, $x_2 = femur$ length, and $x_3 = ulna$ length.
The heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example $Y$, $x_2$, and $x_3$ are quantitative variables. If $x_4 = \text{gender}$ is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

**Definition 2.4.** Suppose that the response variable $Y$ and at least one predictor variable $x_i$ are quantitative. Then the **multiple linear regression** (MLR) model is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = x_i^T\beta + e_i$$

for $i = 1, \ldots, n$. Here $n$ is the *sample size* and the random variable $e_i$ is the *ith error*. Suppressing the subscript $i$, the model is $Y = x^T\beta + e$.

In matrix notation, these $n$ equations become

$$Y = X\beta + e,$$

where $Y$ is an $n \times 1$ vector of dependent variables, $X$ is an $n \times p$ matrix of predictors, $\beta$ is a $p \times 1$ vector of unknown coefficients, and $e$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$ (2.3)

Often the first column of $X$ is $X_1 = 1$, the $n \times 1$ vector of ones. The *ith case* $(x_i^T, Y_i) = (x_{i,1}, x_{i,2}, \ldots, x_{i,p}, Y_i)$ corresponds to the *ith row* $x_i^T$ of $X$ and the *ith element* of $Y$ (if $x_{i,1} \equiv 1$, then $x_{i,1}$ could be omitted). In the MLR model $Y = x^T\beta + e$, the $Y$ and $e$ are random variables, but we only have observed values $Y_i$ and $x_i$. If the $e_i$ are iid (independent and identically distributed) with zero mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \text{VAR}(e_i) = \sigma^2$, then regression is used to estimate the unknown parameters $\beta$ and $\sigma^2$.

**Definition 2.5.** The **constant variance MLR model** uses the assumption that the errors $e_1, \ldots, e_n$ are iid with mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $x_i$. The predictor variables $x_i$ are assumed to be fixed and measured without error. The cases $(x_i^T, Y_i)$ are independent for $i = 1, \ldots, n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $x_i$. That is, observe the $x_i$ and then act as if the observed $x_i$ are fixed.
Definition 2.6. The unimodal MLR model has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors \( e_1, \ldots, e_n \) are iid from a unimodal distribution that is not highly skewed. Note that \( E(e_i) = 0 \) and \( V(e_i) = \sigma^2 < \infty \).

Definition 2.7. The normal MLR model or Gaussian MLR model has the same assumptions as the unimodal MLR model but adds the assumption that the errors \( e_1, \ldots, e_n \) are iid \( N(0, \sigma^2) \) random variables. That is, the \( e_i \) are iid normal random variables with zero mean and variance \( \sigma^2 \).

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares (OLS).

Notation. The symbol \( A \equiv B = f(c) \) means that \( A \) and \( B \) are equivalent and equal, and that \( f(c) \) is the formula used to compute \( A \) and \( B \).

Definition 2.8. Given an estimate \( b \) of \( \beta \), the corresponding vector of predicted values or fitted values is \( \hat{Y} \equiv \hat{Y}(b) = Xb \). Thus the ith fitted value

\[
\hat{Y}_i \equiv \hat{Y}_i(b) = x_i^T b = x_{i,1} b_1 + \cdots + x_{i,p} b_p.
\]

The vector of residuals is \( r \equiv r(b) = Y - \hat{Y}(b) \). Thus ith residual \( r_i \equiv r_i(b) = Y_i - \hat{Y}_i(b) = Y_i - x_{i,1} b_1 - \cdots - x_{i,p} b_p \).

Most regression methods attempt to find an estimate \( \hat{\beta} \) of \( \beta \) which minimizes some criterion function \( Q(b) \) of the residuals.

Definition 2.9. The ordinary least squares (OLS) estimator \( \hat{\beta}_{OLS} \) minimizes

\[
Q_{OLS}(b) = \sum_{i=1}^{n} r_i^2(b), \tag{2.4}
\]

and \( \hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \).

The vector of predicted or fitted values \( \hat{Y}_{OLS} = X \hat{\beta}_{OLS} = H Y \) where the hat matrix \( H = X(X^T X)^{-1} X^T \) provided the inverse exists. Typically the subscript OLS is omitted, and the least squares regression equation is \( \hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \) where \( x_1 \equiv 1 \) if the model contains a constant.

There are many statistical models besides the MLR model, and you should learn how to quickly recognize an MLR model. A “regression” model has a response variable \( Y \) and the conditional distribution of \( Y \) given the predictors \( x = (x_1, \ldots, x_p)^T \) is of interest. Regression models are used to predict \( Y \) and to summarize the relationship between \( Y \) and \( x \). If a constant \( x_{i,1} = 1 \) (this notation means that \( x_{i,1} = 1 \) for \( i = 1, \ldots, n \)) is in the model, then \( x_{i,1} \) is often called the trivial predictor, and the MLR model is said to have
a constant or intercept. All nonconstant predictors are called nontrivial predictors. The term “multiple” is used if the model uses one or more nontrivial predictors. (Some authors use multivariate instead of multiple, but in this text a multivariate linear regression model will have \( m \geq 2 \) response variables. See Chapter 12.) The simple linear regression model is a special case of the MLR model that uses exactly one nontrivial predictor. Suppose the response variable is \( Y \) and data has been collected on additional variables \( x_1, \ldots, x_p \).

An MLR model is “linear” in the unknown coefficients \( \beta \). Thus the model is an MLR model in \( Y \) and \( \beta \) if we can write

\[
Y_i = x_i^T \beta + e_i
\]

where each \( w_i \) is a function of \( x_1, \ldots, x_p \). Symbols other than \( w \) or \( x \) may be used. Alternatively, the model is linear in the parameters \( \beta \) if \( \partial Y / \partial \beta_i \) does not depend on the parameters. If

\[
Y = x^T \beta + e = x_1 \beta_1 + \cdots + x_p \beta_p + e,
\]

then \( \partial Y / \partial \beta_i = x_i \). Similarly, if \( Y = w^T \beta + e \), then \( \partial Y / \partial \beta_i = w_i \).

**Example 2.2.**

a) Suppose that interest is in predicting a function of \( Z \) from functions of \( w_1, \ldots, w_k \). If \( Y = t(Z) = x^T \beta + e \) where \( t \) is a function and each \( x_i \) is some function of \( w_1, \ldots, w_k \), then there is an MLR model in \( Y \) and \( \beta \). Similarly, \( Z = t(Y) = w^T \beta + e \) is an MLR model in \( Z \) and \( \beta \).

b) To see that \( Y = \beta_1 + \beta_2 x + \beta_3 x^2 + e \) is an MLR model in \( Y \) and \( \beta \), take \( w_1 = 1 \), \( w_2 = x \), and \( w_3 = x^2 \). Then \( Y = w^T \beta + e \).

c) If \( Y = \beta_1 + \beta_2 \exp(\beta_3 x) + e \), then the model is a nonlinear regression model that is not an MLR model in \( Y \) and \( \beta \). Notice that the model can not be written in the form \( Y = w^T \beta + e \) and that \( \partial Y / \partial \beta_2 = \exp(\beta_3 x) \) and \( \partial Y / \partial \beta_3 = \beta_2 x \exp(\beta_3 x) \) depend on the parameters.

### 2.2 Checking Goodness of Fit

It is crucial to realize that an MLR model is not necessarily a useful model for the data, even if the data set consists of a response variable and several predictor variables. For example, a nonlinear regression model or a much more complicated model may be needed. Chapters 1 and 13 describe several alternative models. Let \( p \) be the number of predictors and \( n \) the number of cases. Assume that \( n \geq 5p \), then plots can be used to check whether the MLR model is useful for studying the data. This technique is known as checking the goodness of fit of the MLR model.

**Notation.** Plots will be used to simplify regression analysis, and in this text a plot of \( W \) versus \( Z \) uses \( W \) on the horizontal axis and \( Z \) on the vertical axis.

**Definition 2.10.** A scatterplot of \( X \) versus \( Y \) is a plot of \( X \) versus \( Y \) and is used to visualize the conditional distribution \( Y|X \) of \( Y \) given \( X \).
Definition 2.11. A response plot is a plot of a variable $w_i$ versus $Y_i$. Typically $w_i$ is a linear combination of the predictors: $w_i = x_i^T \eta$ where $\eta$ is a known $p \times 1$ vector. The most commonly used response plot is a plot of the fitted values $\hat{Y}_i$ versus the response $Y_i$.

Proposition 2.1. Suppose that the regression estimator $b$ of $\beta$ is used to find the residuals $r_i \equiv r_i(b)$ and the fitted values $\hat{Y}_i \equiv \hat{Y}_i(b) = x_i^T b$. Then in the response plot of $\hat{Y}_i$ versus $Y_i$, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(b)$.

Proof. The identity line in the response plot is $Y = x^T b$. Hence the vertical deviation is $Y_i - x_i^T b = r_i(b)$. □

Definition 2.12. A residual plot is a plot of a variable $w_i$ versus the residuals $r_i$. The most commonly used residual plot is a plot of $\hat{Y}_i$ versus $r_i$.

Notation: For MLR, “the residual plot” will often mean the residual plot of $\hat{Y}_i$ versus $r_i$, and “the response plot” will often mean the plot of $\hat{Y}_i$ versus $Y_i$.

If the unimodal MLR model as estimated by least squares is useful, then in the response plot the plotted points should scatter about the identity line while in the residual plot of $\hat{Y}$ versus $r$ the plotted points should scatter about the $r = 0$ line (the horizontal axis) with no other pattern. Figures 1.2 and 1.3 show what a response plot and residual plot look like for an artificial MLR data set where the MLR regression relationship is rather strong in that the sample correlation $\text{corr}(\hat{Y}, Y)$ is near 1. Figure 1.4 shows a response plot where the response $Y$ is independent of the nontrivial predictors in the model. Here $\text{corr}(\hat{Y}, Y)$ is near 0 but the points still scatter about the identity line. When the MLR relationship is very weak, the response plot will look like the residual plot.

The above ideal shapes for the response and residual plots are for when the unimodal MLR model gives a good approximation for the data. If the plots have the ideal shapes and $n \geq 10p$, then expect inference, except for classical prediction intervals, to be approximately correct for many unimodal distributions that are close to the normal distribution.

If the response and residual plots suggest an MLR model with iid skewed errors, then add lowess to both plots. The scatterplot smoother tries to estimate the mean function $E(Y|\hat{Y})$ or $E(r|\hat{Y})$ without using any model. If the lowess curve is close to the identity line in the response plot and close to the $r = 0$ line in the residual plot, then the constant variance MLR model may be a good approximation to the data, but sample sizes much larger than $n = 10p$ may be needed before inference is approximately correct. Such skewed data sets seem rather rare, but see Chen et al. (2009) and see Problem 2.28.

Remark 2.1. For any MLR analysis, always make the response plot and the residual plot of $\hat{Y}_i$ versus $Y_i$ and $r_i$, respectively.
**Definition 2.13.** An outlier is an observation that lies far away from the bulk of the data.

**Remark 2.2.** For MLR, the response plot is important because MLR is the study of the conditional distribution of $Y|\mathbf{x}^T\beta$, and the response plot is used to visualize the conditional distribution of $Y|\mathbf{x}^T\beta$ since $\hat{Y} = \mathbf{x}^T\hat{\beta}$ is a good estimator of $\mathbf{x}^T\beta$ if $\hat{\beta}$ is a good estimator of $\beta$.

If the MLR model is useful, then the plotted points in the response plot should be linear and scatter about the identity line with no gross outliers. Suppose the fitted values range in value from $w_L$ to $w_H$ with no outliers. Fix the fit $= w$ in this range and mentally add a narrow vertical strip centered at $w$ to the response plot. The plotted points in the vertical strip should have a mean near $w$ since they scatter about the identity line. Hence $Y|\text{fit} = w$ is like a sample from a distribution with mean $w$. The following example helps illustrate this remark.

**Example 2.3.** Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases because of missing values and used *height* as the response variable $Y$. Along with a
constant \( x_{i,1} \equiv 1 \), the five additional predictor variables used were height when sitting, height when kneeling, head length, nasal breadth, and span (perhaps from left hand to right hand). Figure 2.1 presents the OLS response and residual plots for this data set. These plots show that an MLR model should be a useful model for the data since the plotted points in the response plot are linear and follow the identity line while the plotted points in the residual plot follow the \( r = 0 \) line with no other pattern (except for a possible outlier marked 44).

To use the response plot to visualize the conditional distribution of \( Y | \mathbf{x}^T \beta \), use the fact that the fitted values \( \hat{Y} = \mathbf{x}^T \hat{\beta} \). For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1685 to 1715. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit = \( w \) for \( w \) between 1500 and 1850, the cases have heights near \( w \), on average.

Cases 3, 44, and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as outliers. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response plot and about the horizontal line for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the \( r = 0 \) line. In Figure 2.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The identity line can also pass through or near an outlier or a cluster of outliers. Then the outliers will be in the upper right or lower left of the response plot, and there will be a large gap between the cluster of outliers and the bulk of the data. See Figure 3.11. Figure 2.1 was made with the following R commands, using bregpack function MLRplot and the major.lsp data set from the text’s webpage.

```r
major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
# copy and paste the data set then press enter
major <- major[,,-1]
X<-major[,,-6]
Y <- major[,6]
MLRplot(X,Y) # left click the 3 highlighted cases,
# then right click Stop for each of the two plots
```
2.3 Checking Lack of Fit

The response plot may look good while the residual plot suggests that the unimodal MLR model can be improved. Examining plots to find model violations is called checking for lack of fit. Again assume that \( n \geq 5p \).

The unimodal MLR model often provides a useful model for the data, but the following assumptions do need to be checked.

i) Is the MLR model appropriate?
ii) Are outliers present?
iii) Is the error variance constant or nonconstant? The constant variance assumption \( \text{VAR}(e_i) \equiv \sigma^2 \) is known as homoscedasticity. The nonconstant variance assumption \( \text{VAR}(e_i) = \sigma_i^2 \) is known as heteroscedasticity.
iv) Are any important predictors left out of the model?
v) Are the errors \( e_1, \ldots, e_n \) iid?
vii) Are the errors \( e_i \) independent of the predictors \( x_i \)?

Make the response plot and the residual plot to check i), ii), and iii). An MLR model is reasonable if the plots look like Figures 1.2, 1.3, 1.4, and 2.1. A response plot that looks like Figure 13.7 suggests that the model is not linear. If the plotted points in the residual plot do not scatter about the \( r = 0 \) line with no other pattern (i.e., if the cloud of points is not ellipsoidal or rectangular with zero slope), then the unimodal MLR model is not sustained.

The \( i \)th residual \( r_i \) is an estimator of the \( i \)th error \( e_i \). The constant variance assumption may have been violated if the variability of the point cloud in the residual plot depends on the value of \( \hat{Y} \). Often the variability of the residuals increases as \( \hat{Y} \) increases, resulting in a right opening megaphone shape. (Figure 4.1b has this shape.) Often the variability of the residuals decreases as \( \hat{Y} \) increases, resulting in a left opening megaphone shape. Sometimes the variability decreases then increases again, and sometimes the variability increases then decreases again (like a stretched or compressed football).

2.3.1 Residual Plots

Remark 2.3. Residual plots magnify departures from the model while the response plot emphasizes how well the MLR model fits the data.

Since the residuals \( r_i = \hat{e}_i \) are estimators of the errors, the residual plot is used to visualize the conditional distribution \( e|SP \) of the errors given the sufficient predictor \( SP = x^T \beta \), where \( SP \) is estimated by \( \hat{Y} = x^T \hat{\beta} \). For the unimodal MLR model, there should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change.
Notation. A rule of thumb is a rule that often but not always works well in practice.

Rule of thumb 2.1. If the residual plot would look good after several points have been deleted, and if these deleted points were not gross outliers (points far from the point cloud formed by the bulk of the data), then the residual plot is probably good. Beginners often find too many things wrong with a good model. For practice, use the \texttt{bregpack} function \texttt{MLRsims} to generate several MLR data sets, and make the response and residual plots for these data sets: type \texttt{MLRsims(nruns=10)} in \texttt{R} and right click Stop for each plot (20 times) to generate 10 pairs of response and residual plots. This exercise will help show that the plots can have considerable variability even when the MLR model is good. See Problem 2.30.

Rule of thumb 2.2. If the plotted points in the residual plot look like a left or right opening megaphone, the first model violation to check is the assumption of nonconstant variance. (This is a rule of thumb because it is possible that such a residual plot results from another model violation such as nonlinearity, but nonconstant variance is much more common.)

The residual plot of $\hat{Y}$ versus $r$ should always be made. It is also a good idea to plot each nontrivial predictor $x_j$ versus $r$ and to plot potential predictors $w_j$ versus $r$. If the predictor is quantitative, then the residual plot of $x_j$ versus $r$ should look like the residual plot of $\hat{Y}$ versus $r$. If the predictor is qualitative, e.g. gender, then interpreting the residual plot is much more difficult; however, if each category contains many observations, then the plotted points for each category should form a vertical line centered at $r = 0$ with roughly the same variability (spread or range).

Rule of thumb 2.3. Suppose that the MLR model uses predictors $x_j$ and that data has been collected on variables $w_j$ that are not included in the MLR model. To check whether important predictors have been left out, make residual plots of $x_j$ and $w_j$ versus $r$. If these plots scatter about the $r = 0$ line with no other pattern, then there is no evidence that $x_j^2$ or $w_j$ are needed in the model. If the plotted points scatter about a parabolic curve, try adding $x_j^2$ or $w_j$ and $w_j^2$ to the MLR model. If the plot of the potential predictor $w_j$ versus $r$ has a linear trend, try adding $w_j$ to the MLR model. The additive error regression model and EE plot in Section 13.7 can also be used to check whether important predictors have been left out.

Rule of thumb 2.4. To check that the errors are independent of the predictors, make residual plots of $x_j$ versus $r$. If the plot of $x_j$ versus $r$ scatters about the $r = 0$ line with no other pattern, then there is no evidence that the errors depend on $x_j$. If the variability of the residuals changes with the value of $x_j$, e.g. if the plot resembles a left or right opening megaphone, the errors may depend on $x_j$. Some remedies for nonconstant variance are considered in Chapter 4.
Multiple Linear Regression

To study residual plots, some notation and properties of the least squares estimator are needed. MLR is the study of the conditional distribution of $Y_i | x_i^T \beta$, and the MLR model is $Y = X\beta + e$ where $X$ is an $n \times p$ matrix of full rank $p$. Hence the number of predictors $p \leq n$. The $i$th row of $X$ is $x_i^T = (x_{i,1}, \ldots, x_{i,p})$ where $x_{i,k}$ is the value of the $i$th observation on the $k$th predictor $x_k$. We will denote the $j$th column of $X$ by $X_j \equiv v_j$ which corresponds to the $j$th variable or predictor $x_j$.

Example 2.4. If $Y$ is brain weight in grams, $x_1 \equiv 1$, $x_2$ is age, and $x_3$ is the size of the head in (mm$^3$), then for the Gladstone (1905) data

$$Y = \begin{bmatrix} 3738 \\ 4261 \\ \vdots \\ 3306 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 39 & 149.5 \\ 1 & 35 & 152.5 \\ \vdots & \vdots & \vdots \\ 1 & 19 & 141 \end{bmatrix} = [v_1 \ v_2 \ v_3].$$

Hence the first person had brain weight $= 3738$, age $= 39$, and size $= 149.5$.

After deleting observations with missing values, there were $n = 267$ cases (people measured on brain weight, age, and size), and $x_{267} = (1, 19, 141)^T$. The second predictor $x_2 = age$ corresponds to the 2nd column of $X$ and is $X_2 = v_2 = (39, 35, \ldots, 19)^T$. Notice that $X_1 \equiv v_1 = 1 = (1, \ldots, 1)^T$ corresponds to the constant $x_1$.

The results in the following proposition are properties of least squares (OLS), not of the underlying MLR model. See Chapter 11 for more linear model theory. Definitions 2.8 and 2.9 define the hat matrix $H$, vector of fitted values $\hat{Y}$, and vector of residuals $r$. Parts f) and g) make residual plots useful. If the plotted points are linear with roughly constant variance and the correlation is zero, then the plotted points scatter about the $r = 0$ line with no other pattern. If the plotted points in a residual plot of $w$ versus $r$ do show a pattern such as a curve or a right opening megaphone, zero correlation will usually force symmetry about either the $r = 0$ line or the $w = \text{median}(w)$ line. Hence departures from the ideal plot of random scatter about the $r = 0$ line are often easy to detect.

**Warning:** If $n > p$, as is usually the case, $X$ is not square, so $(X^T X)^{-1} \neq X^{-1}(X^T)^{-1}$ since $X^{-1}$ does not exist.

**Proposition 2.2.** Suppose that $X$ is an $n \times p$ matrix of full rank $p$. Then

a) $H$ is symmetric: $H = H^T$.

b) $H$ is idempotent: $HH = H$.

c) $X^T r = 0$ so that $X_j^T r = v_j^T r = 0$.

d) If there is a constant $X_1 \equiv v_1 = 1$ in the model, then the sum of the residuals is zero: $\sum_{i=1}^n r_i = 0$.

e) $r^T \hat{Y} = 0$.

f) If there is a constant in the model, then the sample correlation of the fitted values and the residuals is 0: $\text{corr}(r, \hat{Y}) = 0$. 
2.3 Checking Lack of Fit

If there is a constant in the model, then the sample correlation of the
jth predictor with the residuals is 0: \( \text{corr}(r, v_j) = 0 \) for \( j = 1, \ldots, p \).

Proof. a) \( X^T X \) is symmetric since \( (X^T X)^T = X^T (X^T)^T = X^T X \).
Hence \( X^T X \)^{-1} is symmetric since the inverse of a symmetric matrix is
symmetric. (Recall that if \( A \) has an inverse then \( (A^T)^{-1} = (A^{-1})^T \).) Thus
using \( (A^T)^T = A \) and \( (ABC)^T = C^T B^T A^T \) shows that
\[
H^T = X^T (X^T X)^{-1} X^T X (X^T X)^{-1} X^T = H.
\]

b) \( HH = X (X^T X)^{-1} X^T X (X^T X)^{-1} X^T = H \) since \( (X^T X)^{-1} X^T X = I_p \),
the \( p \times p \) identity matrix.

c) \( X^T r = X^T (I_p - H) Y = [X^T - X^T X (X^T X)^{-1} X^T] Y = [X^T - X^T] Y = 0 \).
Since \( v_j \) is the \( j \)th column of \( X \), \( v_j^T \) is the \( j \)th row of \( X^T \)
and \( v_j^T r = 0 \) for \( j = 1, \ldots, p \).

d) Since \( v_1 = 1 \), \( v_1^T r = \sum_{i=1}^n r_i = 0 \) by c).

e) \( r^T \hat{Y} = [(I_n - H) Y]^T HY = Y^T (I_n - H) HY = Y^T (H - H) Y = 0 \).

f) The sample correlation between \( W \) and \( Z \) is \( \text{corr}(W, Z) = \frac{\sum_{i=1}^n (w_i - \bar{w})(z_i - \bar{z})}{(n - 1)s_w s_z} = \frac{\sum_{i=1}^n (w_i - \bar{w})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^n (w_i - \bar{w})^2 \sum_{i=1}^n (z_i - \bar{z})^2}} \)
where \( s_m \) is the sample standard deviation of \( m \) for \( m = w, z \). So the result
follows if \( A = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(r_i - \bar{r}) = 0 \). Now \( \bar{r} = 0 \) by d), and thus
\[
A = \sum_{i=1}^n \hat{Y}_i r_i - \bar{Y} \sum_{i=1}^n r_i = \sum_{i=1}^n \hat{Y}_i r_i
\]
by d) again. But \( \sum_{i=1}^n \hat{Y}_i r_i = r^T \hat{Y} = 0 \) by e).

g) Following the argument in f), the result follows if \( A = \sum_{i=1}^n (x_{i,j} - \bar{x}_j)(r_i - \bar{r}) = 0 \) where \( \bar{x}_j = \sum_{i=1}^n x_{i,j}/n \) is the sample mean of
the \( j \)th predictor. Now \( \bar{r} = \sum_{i=1}^n r_i/n = 0 \) by d), and thus
\[
A = \sum_{i=1}^n x_{i,j} r_i - \bar{x}_j \sum_{i=1}^n r_i = \sum_{i=1}^n x_{i,j} r_i
\]
by d) again. But \( \sum_{i=1}^n x_{i,j} r_i = X_j^T r = v_j^T r = 0 \) by c). □
2.3.2 Other Model Violations

Without loss of generality, $E(e) = 0$ for the unimodal MLR model with a constant, in that if $E(\tilde{e}) = \mu \neq 0$, then the MLR model can always be written as $Y = x^T \beta + e$ where $E(e) = 0$ and $E(Y) = E(Y|x) = x^T \beta$. To see this claim notice that

\[
Y = \tilde{\beta}_1 + x_2 \beta_2 + \cdots + x_p \beta_p + \tilde{e} = \tilde{\beta}_1 + E(\tilde{e}) + x_2 \beta_2 + \cdots + x_p \beta_p + \tilde{e} - E(\tilde{e})
\]

\[
= \beta_1 + x_2 \beta_2 + \cdots + x_p \beta_p + e
\]

where $\beta_1 = \tilde{\beta}_1 + E(\tilde{e})$ and $e = \tilde{e} - E(\tilde{e})$. For example, if the errors $\tilde{e}_i$ are iid exponential ($\lambda$) with $E(\tilde{e}_i) = \lambda$, use $e_i = \tilde{e}_i - \lambda$.

For least squares, it is crucial that $\sigma^2$ exists. For example, if the $e_i$ are iid Cauchy(0,1), then $\sigma^2$ does not exist and the least squares estimators tend to perform very poorly.

The performance of least squares is analogous to the performance of $\overline{Y}$. The sample mean $\overline{Y}$ is a very good estimator of the population mean $\mu$ if the $Y_i$ are iid $N(\mu, \sigma^2)$, and $\overline{Y}$ is a good estimator of $\mu$ if the sample size is large and the $Y_i$ are iid with mean $\mu$ and variance $\sigma^2$. This result follows from the central limit theorem (CLT), but how “large is large” depends on the underlying distribution. The $n > 30$ rule tends to hold for distributions that are close to normal in that they take on many values and $\sigma^2$ is not huge. Error distributions that are highly nonnormal with tiny $\sigma^2$ often need $n >> 30$. For example, if $Y_1, \ldots, Y_n$ are iid Gamma($1/m, 1$), then $n > 25m$ may be needed. Another example is distributions that take on one value with very high probability, e.g. a Poisson random variable with very small variance. Bimodal and multimodal distributions and highly skewed distributions with large variances also need larger $n$. Chihara and Hesterberg (2011, p. 177) suggest using $n > 5000$ for moderately skewed distributions.

There are central limit type theorems for the least squares estimators that depend on the error distribution of the iid errors $e_i$. See Theorems 2.8, 11.25, and 12.7. We always assume that the $e_i$ are continuous random variables with a probability density function. Error distributions that are close to normal may give good results for moderate $n$ if $n \geq 10p$ and $n - p \geq 30$ where $p$ is the number of predictors. Error distributions that need large $n$ for the CLT to apply for $\overline{e}$, will tend to need large $n$ for the limit theorems for least squares to apply (to give good approximations).

Checking whether the errors are iid is often difficult. The iid assumption is often reasonable if measurements are taken on different objects, e.g. people. In industry often several measurements are taken on a batch of material. For example a batch of cement is mixed and then several small cylinders of concrete are made from the batch. Then the cylinders are tested for strength.
Experience from such experiments suggests that objects (e.g., cylinders) from different batches are independent, but objects from the same batch are not independent.

One check on independence can also be made if the time order of the observations is known. Let \( r[t] \) be the residual where \([t]\) is the time order of the trial. Hence \([1]\) was the 1st and \([n]\) was the last trial. Plot the time order \( t \) versus \( r[t] \) if the time order is known. Again, trends and outliers suggest that the model could be improved. A box shaped plot with no trend suggests that the MLR model is good. A plot similar to the Durbin Watson test plots \( r[t−1] \) versus \( r[t] \) for \( t = 2, \ldots, n \). Linear trend suggests serial correlation while random scatter suggests that there is no lag 1 autocorrelation. As a rule of thumb, if the OLS slope \( b \) is computed for the plotted points, \( b > 0.25 \) gives some evidence that there is positive correlation between \( r[t−1] \) and \( r[t] \). Time series plots, such as the ACF or PACF of the residuals, may be useful.

If it is assumed that the error distribution is symmetric, make a histogram of the residuals. Check whether the histogram is roughly symmetric or clearly skewed. If it is assumed that the errors \( e_i \) are iid \( N(0, \sigma^2) \) again check whether the histogram is mound shaped with “short tails.” A commonly used alternative is to make a normal probability plot of the residuals. Let \( r(1) < r(2) < \cdots < r(n) \) denote the residuals ordered from smallest to largest. Hence \( r(1) \) is the value of the smallest residual. The normal probability plot plots the \( \tilde{e}(i) \) versus \( r(i) \) where the \( \tilde{e}(i) \) are the expected values of the order statistics from a sample of size \( n \) from an \( N(0, 1) \) distribution. (Often the \( \tilde{e}(i) \) are the standard normal percentiles that satisfy \( P(Z \leq \tilde{e}(i)) = (i−0.5)/n \) where \( Z \sim N(0, 1) \).

Rules of thumb: i) if the plotted points scatter about some straight line in the normal probability plot, then there is no evidence against the normal assumption. ii) if the plotted points have an “ess shape” (concave up then concave down), then the error distribution is symmetric with lighter tails than the normal distribution. iii) If the plot resembles a cubic function, then the error distribution is symmetric with heavier tails than the normal distribution. iv) If the plotted points look concave up (e.g. like \( x^2 \) where \( x > 0 \)), then the error distribution is right skewed.

2.4 The ANOVA F Test

After fitting least squares and checking the response and residual plots to see that an MLR model is reasonable, the next step is to check whether there is an MLR relationship between \( Y \) and the nontrivial predictors \( x_2, \ldots, x_p \). If at least one of these predictors is useful, then the OLS fitted values \( \hat{Y}_i \) should be used. If none of the nontrivial predictors is useful, then \( \bar{Y} \) will give as good predictions as \( \hat{Y}_i \). Here the sample mean
\[ \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \]  

(2.5)

In the definition below, \( SSE \) is the sum of squared residuals and a residual \( r_i = \hat{e}_i = \text{“errorhat.”} \) In the literature “errorhat” is often rather misleadingly abbreviated as “error.”

**Definition 2.14.** Assume that a constant is in the MLR model.

a) The total sum of squares

\[ SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \]  

(2.6)

b) The regression sum of squares

\[ SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2. \]  

(2.7)

c) The residual sum of squares or error sum of squares is

\[ SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \]  

(2.8)

The result in the following proposition is a property of least squares (OLS), not of the underlying MLR model. An obvious application is that given any two of \( SSTO, SSE, \) and \( SSR, \) the 3rd sum of squares can be found using the formula \( SSTO = SSE + SSR. \)

**Proposition 2.3.** Assume that a constant is in the MLR model. Then \( SSTO = SSE + SSR. \)

**Proof.**

\[ SSTO = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = SSE + SSR + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}). \]

Hence the result follows if

\[ A = \sum_{i=1}^{n} r_i (\hat{Y}_i - \bar{Y}) = 0. \]

But

\[ A = \sum_{i=1}^{n} r_i \hat{Y}_i - \bar{Y} \sum_{i=1}^{n} r_i = 0 \]

by Proposition 2.2 d) and e). □
Definition 2.15. Assume that a constant is in the MLR model and that SSTO \( \neq 0 \). The coefficient of multiple determination

\[
R^2 = [\text{corr}(Y_i, \hat{Y}_i)]^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}
\]

where \( \text{corr}(Y_i, \hat{Y}_i) \) is the sample correlation of \( Y_i \) and \( \hat{Y}_i \).

Warnings: i) \( 0 \leq R^2 \leq 1 \), but small \( R^2 \) does not imply that the MLR model is bad.

ii) If the MLR model contains a constant, then there are several equivalent formulas for \( R^2 \). If the model does not contain a constant, then \( R^2 \) depends on the software package.

iii) \( R^2 \) does not have much meaning unless the response plot and residual plot both look good.

iv) \( R^2 \) tends to be too high if \( n \) is small.

v) \( R^2 \) tends to be too high if there are two or more separated clusters of data in the response plot.

vi) \( R^2 \) is too high if the number of predictors \( p \) is close to \( n \).

vii) In large samples \( R^2 \) will be large (close to one) if \( \sigma^2 \) is small compared to the sample variance \( S_Y^2 \) of the response variable \( Y \). \( R^2 \) is also large if the sample variance of \( \hat{Y} \) is close to \( S_Y^2 \). Thus \( R^2 \) is sometimes interpreted as the proportion of the variability of \( Y \) explained by conditioning on \( x \), but warnings i) - v) suggest that \( R^2 \) may not have much meaning.

The following 2 propositions suggest that \( R^2 \) does not behave well when many predictors that are not needed in the model are included in the model. Such a variable is sometimes called a noise variable and the MLR model is “fitting noise.” Proposition 2.5 appears, for example, in Cramér (1946, pp. 414–415), and suggests that \( R^2 \) should be considerably larger than \( p/n \) if the predictors are useful. Note that if \( n = 10p \) and \( p \geq 2 \), then under the conditions of Proposition 2.5, \( E(R^2) \leq 0.1 \).

Proposition 2.4. Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) \( R^2 \).

Proposition 2.5. Assume that a constant \( \beta_1 \) is in the MLR model, that \( \beta_2 = \cdots = \beta_p = 0 \) and that the \( e_i \) are iid \( N(0, \sigma^2) \). Hence the \( Y_i \) are iid \( N(\beta_1, \sigma^2) \). Then

a) \( R^2 \) follows a beta distribution: \( R^2 \sim \text{beta}(\frac{p-1}{2}, \frac{n-p}{2}) \).

b)

\[
E(R^2) = \frac{p-1}{n-1}.
\]

c)

\[
\text{VAR}(R^2) = \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}.
\]
Notice that each $SS/n$ estimates the variability of some quantity. $SSTO/n \approx S^2_Y$, $SSE/n \approx S^2_e = \sigma^2$, and $SSR/n \approx S^2_{\hat{Y}}$.

**Definition 2.16.** Assume that a constant is in the MLR model. Associated with each SS in Definition 2.14 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, $df = n - 1$ and $MSTO = SSTO/(n - 1)$. For SSR, $df = p - 1$ and $MSR = SSR/(p - 1)$. For SSE, $df = n - p$ and $MSE = SSE/(n - p)$.

Seber and Lee (2003, pp. 44–47) show that when the MLR model holds, MSE is often a good estimator of $\sigma^2$. Under regularity conditions, the MSE is one of the best unbiased quadratic estimators of $\sigma^2$. For the normal MLR model, MSE is the uniformly minimum variance unbiased estimator of $\sigma^2$. Seber and Lee also give the following theorem that shows that the MSE is an unbiased estimator of $\sigma^2$ under very weak assumptions if the MLR model is appropriate. From Theorem 12.7 MSE is a $\sqrt{n}$ consistent estimator of $\sigma^2$.

**Theorem 2.6.** If $Y = X\beta + e$ where $X$ is an $n \times p$ matrix of full rank $p$, if the $e_i$ are independent with $E(e_i) = 0$, and $\text{VAR}(e_i) = \sigma^2$, then $\hat{\sigma}^2 = MSE$ is an unbiased estimator of $\sigma^2$.

The ANOVA $F$ test tests whether any of the nontrivial predictors $x_2, \ldots, x_p$ are needed in the OLS MLR model, that is, whether $Y_i$ should be predicted by the OLS fit $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p$ or with the sample mean $\bar{Y}$. ANOVA stands for analysis of variance, and the computer output needed to perform the test is contained in the ANOVA table. Below is an ANOVA table given in symbols. Sometimes “Regression” is replaced by “Model” and “Residual” by “Error.”

**Summary Analysis of Variance Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>p-1</td>
<td>SSR</td>
<td>MSR</td>
<td>$F_o = MSR/MSE$ for $H_0$: $\beta_2 = \cdots = \beta_p = 0$</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Remark 2.4.** Recall that for a 4 step test of hypotheses, the p–value is the probability of getting a test statistic as extreme as the test statistic actually observed and that $H_0$ is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if $\delta$ is not given. The 4th step is the nontechnical conclusion which is crucial for presenting your results to people who are not familiar with MLR. Replace $Y$ and $x_2, \ldots, x_p$ by the actual variables used in the MLR model. Follow Example 2.5.

**Notation.** The p–value $\equiv \text{pval}$ given by output tends to only be correct for the normal MLR model. Hence the output is usually only giving an estimate of the pval, which will often be denoted by $\text{pval}$. So reject $H_0$ if $\text{pval} \leq \delta$. Often
2.4 The ANOVA F Test

\[ p_{\text{val}} \xrightarrow{\text{P}} 0 \]

(converges to 0 in probability, so pval is a consistent estimator of pvalue) as the sample size \( n \to \infty \). See Theorem 11.25, Section 11.6, and Chang and Olive (2010). Then the computer output pval is a good estimator of the unknown pvalue. We will use \( H_0 \equiv H_0 \) and \( H_a \equiv H_A \equiv H_1 \).

**Be able to perform the 4 step ANOVA F test of hypotheses.**

i) State the hypotheses \( H_0: \beta_2 = \cdots = \beta_p = 0 \) \( Ha: \) not \( H_0 \).

ii) Find the test statistic \( F_o = \frac{MSR}{MSE} \) or obtain it from output.

iii) Find the pval from output or use the \( F \)-table: \( p_{\text{val}} = P(F_{p-1,n-p} > F_o) \).

iv) State whether you reject \( H_0 \) or fail to reject \( H_0 \). If \( H_0 \) is rejected, conclude that there is an MLR relationship between \( Y \) and the predictors \( x_2, \ldots, x_p \). If you fail to reject \( H_0 \), conclude that there is not an MLR relationship between \( Y \) and the predictors \( x_2, \ldots, x_p \). (Or there is not enough evidence to conclude that there is an MLR relationship between \( Y \) and the predictors.)

**Example 2.5.** For the Gladstone (1905) data, the response variable \( Y = \text{brain weight}, x_1 = 1, x_2 = \text{size of head}, x_3 = \text{sex}, x_4 = \text{breadth of head}, x_5 = \text{circumference of head} \). Assume that the response and residual plots look good and test whether at least one of the nontrivial predictors is needed in the model using the output shown below.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4</td>
<td>5396942.</td>
<td>1349235.</td>
<td>196.24</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>262</td>
<td>1801333.</td>
<td>6875.32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Solution: i) \( H_0: \beta_2 = \cdots = \beta_5 = 0 \) \( Ha: \) not \( H_0 \)

ii) \( F_o = 196.24 \) from output.

iii) The pval = 0.0 from output.

iv) The pval < \( \delta \) (= 0.05 since \( \delta \) was not given). So reject \( H_0 \). Hence there is an MLR relationship between brain weight and the predictors size, sex, breadth, and circumference.

**Remark 2.5.** There is a close relationship between the response plot and the ANOVA \( F \) test. If \( n \geq 10p \) and \( n - p \geq 30 \) and if the plotted points follow the identity line, typically \( H_0 \) will be rejected if the identity line fits the plotted points better than any horizontal line (in particular, the line \( Y = \bar{Y} \)). If a horizontal line fits the plotted points about as well as the identity line, as in Figure 1.4, this graphical diagnostic is inconclusive (sometimes the ANOVA \( F \) test will reject \( H_0 \) and sometimes fail to reject \( H_0 \)), but the MLR relationship is at best weak. In Figures 1.2 and 2.1, the ANOVA \( F \) test should reject \( H_0 \) since the identity line fits the plotted points better than any horizontal line. Under the above conditions, a graphical ANOVA \( F \) test
rejects Ho if the response plot is not similar to the residual plot. The graphical test is inconclusive if the response plot looks similar to the residual plot. The graphical test is also useful for multiple linear regression methods other than least squares, such as M-estimators and other robust regression estimators.

**Definition 2.17.** An RR plot is a plot of residuals from 2 different models or fitting methods.

**Remark 2.6.** If the RR plot of the residuals \( Y_i - \bar{Y} \) versus the OLS residuals \( r_i = Y_i - \hat{Y}_i \) shows tight clustering about the identity line, then the MLR relationship is weak: \( \bar{Y} \) fits the data about as well as the OLS fit.

**Example 2.6.** Cook and Weisberg (1999a, pp. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The response \( Y \) was the fraction of the drug recovered from the rat’s liver. The three predictors were the body weight of the rat, the dose of the drug, and the liver weight. A constant was also used. The experimenter expected the response to be independent of the predictors, and 19 cases were used. However, the ANOVA \( F \) test suggested that the predictors were important. The third case was an outlier and easily detected in the response and residual plots (not shown). After deleting the outlier, the response and residual plots looked ok and the following output was obtained.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>0.00184396</td>
<td>0.000614652</td>
<td>0.10</td>
<td>0.9585</td>
</tr>
<tr>
<td>Residual</td>
<td>14</td>
<td>0.0857172</td>
<td>0.00612265</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 4 step ANOVA \( F \) test is
i) \( \text{Ho}: \beta_2 = \cdots = \beta_4 = 0 \) Ha: not Ho
ii) \( F_o = 0.10 \).
iii) pval = 0.9585.
iv) The pval > \( \delta \) (= 0.05 since \( \delta \) was not given). So fail to reject Ho. Hence there is not an MLR relationship between fraction of drug recovered and the predictors body weight, dose, and liver weight. (More accurately, there is not enough statistical evidence to conclude that there is an MLR relationship: failing to reject Ho is not the same as accepting Ho; however, it may be a good idea to keep the nontechnical conclusions nontechnical.)

Figure 2.2 shows the RR plot where the residuals from the full model are plotted against \( Y_i - \bar{Y} \), the residuals from the model using no nontrivial predictors. This plot reinforces the conclusion that the response \( Y \) is independent of the nontrivial predictors. The identity line and the OLS line from regressing \( r_i \) on \( Y_i - \bar{Y} \) (that is, use \( \bar{Y}_i = r_i \), a constant and \( \hat{x}_{i,2} = Y_i - \bar{Y} \), find the OLS line and then plot it) are shown as visual aids. If the OLS line and identity line nearly coincide in that it is difficult to tell that the two lines intersect at the origin, then the 2 sets of residuals are “close.”
Some assumptions are needed on the ANOVA $F$ test. Assume that both the response and residual plots look good. It is crucial that there are no outliers. Then a rule of thumb is that if $n - p$ is large, then the ANOVA $F$ test $p$–value is approximately correct. An analogy can be made with the central limit theorem, $\bar{Y}$ is a good estimator for $\mu$ if the $Y_i$ are iid $N(\mu, \sigma^2)$ and also a good estimator for $\mu$ if the data are iid with mean $\mu$ and variance $\sigma^2$ if $n$ is large enough. Also see Theorem 11.25. More on the robustness and lack of robustness of the ANOVA $F$ test can be found in Wilcox (2012).

If all of the $x_i$ are different (no replication) and if the number of predictors $p = n$, then the OLS fit $\hat{Y}_i = Y_i$ and $R^2 = 1$. Notice that Ho is rejected if the statistic $F_o$ is large. More precisely, reject Ho if

$$F_o > F_{p-1,n-p,1-\delta}$$

where

$$P(F \leq F_{p-1,n-p,1-\delta}) = 1 - \delta$$

when $F \sim F_{p-1,n-p}$. Since $R^2$ increases to 1 while $(n - p)/(p - 1)$ decreases to 0 as $p$ increases to $n$, Theorem 2.7a below implies that if $p$ is large then the $F_o$ statistic may be small even if some of the predictors are very good. It is a good idea to use $n \geq 10p$ or at least $n \geq 5p$ if possible. Theorem 11.25 can be used to show that $pval$ is a consistent estimator of the $p$-value under reasonable conditions.
**Theorem 2.7.** Assume that the MLR model has a constant \( \beta_1 \).

a) 
\[
F_o = \frac{MSR}{MSE} = \frac{R^2}{1-R^2} \frac{n-p}{p-1}.
\]

b) If the errors \( e_i \) are iid \( N(0, \sigma^2) \), and if \( H_0: \beta_2 = \cdots = \beta_p = 0 \) is true, then \( F_o \) has an \( F \) distribution with \( p-1 \) numerator and \( n-p \) denominator degrees of freedom: \( F_o \sim F_{p-1,n-p} \).

c) If the errors are iid with mean 0 and variance \( \sigma^2 \), if the error distribution is close to normal, and if \( n-p \) is large enough, and if \( H_0 \) is true, then \( F_o \approx F_{p-1,n-p} \) in that the p-value from the software (pval) is approximately correct.

**Remark 2.7.** When a constant is not contained in the model (i.e., \( x_{i,1} \) is not equal to 1 for all \( i \)), then the computer output still produces an ANOVA table with the test statistic and p-value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now \( H_0: \beta_1 = \cdots = \beta_p = 0 \) Ha: not \( H_0 \), and you are testing whether or not there is an MLR relationship between \( Y \) and \( x_1, \ldots, x_p \). An MLR model without a constant (no intercept) is sometimes called a “regression through the origin.” See Section 2.10.

### 2.5 Prediction

This section gives estimators for predicting a future or new value \( Y_f \) of the response variable given the predictors \( x_f \), and for estimating the mean \( E(Y_f) \equiv E(Y_f|x_f) \). This mean is conditional on the values of the predictors \( x_f \), but the conditioning is often suppressed.

**Warning:** All too often the MLR model seems to fit the “training data” \( (Y_1, x_1), \ldots, (Y_n, x_n) \) well, but when new “test data” is collected, a very different MLR model is needed to fit the new data well. In particular, the MLR model seems to fit the data \( (Y_i, x_i) \) well for \( i = 1, \ldots, n \), but when the researcher tries to predict \( Y_f \) for a new vector of predictors \( x_f \), the prediction is very poor in that \( \hat{Y}_f \) is not close to the \( Y_f \) actually observed. **Wait until after the MLR model has been shown to make good predictions before claiming that the model gives good predictions!**

There are several reasons why the MLR model may not fit new data well.  
i) The model building process is usually iterative. Data \( Z, w_1, \ldots, w_r \) is collected. If the model is not linear, then functions of \( Z \) are used as a potential response variable and functions of the \( w_i \) as potential predictors. After trial and error, the functions are chosen, resulting in a final MLR model using \( Y \) and \( x_1, \ldots, x_p \). Since the same data set was used during the model building...
process, biases are introduced and the MLR model fits the “training data” better than it fits new test data. Suppose that $Y, x_1, \ldots, x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

ii) If $(Y_f, x_f)$ come from a different population than the population of $(Y_1, x_1), \ldots, (Y_n, x_n)$, then prediction for $Y_f$ can be arbitrarily bad.

iii) Even a good MLR model may not provide good predictions for an $x_f$ that is far from the $x_i$ (extrapolation).

iv) The MLR model may be missing important predictors (underfitting).

v) The MLR model may contain unnecessary predictors (overfitting).

Three remedies for i) are a) use previously published studies to select an MLR model before gathering data. Unfortunately, often the population from a published study is quite different from the population of the data collected by the researcher. Then the MLR model from the published study is not adequate. b) Do a trial study. Collect some data, build an MLR model using the iterative process. Then use this model as the prespecified model and collect data for the main part of the study. Better yet, do a trial study, specify a model, collect more trial data, improve the specified model, and repeat until the latest specified model works well. Unfortunately, trial studies are often too expensive or not possible because the data is difficult to collect. c) If the data set is large enough, use a “training set” of a random sample of $k$ of the $n$ cases to build a model where $10p \leq n/2 \leq k \leq 0.9n$. Then use “validation set” of the other $n - k$ cases to confirm that the model built with the “training set” is good. This technique may help reduce biases, but needs $n \geq 20p$. See James et al. (2013, pp. 176–178). In particular, build the model with the training set, then check the asymptotically optimal prediction interval (2.20), derived later in this section, on the validation set.

**Definition 2.18.** Consider the MLR model $Y = X\beta + e$ and the hat matrix $H = X(X^T X)^{-1}X^T$. Let $h_i = h_{ii}$ be the $i$th diagonal element of $H$ for $i = 1, \ldots, n$. Then $h_i$ is called the $i$th **leverage** and $h_i = x_i^T(X^T X)^{-1}x_i$. Suppose new data is to be collected with predictor vector $x_f$. Then the leverage of $x_f$ is $h_f = x_f^T(X^T X)^{-1}x_f$. **Extrapolation** occurs if $x_f$ is far from the $x_1, \ldots, x_n$.

**Rule of thumb 2.5.** Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, \ldots, h_n)$. This rule works best if the predictors are linearly related in that a plot of $x_i$ versus $x_j$ should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $x_f$ could be far from the $x_i$ but still have $h_f < \max(h_1, \ldots, h_n)$. 
Example 2.7. Consider predicting $Y = \text{weight}$ from $x = \text{height}$ and a constant from data collected on men between 18 and 24 where the minimum height was 57 and the maximum height was 79 inches. The OLS equation was $\hat{Y} = -167 + 4.7x$. If $x = 70$, then $\hat{Y} = -167 + 4.7(70) = 162$ pounds. If $x = 1$ inch, then $\hat{Y} = -167 + 4.7(1) = -162.3$ pounds. It is impossible to have negative weight, but it is also impossible to find a 1 inch man. This MLR model should not be used for $x$ far from the interval (57, 79).

Definition 2.19. Consider the unimodal MLR model $Y = x^T \beta + e$ where $E(e) = 0$. Then regression function is the hyperplane $E(Y) \equiv E(Y|x) = x_1 \beta_1 + x_2 \beta_2 + \cdots + x_p \beta_p = x^T \beta$. (2.9)

Assume OLS is used to find $\hat{\beta}$. Then the point estimator of $Y_f$ given $x = x_f$ is

$$\hat{Y}_f = x_{f,1} \hat{\beta}_1 + \cdots + x_{f,p} \hat{\beta}_p = x_f^T \hat{\beta}. \quad (2.10)$$

The point estimator of $E(Y_f) \equiv E(Y_f|x_f)$ given $x = x_f$ is also $\hat{Y}_f = x_f^T \hat{\beta}$. Assume that the MLR model contains a constant $\beta_1$ so that $x_1 \equiv 1$. The large sample 100 $(1 - \delta)$% confidence interval (CI) for $E(Y_f|x_f) = x_f^T \beta = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2} \text{se}(\hat{Y}_f) \quad (2.11)$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if $T$ has a $t$ distribution with $n - p$ degrees of freedom. Generally $\text{se}(\hat{Y}_f)$ will come from output, but

$$\text{se}(\hat{Y}_f) = \sqrt{\text{MSE} h_f} = \sqrt{\text{MSE} x_f^T (X^TX)^{-1} x_f}. $$

Recall the interpretation of a 100 $(1 - \delta)$% CI for a parameter $\mu$ is that if you collect data then form the CI, and repeat for a total of $k$ times where the $k$ trials are independent from the same population, then the probability that $m$ of the CIs will contain $\mu$ follows a binomial($k, \rho = 1 - \delta$) distribution. Hence if 100 95% CIs are made, $\rho = 0.95$ and about 95 of the CIs will contain $\mu$ while about 5 will not. Any given CI may (good sample) or may not (bad sample) contain $\mu$, but the probability of a “bad sample” is $\delta$.

The following theorem is analogous to the central limit theorem and the theory for the $t$–interval for $\mu$ based on $\bar{Y}$ and the sample standard deviation (SD) $S_Y$. If the data $Y_1, \ldots, Y_n$ are iid with mean 0 and variance $\sigma^2$, then $\bar{Y}$ is asymptotically normal and the $t$–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\beta}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $X$ or if the $x_i$ are iid random vectors from a well-behaved population. Outliers can cause the condition to fail. Convergence in distribution, $Z_n \overset{D}{\to} N_p(0, \Sigma)$, means the multivariate normal approximation can be used for probability
calculations involving $Z_n$. When $p = 1$, the univariate normal distribution can be used. See Sen and Singer (1993, p. 280) for the theorem, which implies that \( \hat{\beta} \approx N_p(\beta, \sigma^2 (X^T X)^{-1}) \). See Chapter 10 for the multivariate normal distribution.

**Theorem 2.8, LS CLT (Least Squares Central Limit Theorem):**
Consider the MLR model $Y_i = x_i^T \beta + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, \ldots, h_n) \to 0$ in probability as $n \to \infty$ and

$$\frac{X^T X}{n} \to W^{-1}$$

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\beta}$ satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N_p(0, \sigma^2 W).$$

(2.12)

Equivalently,

$$(X^T X)^{1/2}(\hat{\beta} - \beta) \xrightarrow{D} N_p(0, \sigma^2 I_p).$$

(2.13)

**Definition 2.20.** A large sample $100(1 - \delta)\%$ prediction interval (PI) has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable $Y_f$ is independent of $Y_1, \ldots, Y_n$. Then the $100(1 - \delta)\%$ PI for $Y_f$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2} se(ped)$$

(2.14)

where $P(T \leq t_{n-p,\delta}) = \delta$ if $T$ has a $t$ distribution with $n - p$ degrees of freedom. Generally $se(ped)$ will come from output, but

$$se(ped) = \sqrt{MSE(1 + h_f)}.$$

The interpretation of a $100(1 - \delta)\%$ PI for a random variable $Y_f$ is similar to that of a CI. Collect data, then form the PI, and repeat for a total of $k$ times where $k$ trials are independent from the same population. If $Y_{fi}$ is the $i$th random variable and $PI_i$ is the $i$th PI, then the probability that $Y_{fi} \in PI_i$ for $m$ of the PIs follows a binomial($k, \rho = 1 - \delta$) distribution. Hence if 100 $95\%$ PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size $n$ goes to $\infty$ while the length of the PI converges to some nonzero number $J$, say. Secondly, the CI for $E(Y_f|x_f)$ given in Definition 2.19 tends to work well for the unimodal MLR model if the sample size is large while the PI in Definition 2.20 is made under the assumption that the $e_i$ are iid $N(0, \sigma^2)$ and may not perform well if the normality assumption is violated.
To see this, consider \( \mathbf{x}_f \) such that the heights \( Y \) of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for \( E(Y|\mathbf{x}_f) \) should be centered at about 66 and the length should go to zero as \( n \) gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval 66 ± 1.96(3). This result follows because if \( Y \sim N(66, 9) \) then \( P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) \approx 0.025 \). In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Assume that the unimodal MLR model is valid so that \( e \) is from some distribution with 0 mean and variance \( \sigma^2 \). Olive (2007) shows that if \( 1 - \gamma \) is the asymptotic coverage of the classical nominal \( (1 - \delta)100\% \) PI (2.14), then

\[
1 - \gamma = P(-\sigma z_{1-\delta/2} < e < \sigma z_{1-\delta/2}) \geq 1 - \frac{1}{z_{1-\delta/2}^2} \tag{2.15}
\]

where the inequality follows from Chebyshev’s inequality. Hence the asymptotic coverage of the nominal 95% PI is at least 73.9%. The 95% PI (2.14) was often quite accurate in that the asymptotic coverage was close to 95% for a wide variety of error distributions. The 99% and 90% PIs did not perform as well.

Let \( \xi_\delta \) be the \( \delta \) percentile of the error \( e \), i.e. \( P(e \leq \xi_\delta) = \delta \). Let \( \hat{\xi}_\delta \) be the sample \( \delta \) percentile of the residuals. Then the results from Theorem 2.8 suggest that the residuals \( r_i \) estimate the errors \( e_i \), and that the sample percentiles of the residuals \( \hat{\xi}_\delta \) estimate \( \xi_\delta \). For many error distributions,

\[
E(MSE) = E \left( \sum_{i=1}^{n} \frac{r_i^2}{n-p} \right) = \sigma^2 = E \left( \sum_{i=1}^{n} \frac{e_i^2}{n} \right).
\]

This result suggests that

\[
\sqrt{\frac{n}{n-p}} \ r_i \approx e_i.
\]

Using

\[
a_n = \left( 1 + \frac{15}{n} \right) \sqrt{\frac{n}{n-p}} \sqrt{(1 + h_f)}, \tag{2.16}
\]

a large sample semiparametric 100(1 − \( \delta \))% PI for \( Y_f \) is

\[
[\hat{Y}_f + a_n \hat{\xi}_{\delta/2}, \hat{Y}_f + a_n \hat{\xi}_{1-\delta/2}]. \tag{2.17}
\]

This PI is very similar to the classical PI except that \( \hat{\xi}_\delta \) is used instead of \( \sigma z_{\delta} \) to estimate the error percentiles \( \xi_\delta \). The large sample coverage \( 1 - \gamma \) of this nominal 100(1 − \( \delta \))% PI is asymptotically correct: \( 1 - \gamma = 1 - \delta \).
Example 2.8. For the Buxton (1920) data suppose that the response $Y = \text{height}$ and the predictors were a constant, head length, nasal height, bigausal breadth, and cephalic index. Five outliers were deleted leaving 82 cases. Figure 2.3 shows a response plot of the fitted values versus the response $Y$ with the identity line added as a visual aid. The plot suggests that the model is good since the plotted points scatter about the identity line in an evenly populated band although the relationship is rather weak since the correlation of the plotted points is not very high. The triangles represent the upper and lower limits of the semiparametric 95% PI (2.17). For this example, 79 (or 96%) of the $Y_i$ fell within their corresponding PI while 3 $Y_i$ did not. A plot using the classical PI (2.14) would be very similar for this data. The plot was made with the following $R$ commands, using the \texttt{lregpack} function \texttt{piplot}.

```r
x <- buxx[-c(61,62,63,64,65),]
Y <- buxy[-c(61,62,63,64,65)]
piplot(x,Y)
```

![Fig. 2.3 95% PI Limits for Buxton Data](image)

Given output showing $\hat{\beta}_i$ and given $x_f$, se(pred), and se($\hat{Y}_f$), Example 2.9 shows how to find $\hat{Y}_f$, a CI for $E(Y_f|x_f)$, and the classical PI (2.14) for $Y_f$. 
Below is shown typical output in symbols. Sometimes “Label” is replaced by “Predictor” and “Estimate” by “coef” or “Coefficients.”

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_1$</td>
<td>$se(\hat{\beta}_1)$</td>
<td>$t_{o,1}$</td>
<td>for $Ho: \beta_1 = 0$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\hat{\beta}_2$</td>
<td>$se(\hat{\beta}_2)$</td>
<td>$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$</td>
<td>for $Ho: \beta_2 = 0$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$x_p$</td>
<td>$\hat{\beta}_p$</td>
<td>$se(\hat{\beta}_p)$</td>
<td>$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$</td>
<td>for $Ho: \beta_p = 0$</td>
</tr>
</tbody>
</table>

**Example 2.9.** The Rouncefield (1995) data povc.1sp are female and male life expectancies from $n = 91$ countries where 6 cases with missing GNP were deleted. Suppose that it is desired to predict female life expectancy $Y$ from male life expectancy $X$. Suppose that if $X_f = 60$, then $se(\text{pred}) = 2.1285$, and $se(\hat{Y}_f) = 0.2241$. Below is some output.

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$-2.93739$</td>
<td>$1.42523$</td>
<td>$-2.061$</td>
<td>$0.0422$</td>
</tr>
<tr>
<td>$mlife$</td>
<td>$1.12359$</td>
<td>$0.0229362$</td>
<td>$48.988$</td>
<td>$0.0000$</td>
</tr>
</tbody>
</table>

a) Find $\hat{Y}_f$ if $X_f = 60$.

Solution: In this example, $x_f = (1, X_f)^T$ since a constant is in the output above. Thus $\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f = -2.93739 + 1.12359(60) = 64.478$.

b) If $X_f = 60$, find a 90% confidence interval for $E(Y) \equiv E(Y_f|x_f)$.

Solution: The CI is $\hat{Y}_f \pm t_{n-2,1-\delta/2}se(\hat{Y}_f) = 64.478 \pm 1.645(0.2241) = 64.478 \pm 0.3686 = [64.1094, 64.8466]$. To use the $t$-table on the last page of Chapter 14, use the 2nd to last row marked by $Z$ since $d = df = n - 2 = 89 > 30$. In the last row find CI = 90% and intersect the 90% column and the Z row to get the value of $t_{89,0.95} \approx z_{95} = 1.645$.

c) If $X_f = 60$, find a 90% prediction interval for $Y_f$.

Solution: The PI is $\hat{Y}_f \pm t_{n-2,1-\delta/2}se(pred) = 64.478 \pm 1.645(2.1285) = 64.478 \pm 3.5014 = [60.9766, 67.9794]$.

Two more PIs will be defined and then the 4 PIs (2.14), (2.17), (2.18), and (2.20) will be compared via simulation. An asymptotically conservative (ac) 100$(1 - \delta)%$ PI has asymptotic coverage $1 - \gamma \geq 1 - \delta$. We used the (ac) 100$(1 - \delta)%$ PI

$$\hat{Y}_f \pm \sqrt{\frac{n}{n-p}} \max(|\hat{\xi}_{\delta/2}|, |\hat{\xi}_{1-\delta/2}|)\sqrt{1+h_f} \quad (2.18)$$

which has asymptotic coverage.
2.5 Prediction

\[ 1 - \gamma = P\left[ -\max(|\xi_{\delta}/2|, |\xi_{1-\delta}/2|) < e < \max(|\xi_{\delta}/2|, |\xi_{1-\delta}/2|) \right]. \] \hspace{1cm} (2.19)

Notice that \( 1 - \delta \leq 1 - \gamma \leq 1 - \delta/2 \) and \( 1 - \gamma = 1 - \delta \) if the error distribution is symmetric.

In the simulations described below, \( \hat{\xi}_{\delta} \) will be the sample percentile for the PIs (2.17) and (2.18). A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where \( c = \lceil n(1 - \delta) \rceil \) and \( \lceil x \rceil \) is the smallest integer \( \geq x \), e.g., \( \lceil 7.7 \rceil = 8 \). That is, let \( r(1), \ldots, r(n) \) be the order statistics of the residuals. Compute \( r(c) - r(1), r(c+1) - r(2), \ldots, r(n) - r(n-c+1) \). Let \( [r(d), r(d+c-1)] = [\hat{\xi}_{\delta_1}, \hat{\xi}_{1-\delta_2}] \) correspond to the interval with the smallest distance. Then the large sample asymptotically optimal 100 (1 - \( \delta \))\% PI for \( Y_f \) is

\[ [\hat{Y}_f + a_n \hat{\xi}_{\delta_1}, \hat{Y}_f + a_n \hat{\xi}_{1-\delta_2}] \] \hspace{1cm} (2.20)

where \( a_n \) is given by (2.16).

**Remark 2.8.** We recommend using the asymptotically optimal PI (2.20) instead of the classical PI (2.14). The *lregpack* function *pisim* can be used to recreate the simulation described below. See Problem 2.29.

A small simulation study compares the PI lengths and coverages for sample sizes \( n = 50, 100 \), and 1000 for several error distributions. The value \( n = \infty \) gives the asymptotic coverages and lengths. The MLR model with \( E(Y_i) = 1 + x_{i2} + \cdots + x_{i8} \) was used. The vectors \( (x_2, \ldots, x_8)^T \) were iid \( N_7(0, I_7) \). The error distributions were N(0,1), \( t_3 \), and exponential(1) -1. Also, a small sensitivity study to examine the effects of changing \( (1 + 15/n) \) to \( (1 + k/n) \) on the 99% PIs (2.17) and (2.20) was performed. For \( n = 50 \) and \( k \) between 10 and 20, the coverage increased by roughly 0.001 as \( k \) increased by 1.

**Table 2.1** N(0,1) Errors.

<table>
<thead>
<tr>
<th>( \delta )</th>
<th>n</th>
<th>clen</th>
<th>slen</th>
<th>alen</th>
<th>olen</th>
<th>ccov</th>
<th>scov</th>
<th>acov</th>
<th>ocov</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>50</td>
<td>5.860</td>
<td>6.172</td>
<td>5.191</td>
<td>6.448</td>
<td>.989</td>
<td>.988</td>
<td>.972</td>
<td>.990</td>
</tr>
<tr>
<td>0.01</td>
<td>100</td>
<td>5.470</td>
<td>5.625</td>
<td>5.257</td>
<td>5.412</td>
<td>.990</td>
<td>.988</td>
<td>.985</td>
<td>.985</td>
</tr>
<tr>
<td>0.01</td>
<td>1000</td>
<td>5.182</td>
<td>5.181</td>
<td>5.263</td>
<td>5.097</td>
<td>.992</td>
<td>.993</td>
<td>.994</td>
<td>.992</td>
</tr>
<tr>
<td>0.01</td>
<td>( \infty )</td>
<td>5.152</td>
<td>5.152</td>
<td>5.152</td>
<td>5.152</td>
<td>.990</td>
<td>.990</td>
<td>.990</td>
<td>.990</td>
</tr>
<tr>
<td>0.05</td>
<td>50</td>
<td>4.379</td>
<td>5.167</td>
<td>4.290</td>
<td>5.111</td>
<td>.948</td>
<td>.974</td>
<td>.940</td>
<td>.968</td>
</tr>
<tr>
<td>0.05</td>
<td>100</td>
<td>4.136</td>
<td>4.531</td>
<td>4.172</td>
<td>4.359</td>
<td>.956</td>
<td>.970</td>
<td>.956</td>
<td>.958</td>
</tr>
<tr>
<td>0.05</td>
<td>1000</td>
<td>3.938</td>
<td>3.977</td>
<td>4.001</td>
<td>3.927</td>
<td>.952</td>
<td>.952</td>
<td>.954</td>
<td>.948</td>
</tr>
<tr>
<td>0.05</td>
<td>( \infty )</td>
<td>3.920</td>
<td>3.920</td>
<td>3.920</td>
<td>3.920</td>
<td>.950</td>
<td>.950</td>
<td>.950</td>
<td>.950</td>
</tr>
<tr>
<td>0.1</td>
<td>50</td>
<td>3.642</td>
<td>4.445</td>
<td>3.658</td>
<td>4.193</td>
<td>.894</td>
<td>.945</td>
<td>.895</td>
<td>.929</td>
</tr>
<tr>
<td>0.1</td>
<td>100</td>
<td>3.455</td>
<td>3.841</td>
<td>3.519</td>
<td>3.690</td>
<td>.900</td>
<td>.930</td>
<td>.905</td>
<td>.913</td>
</tr>
<tr>
<td>0.1</td>
<td>1000</td>
<td>3.304</td>
<td>3.343</td>
<td>3.352</td>
<td>3.304</td>
<td>.901</td>
<td>.903</td>
<td>.907</td>
<td>.901</td>
</tr>
<tr>
<td>0.1</td>
<td>( \infty )</td>
<td>3.290</td>
<td>3.290</td>
<td>3.290</td>
<td>3.290</td>
<td>.900</td>
<td>.900</td>
<td>.900</td>
<td>.900</td>
</tr>
</tbody>
</table>
The simulation compared coverages and lengths of the classical (2.14), semiparametric (2.17), asymptotically conservative (2.18), and asymptotically optimal (2.20) PIs. The latter 3 intervals are asymptotically optimal for symmetric unimodal error distributions in that they have the shortest asymptotic length that gives the desired asymptotic coverage. The PIs (2.17) and (2.20) also give the correct asymptotic coverage if the unimodal errors are not symmetric, while the PI (2.18) gives higher coverage (is conservative). The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where $Y_f$ fell within the nominal $100(1 - \delta)\%$ PI. The count $m\hat{p}$ has a binomial($m = 5000, p = 1 - \gamma_n$) distribution where $1 - \gamma_n$ converges to the asymptotic coverage $(1 - \gamma)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0014, 0.0031, \text{and } 0.0042$ for $p = 0.01, 0.05 \text{ and } 0.1$, respectively. Hence an observed coverage $\hat{p} \in (0.986, 0.994)$ for 99%, $\hat{p} \in (0.941, 0.959)$ for 95%, and $\hat{p} \in (0.887, 0.913)$ for 90% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.
Tables 2.1–2.3 show the results of the simulations for the 3 error distributions. The letters c, s, a, and o refer to intervals (2.14), (2.17), (2.18), and (2.20), respectively. For the normal errors, the coverages were about right and the semiparametric interval tended to be rather long for \( n = 50 \) and 100. The classical PI asymptotic coverage \( 1 - \gamma \) tended to be fairly close to the nominal coverage \( 1 - \delta \) for all 3 distributions and \( \delta = 0.01, 0.05, \) and \( 0.1 \). The asymptotically optimal PI tended to have short length and simulated coverage close to the nominal coverage.

### 2.6 The Partial F Test

Suppose that there is data on variables \( Z, w_1, \ldots, w_r \) and that a useful MLR model has been made using \( Y = t(Z), x_1 \equiv 1, x_2, \ldots, x_p \) where each \( x_i \) is some function of \( w_1, \ldots, w_r \). This useful model will be called the full model. It is important to realize that the full model does not need to use every variable \( w_j \) that was collected. For example, variables with outliers or missing values may not be used. Forming a useful full model is often very difficult, and it is often not reasonable to assume that the candidate full model is good based on a single data set, especially if the model is to be used for prediction.

Even if the full model is useful, the investigator will often be interested in checking whether a model that uses fewer predictors will work just as well. For example, perhaps \( x_p \) is a very expensive predictor but is not needed given that \( x_1, \ldots, x_{p-1} \) are in the model. Also a model with fewer predictors tends to be easier to understand.

**Definition 2.21.** Let the **full model** use \( Y, x_1 \equiv 1, x_2, \ldots, x_p \) and let the **reduced model** use \( Y, x_1, x_{i_2}, \ldots, x_{i_q} \) where \( \{i_2, \ldots, i_q\} \subset \{2, \ldots, p\} \).

The partial \( F \) test is used to test whether the reduced model is good in that it can be used instead of the full model. It is crucial that the reduced model be selected before looking at the data. If the reduced model is selected after looking at output and discarding the worst variables, then the \( p \)-value for the partial \( F \) test will be too high. For (ordinary) least squares, usually a constant is used, and we are assuming that both the full model and the reduced model contain a constant. The partial \( F \) test has null hypothesis \( H_0 : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0 \), and alternative hypothesis \( H_A : \) at least one of the \( \beta_{i_j} \neq 0 \) for \( j > q \). The null hypothesis is equivalent to \( H_0 : \) “the reduced model is good.” Since only the full model and reduced model are being compared, the alternative hypothesis is equivalent to \( H_A : \) “the reduced model is not as good as the full model, so use the full model,” or more simply, \( H_A : \) “use the full model.”

To perform the partial \( F \) test, fit the full model and the reduced model and obtain the ANOVA table for each model. The quantities \( df_F, \) SSE(F)
and MSE(F) are for the full model and the corresponding quantities from the reduced model use an $R$ instead of an $F$. Hence SSE(F) and SSE(R) are the residual sums of squares for the full and reduced models, respectively. Shown below is output only using symbols.

Full model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Fo and p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$p - 1$</td>
<td>SSR</td>
<td>MSR</td>
<td>Fo = MSR/MSE</td>
</tr>
<tr>
<td>Residual</td>
<td>$df_F = n - p$</td>
<td>SSE(F)</td>
<td>MSE(F)</td>
<td>for Ho: $\beta_2 = \cdots = \beta_p = 0$</td>
</tr>
</tbody>
</table>

Reduced model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Fo and p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$q - 1$</td>
<td>SSR</td>
<td>MSR</td>
<td>Fo = MSR/MSE</td>
</tr>
<tr>
<td>Residual</td>
<td>$df_R = n - q$</td>
<td>SSE(R)</td>
<td>MSE(R)</td>
<td>for Ho: $\beta_2 = \cdots = \beta_q = 0$</td>
</tr>
</tbody>
</table>

**Be able to perform the 4 step partial F test of hypotheses.**

i) State the hypotheses. Ho: the reduced model is good Ha: use the full model

ii) Find the test statistic. $F_R = \frac{\left[ SSE(R) - SSE(F) \right]}{df_R - df_F} / MSE(F)$

iii) Find the pval = P($F_{df_R - df_F, df_F} > F_R$). (On exams typically an $F$ table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$, while pval is the estimated $p$-value.)

iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Sometimes software has a shortcut. In particular, the $R$ software uses the `anova` command. See Problem 2.27. As an example, assume that the full model uses $x_2$ and $x_3$ while the reduced model uses $x_2$. Both models contain a constant. Then the following commands will perform the partial $F$ test. (On the computer screen the second command looks more like red $\leftarrow$ `lm(y~x2).`)

```r
full <- lm(y~x2+x3)
red <- lm(y~x2)
anova(red,full)
```

For an $n \times 1$ vector $a$, let

$$\|a\| = \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{a^T a}$$
be the Euclidean norm of $a$. If $r$ and $r_R$ are the vector of residuals from the full and reduced models, respectively, notice that $SSE(F) = \|r\|^2$ and $SSE(R) = \|r_R\|^2$.

The following proposition suggests that $H_0$ is rejected in the partial $F$ test if the change in residual sum of squares $SSE(R) - SSE(F)$ is large compared to $SSE(F)$. If the change is small, then $F_R$ is small and the test suggests that the reduced model can be used.

**Proposition 2.6.** Let $R^2$ and $R^2_R$ be the multiple coefficients of determination for the full and reduced models, respectively. Let $\hat{Y}$ and $\hat{Y}_R$ be the vectors of fitted values for the full and reduced models, respectively. Then the test statistic in the partial $F$ test is

$$F_R = \left[ \frac{SSE(R) - SSE(F)}{df_R - df_F} \right] /MSE(F) =$$

$$\left[ \frac{\|\hat{Y}\|^2 - \|\hat{Y}_R\|^2}{df_R - df_F} \right] /MSE(F) =$$

$$\frac{SSE(R) - SSE(F)}{SSE(F)} \frac{n - p}{p - q} = \frac{R^2 - R^2_R}{1 - R^2} \frac{n - p}{p - q}.$$

**Definition 2.22.** An **FF plot** is a plot of fitted values from 2 different models or fitting methods.

Six plots are useful diagnostics for the partial $F$ test: the RR plot with the full model residuals on the vertical axis and the reduced model residuals on the horizontal axis, the FF plot with the full model fitted values on the vertical axis, and always make the response and residual plots for the full and reduced models. Suppose that the full model is a useful MLR model. If the reduced model is good, then the response plots from the full and reduced models should be very similar, visually. Similarly, the residual plots (of the fitted values versus the residuals) from the full and reduced models should be very similar, visually. Finally, the correlation of the plotted points in the RR and FF plots should be high, $\geq 0.95$, say, and the plotted points in the RR and FF plots should cluster tightly about the identity line. Add the identity line to both the RR and FF plots as a visual aid. Also add the OLS line from regressing $r$ on $r_R$ to the RR plot (the OLS line is the identity line in the FF plot). If the reduced model is good, then the OLS line should nearly coincide with the identity line in that it should be difficult to see that the two lines intersect at the origin, as in Figure 2.2. If the FF plot looks good but the RR plot does not, the reduced model may be good if the main goal of the analysis is to predict $Y$. 

In Chapter 3, Example 3.8 describes the Gladstone (1905) data. Let the reduced model use a constant, \((\text{size})^{1/3}\), \(\text{sex}\), and \(\text{age}\). Then Figure 3.7 shows the response and residual plots for the full and reduced models, and Figure 3.9 shows the RR and FF plots.

**Example 2.10.** For the Buxton (1920) data, \(n = 76\) after 5 outliers and 6 cases with missing values are removed. Assume that the response variable \(Y\) is height, and the explanatory variables are \(x_2 = \text{bigonal breadth}, x_3 = \text{cephalic index}, x_4 = \text{finger to ground}, x_5 = \text{head length}, x_6 = \text{nasal height},\) and \(x_7 = \text{sternal height}.\) Suppose that the full model uses all 6 predictors plus a constant \((x_1)\) while the reduced model uses the constant, cephalic index, and finger to ground. Test whether the reduced model can be used instead of the full model using the output below.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>6</td>
<td>260467</td>
<td>43411.1</td>
<td>87.41</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>69</td>
<td>34267.4</td>
<td>496.629</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>94110.5</td>
<td>47055.3</td>
<td>17.12</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>73</td>
<td>200623.</td>
<td>2748.27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Solution: The 4 step partial F test is shown below.

i) Ho: the reduced model is good Ha: use the full model

\[ F_R = \left( \frac{SSE(R) - SSE(F)}{df_R - df_F} \right) / MSE(F) \]

\[ = \frac{41588.9 / 496.629}{34267.4 / 496.629} = 83.742. \]

iii) pval = \(P(F_{4,69} > 83.742) = 0.00.\)

iv) The pval < \(\delta\) (= 0.05, since \(\delta\) was not given), so reject Ho. The full model should be used instead of the reduced model. (Bigonal breadth, head length, nasal height, and sternal height are needed in the MLR for height given that cephalic index and finger to ground are in the model.)

Using a computer to get the pval makes sense, but for exams you may need to use a table. In ARC, you can use the Calculate probability option from the ARC menu, enter 83.742 as the value of the statistic, 4 and 69 as the degrees of freedom, and select the F distribution. To use the table near the end of Chapter 14, use the bottom row since the denominator degrees of freedom 69 > 30. Intersect with the column corresponding to \(k = 4\) numerator degrees of freedom. The cutoff value is 2.37. If the \(F_R\) statistic was 2.37, then the pval would be 0.05. Since 83.472 > 2.37, the pval < 0.05, and since 83.472 >> 2.37, we can say that the pval ≈ 0.0.
The Wald t Test

Example 2.11. Now assume that the reduced model uses the constant, sternal height, finger to ground, and head length. Using the output below, test whether the reduced model is good.

Summary Analysis of Variance Table for Reduced Model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>259704</td>
<td>86568</td>
<td>177.93</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>72</td>
<td>35030.1</td>
<td>486.528</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Solution: The 4 step partial F test follows.

i) Ho: the reduced model is good  Ha: use the full model

\[ F_R = \frac{\left[ SSE(R) - SSE(F) \right]}{df_R - df_F} / MSE(F) = \frac{\left[ 35030.1 - 34267.4 \right]}{72 - 69} / 496.629 \]

\[ = \frac{254.2333}{496.629} = 0.512. \]

ii) The pval = \( P(F_{3,69} > 0.512) = 0.675. \)

iii) The pval < \( \delta \), so reject fail to reject Ho. The reduced model is good.

To use the F table near the end of Chapter 14, use the bottom row since the denominator degrees of freedom 69 > 30. Intersect with the column corresponding to \( k = 3 \) numerator degrees of freedom. The cutoff value is 2.61. Since 0.512 < 2.61, pval > 0.05, and this is enough information to fail to reject Ho.

Some R commands and output to do the above problem are shown below.

```r
library(ggplot2)
cyp <- matrix(scan(), nrow=76, ncol=8, byrow=T, dimnames=list(c("indx", "ht", "sternal", "finger", "hdlen", "nasal", "bigonal", "cephalic")))
cyp <- cyp[,1:8]
cyp <- as.data.frame(cyp)
full <- lm(ht ~ ., data=cyp)
red <- lm(ht ~ sternal + finger + hdlen, data=cyp)
anova(red, full)
Model 1: ht ~ sternal + Finger + hdlen
Model 2: ht ~ sternal + Finger + hdlen + nasal + bigonal + cephalic
Res.Df RSS Df Sum of Sq F Pr(>F)
1 72 35030
2 69 34267 3 762.67 0.5119 0.675
```

2.7 The Wald t Test

Often investigators hope to examine \( \beta_k \) in order to determine the importance of the predictor \( x_k \) in the model; however, \( \beta_k \) is the coefficient for \( x_k \) given that the other predictors are in the model. Hence \( \beta_k \) depends strongly on the other predictors in the model. Suppose that the model has an intercept:
The predictor \( x_k \) is highly correlated with the other predictors if the OLS regression of \( x_k \) on \( x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p \) has a high coefficient of determination \( R^2_k \). If this is the case, then often \( x_k \) is not needed in the model given that the other predictors are in the model. If at least one \( R^2_k \) is high for \( k \geq 2 \), then there is multicollinearity among the predictors.

As an example, suppose that \( Y = \text{height} \), \( x_1 \equiv 1 \), \( x_2 = \text{left leg length} \), and \( x_3 = \text{right leg length} \). Then \( x_2 \) should not be needed given \( x_3 \) is in the model and \( \beta_2 = 0 \) is reasonable. Similarly \( \beta_3 = 0 \) is reasonable. On the other hand, if the model only contains \( x_1 \) and \( x_2 \), then \( x_2 \) is extremely important with \( \beta_2 \) near 2. If the model contains \( x_1, x_2, x_3, x_4 = \text{height at shoulder}, x_5 = \text{right arm length}, x_6 = \text{head length}, \) and \( x_7 = \text{length of back} \), then \( R^2_i \) may be high for each \( i \geq 2 \). Hence \( x_i \) is not needed in the MLR model for \( Y \) given that the other predictors are in the model.

**Definition 2.23.** The 100 \((1 - \delta)\%\) CI for \( \beta_k \) is \( \hat{\beta}_k \pm t_{n-p,1-\delta/2} se(\hat{\beta}_k) \). If the degrees of freedom \( d = n - p \geq 30 \), the N(0,1) cutoff \( z_{1-\delta/2} \) may be used.

**Know how to do** the 4 step Wald \( t \)-test of hypotheses.

i) State the hypotheses \( H_0: \beta_k = 0 \) \( H_a: \beta_k \neq 0 \).
ii) Find the test statistic \( t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k) \) or obtain it from output.
iii) Find \( pval \) from output or use the \( t \)-table: \( pval = 2P(t_{n-p} < -|t_{o,k}|) = 2P(t_{n-p} > |t_{o,k}|) \).

Use the normal table or the \( d = Z \) line in the \( t \)-table if the degrees of freedom \( d = n - p \geq 30 \). Again \( pval \) is the estimated \( p \)-value.
iv) State whether you reject \( H_0 \) or fail to reject \( H_0 \) and give a nontechnical sentence restating your conclusion in terms of the story problem.

Recall that \( H_0 \) is rejected if the \( pval \leq \delta \). As a benchmark for this textbook, use \( \delta = 0.05 \) if \( \delta \) is not given. If \( H_0 \) is rejected, then conclude that \( x_k \) is needed in the MLR model for \( Y \) given that the other predictors are in the model. If you fail to reject \( H_0 \), then conclude that \( x_k \) is not needed in the MLR model for \( Y \) given that the other predictors are in the model. (Or there is not enough evidence to conclude that \( x_k \) is needed in the MLR model given that the other predictors are in the model.) Note that \( x_k \) could be a very useful individual predictor, but may not be needed if other predictors are added to the model. It is better to use the output to get the test statistic and \( pval \) than to use formulas and the \( t \)-table, but exams may not give the relevant output.

**Definition 2.24.** Assume that there is a constant \( x_1 \equiv 1 \) in the model, and let \( \mathbf{x}_{(k)} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)^T \) be the vector of predictors with the \( k \)th predictor \( x_k \) deleted. Let \( \mathbf{r}_{(k)} \) be the residuals from regressing \( Y \) on \( \mathbf{x}_{(k)} \), that is, on all of the predictor variables except \( x_k \). Let \( \mathbf{r}(x_k|\mathbf{x}_{(k)}) \) denote the residuals from regressing \( x_k \) on \( \mathbf{x}_{(k)} \). Then an added variable plot for \( x_k \) is a plot of \( \mathbf{r}(x_k|\mathbf{x}_{(k)}) \) versus \( \mathbf{r}_{(k)} \) for \( k = 2, \ldots, p \).
The added variable plot (also called a partial regression plot) is used to give information about the test $H_0 : \beta_k = 0$. The points in the plot cluster about a line through the origin with slope $= \hat{\beta}_k$. An interesting fact is that the residuals from this line, i.e. the residuals from regressing $r_{(k)}$ on $r(x_k|x_{(k)})$, are exactly the same as the usual residuals from regressing $Y$ on $x$. The range of the horizontal axis gives information about the collinearity of $x_k$ with the other predictors. Small range implies that $x_k$ is well explained by the other predictors. The $r(x_k|x_{(k)})$ represent the part of $x_k$ that is not explained by the remaining variables while the $r_{(k)}$ represent the part of $Y$ that is not explained by the remaining variables.

An added variable plot with a clearly nonzero slope and tight clustering about a line implies that $x_k$ is needed in the MLR for $Y$ given that the other predictors $x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p$ are in the model. Slope near zero in the added variable plot implies that $x_k$ may not be needed in the MLR for $Y$ given that all other predictors $x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p$ are in the model.

If the zero line with 0 slope and 0 intercept and the OLS line are added to the added variable plot, the variable is probably needed if it is clear that the two lines intersect at the origin. Then the point cloud should be tilted away from the zero line. The variable is probably not needed if the two lines nearly coincide near the origin in that you cannot clearly tell that they intersect at the origin.

Shown below is output only using symbols and the following example shows how to use output to perform the Wald $t$–test.

Response = Y

Coefficient Estimates

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$\hat{\beta}_1$</td>
<td>$se(\hat{\beta}_1)$</td>
<td>$t_{o,1}$</td>
<td>for $H_0: \beta_1 = 0$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\hat{\beta}_2$</td>
<td>$se(\hat{\beta}_2)$</td>
<td>$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$</td>
<td>for $H_0: \beta_2 = 0$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$x_p$</td>
<td>$\hat{\beta}_p$</td>
<td>$se(\hat{\beta}_p)$</td>
<td>$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$</td>
<td>for $H_0: \beta_p = 0$</td>
</tr>
</tbody>
</table>

Output for Ex. 2.12

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-7736.26</td>
<td>2660.36</td>
<td>-2.908</td>
<td>0.0079</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.180225</td>
<td>0.00503871</td>
<td>35.768</td>
<td>0.0000</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-1.89411</td>
<td>2.65789</td>
<td>-0.713</td>
<td>0.4832</td>
</tr>
</tbody>
</table>

R Squared: 0.988, Sigma hat: 4756.08, n = 26

Summary Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>41380950140.</td>
<td>20690475070.</td>
<td>914.69</td>
<td>0.00</td>
</tr>
<tr>
<td>Residual</td>
<td>23</td>
<td>520265969.</td>
<td>22620260.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 2.12. The output above was collected from 26 districts in Prussia in 1843. See Hebbler (1847). The goal is to study the relationship between \( Y = \) the number of women married to civilians in the district with the predictors \( x_2 = \) the population of the district, and \( x_3 = \) military women = number of women married to husbands in the military.

a) Find a 95% confidence interval for \( \beta_2 \) corresponding to population.

The CI is \( \hat{\beta}_k \pm t_{n-p,1-\delta/2} \text{se}(\hat{\beta}_k) \). Since \( n = 26, \text{df} = n-p = 26 - 3 = 23 \). From the \( t \)-table at the end of Chapter 14, intersect the df = 23 row with the column that is labelled by 95% in the CI row near the bottom of the table. Then \( t_{n-p,1-\delta/2} = 2.069 \). Using the output shows that the 95% CI is \( 0.180225 \pm 2.069(0.00503871) = [0.16980, 0.19065] \).

b) Perform a 4 step test for Ho: \( \beta_2 = 0 \) corresponding to population.

i) Ho: \( \beta_2 = 0 \)

ii) \( t_{02} = 35.768 \)

iii) pval = 0.0

iv) Reject Ho, the population is needed in the MLR model for the number of women married to civilians if the number of military women is in the model.

c) Perform a 4 step test for Ho: \( \beta_3 = 0 \) corresponding to military women.

i) Ho: \( \beta_3 = 0 \)

ii) \( t_{03} = -0.713 \)

iii) pval = 0.4883

iv) Fail to reject Ho, the number of military women is not needed in the MLR model for the number of women married to civilians if population is in the model.

Figure 2.4, made with the commands shown below, shows the added variable plots for \( x_2 \) and \( x_3 \). The plot for \( x_2 \) strongly suggests that \( x_2 \) is needed in the MLR model while the plot for \( x_3 \) indicates that \( x_3 \) does not seem to be very important. The slope of the OLS line in a) is 0.1802 while the slope of the line in b) is -1.894.

```r
source("G:/lregdata.txt")
x2 <- marry[,1]
x3 <- marry[,5]
y <- marry[,3]

#par(mfrow=c(1,2),pty="s")
#square plots look nice but have too much white space
par(mfrow=c(1,2))
resy2 <- residuals(lm(y~x3))
resx2 <- residuals(lm(x2~x3))
plot(resx2,resy2)
abline(lsfit(resx2,resy2)$coef)
title("a) Added Variable Plot for x2")
resy3 <- residuals(lm(y~x2))
```
2.8 The OLS Criterion

The OLS estimator $\hat{\beta}$ minimizes the OLS criterion

$$Q_{OLS}(\eta) = \sum_{i=1}^{n} r_{i}^2(\eta)$$

resx3 <- residuals(lm(x3~x2))
plot(resx3,resy3)
abline(lmfit(resx3,resy3)$coef)
title("b) Added Variable Plot for x3")
par(mfrow=c(1,1))

If the predictor $x_k$ is categorical, e.g. gender, the added variable plot may look like two spheres, but if the OLS line is added to the plot, it will have slope equal to $\hat{\beta}_k$.
Multiple Linear Regression

where the residual $r_i(\eta) = Y_i - x_i^T \eta$. In other words, let $r_i = r_i(\beta)$ be the OLS residuals. Then $\sum_{i=1}^n r_i^2 \leq \sum_{i=1}^n r_i^2(\eta)$ for any $p \times 1$ vector $\eta$, and the equality holds iff $\eta = \beta$ if the $n \times p$ design matrix $X$ is of full rank $p \leq n$. In particular, if $X$ has full rank $p$, then $\sum_{i=1}^n r_i^2 < \sum_{i=1}^n r_i^2(\beta) = \sum_{i=1}^n e_i^2$ even if the MLR model $Y = X\beta + e$ is a good approximation to the data.

**Example 2.13.** When a model depends on the predictors $x$ only through the linear combination $x^T \beta$, then $x^T \beta$ is called a sufficient predictor and $x^T \hat{\beta}$ is called an estimated sufficient predictor (ESP). For OLS the model is $Y = x^T \beta + e$, and the fitted value $\hat{Y} = ESP$. To illustrate the OLS criterion graphically, consider the Gladstone (1905) data where we used brain weight as the response. A constant, $x_2 = age$, $x_3 = sex$, and $x_4 = (size)^{1/3}$ were used as predictors after deleting five "infants" from the data set. In Figure 2.5a, the OLS response plot of the OLS ESP $\hat{Y}$ versus $Y$ is shown. The vertical deviations from the identity line are the residuals, and OLS minimizes the sum of squared residuals. If any other ESP $x^T \eta$ is plotted versus $Y$, then the vertical
deviations from the identity line are the residuals \( r_i(\eta) \). For this data, the OLS estimator \( \hat{\beta} = (498.726, -1.597, 30.462, 0.696)^T \). Figure 2.5b shows the response plot using the ESP \( x^T \eta \) where \( \eta = (498.726, -1.597, 30.462, 0.796)^T \). Hence only the coefficient for \( x_4 \) was changed; however, the residuals \( r_i(\eta) \) in the resulting plot are much larger in magnitude on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting \( \eta \) will be such that the squared residuals are massive.

**Proposition 2.10.** The OLS estimator \( \hat{\beta} \) is the unique minimizer of the OLS criterion if \( X \) has full rank \( p \leq n \).

**Proof:** Seber and Lee (2003, pp. 36–37). Recall that the hat matrix \( H = X(X^T X)^{-1} X^T \) and notice that \( (I - H)H = 0 \) and that \( HX = X \). Let \( \eta \) be any \( p \times 1 \) vector. Then

\[
(Y - X\hat{\beta})^T (X\hat{\beta} - X\eta) = (Y - HY)^T (HY - HX\eta) = 
Y^T (I - H)H (Y - X\eta) = 0.
\]

Thus \( Q_{OLS}(\eta) = \|Y - X\eta\|^2 = \|Y - X\hat{\beta} + X\hat{\beta} - X\eta\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\eta\|^2 + 2(Y - X\beta)^T (X\hat{\beta} - X\eta). \)

Hence

\[
\|Y - X\eta\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\eta\|^2.
\] (2.21)

So

\[
\|Y - X\eta\|^2 \geq \|Y - X\hat{\beta}\|^2
\]

with equality iff \( X(\hat{\beta} - \eta) = 0 \)

iff \( \hat{\beta} = \eta \) since \( X \) is full rank. \( \square \)

Alternatively calculus can be used. Notice that \( r_i(\eta) = Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p \). Recall that \( x_i^T \) is the \( i \)th row of \( X \) while \( v_j \) is the \( j \)th column. Since \( Q_{OLS}(\eta) = \)

\[
\sum_{i=1}^{n} (Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p)^2,
\]

the \( j \)th partial derivative

\[
\frac{\partial Q_{OLS}(\eta)}{\partial \eta_j} = -2\sum_{i=1}^{n} x_{i,j}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p) = -2(v_j)^T (Y - X\eta)
\]
for \( j = 1, \ldots, p \). Combining these equations into matrix form, setting the derivative to zero and calling the solution \( \hat{\beta} \) gives
\[
X^T Y - X^T X \hat{\beta} = 0,
\]
or
\[
X^T X \hat{\beta} = X^T Y. \tag{2.22}
\]
Equation (2.22) is known as the normal equations. If \( X \) has full rank, then \( \hat{\beta} = (X^T X)^{-1} X^T Y \). To show that \( \hat{\beta} \) is the global minimizer of the OLS criterion, use the argument following Equation (2.21).

2.9 Two Important Special Cases

When studying a statistical model, it is often useful to try to understand the model that contains a constant but no nontrivial predictors, then try to understand the model with a constant and one nontrivial predictor, then the model with a constant and two nontrivial predictors, and then the general model with many predictors. In this text, most of the models are such that \( Y \) is independent of \( x \) given \( x^T \beta \), written
\[
Y \perp x | x^T \beta.
\]
Then \( w_i = x_i^T \hat{\beta} \) is a scalar, and trying to understand the model in terms of \( x_i^T \hat{\beta} \) is about as easy as trying to understand the model in terms of one nontrivial predictor. In particular, the response plot of \( x_i^T \hat{\beta} \) versus \( Y_i \) is essential.

For MLR, the two main benefits of studying the MLR model with one nontrivial predictor \( X \) are that the data can be plotted in a scatterplot of \( X_i \) versus \( Y_i \) and that the OLS estimators can be computed by hand with the aid of a calculator if \( n \) is small.

2.9.1 The Location Model

The location model
\[
Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{2.23}
\]
is a special case of the multiple linear regression model where \( p = 1, X = 1, \) and \( \beta = \beta_1 = \mu \). This model contains a constant but no nontrivial predictors.

In the location model, \( \hat{\beta}_{OLS} = \hat{\beta}_1 = \hat{\mu} = \bar{Y} \). To see this, notice that
2.9 Two Important Special Cases

\[ Q_{OLS}(\eta) = \sum_{i=1}^{n} (Y_i - \eta)^2 \] and \[ \frac{dQ_{OLS}(\eta)}{d\eta} = -2 \sum_{i=1}^{n} (Y_i - \eta) \).

Setting the derivative equal to 0 and calling the solution \( \hat{\mu} \) gives \( \sum_{i=1}^{n} Y_i = n \hat{\mu} \) or \( \hat{\mu} = \bar{Y} \). The second derivative \[ \frac{d^2Q_{OLS}(\eta)}{d\eta^2} = 2n > 0, \]

hence \( \hat{\mu} \) is the global minimizer.

2.9.2 Simple Linear Regression

The simple linear regression (SLR) model is

\[ Y_i = \beta_1 + \beta_2 X_i + e_i = \alpha + \beta X_i + e_i \]

where the \( e_i \) are iid with \( E(e_i) = 0 \) and \( \text{VAR}(e_i) = \sigma^2 \) for \( i = 1,\ldots,n \). The \( Y_i \) and \( e_i \) are random variables while the \( X_i \) are treated as known constants. The parameters \( \beta_1, \beta_2, \) and \( \sigma^2 \) are unknown constants that need to be estimated. (If the \( X_i \) are random variables, then the model is conditional on the \( X_i \)'s provided that the errors \( e_i \) are independent of the \( X_i \). Hence the \( X_i \)'s are still treated as constants.)

The SLR model is a special case of the MLR model with \( p = 2 \), \( x_{i,1} \equiv 1 \), and \( x_{i,2} = X_i \). The normal SLR model adds the assumption that the \( e_i \) are iid N(0, \( \sigma^2 \)). That is, the error distribution is normal with zero mean and constant variance \( \sigma^2 \). The response variable \( Y \) is the variable that you want to predict while the predictor variable \( X \) is the variable used to predict the response. For SLR, \( E(Y_i) = \beta_1 + \beta_2 X_i \) and the line \( E(Y) = \beta_1 + \beta_2 X \) is the regression function. \( \text{VAR}(Y_i) = \sigma^2 \).

For SLR, the least squares estimators \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) minimize the least squares criterion \( Q(\eta_1, \eta_2) = \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)^2 \). For a fixed \( \eta_1 \) and \( \eta_2 \), \( Q \) is the sum of the squared vertical deviations from the line \( Y = \eta_1 + \eta_2 X \).

The least squares (OLS) line is \( \hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X \) where the slope

\[ \hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]

and the intercept \( \hat{\beta}_1 \equiv \hat{\alpha} = \bar{Y} - \hat{\beta}_2 \bar{X} \).

By the chain rule,

\[ \frac{\partial Q}{\partial \eta_1} = -2 \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i) \]
and
\[ \frac{\partial^2 Q}{\partial \eta_1^2} = 2n. \]

Similarly,
\[ \frac{\partial Q}{\partial \eta_2} = -2 \sum_{i=1}^{n} X_i (Y_i - \eta_1 - \eta_2 X_i) \]

and
\[ \frac{\partial^2 Q}{\partial \eta_2^2} = 2 \sum_{i=1}^{n} X_i^2. \]

Setting the first partial derivatives to zero and calling the solutions \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) shows that the OLS estimators \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) satisfy the normal equations:

\[ \sum_{i=1}^{n} Y_i = n \hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \]

and
\[ \sum_{i=1}^{n} X_i Y_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2. \]

The first equation gives \( \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \).

There are several equivalent formulas for the slope \( \hat{\beta}_2 \).

\[ \hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n} (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n} (\sum_{i=1}^{n} X_i)^2} \]

\[ = \frac{\sum_{i=1}^{n} (X_i - \bar{X})Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2} = \hat{\rho}_s Y / s_X. \]

Here the sample correlation \( \hat{\rho} \equiv \hat{\rho}(X, Y) = \text{corr}(X, Y) = \)

\[ \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_X s_Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \]

where the sample standard deviation

\[ s_W = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2} \]

for \( W = X, Y \). Notice that the term \( n - 1 \) that occurs in the denominator of \( \hat{\rho}, s_Y^2, \) and \( s_X^2 \) can be replaced by \( n \) as long as \( n \) is used in all 3 quantities.

Also notice that the slope \( \hat{\beta}_2 = \sum_{i=1}^{n} k_i Y_i \) where the constants
2.10 The No Intercept MLR Model

The no intercept MLR model, also known as regression through the origin, is still \( Y = X\beta + e \), but there is no intercept in the model, so \( X \) does not contain a column of ones \( 1 \). Hence the intercept term \( \beta_1 = \beta_1(1) \) is replaced by \( \beta_1 x_{i1} \). Software gives output for this model if the “no intercept” or “intercept = F” option is selected. For the no intercept model, the assumption \( E(e) = 0 \) is important, and this assumption is rather strong.

Many of the usual MLR results still hold: \( \hat{\beta}_{OLS} = (X^T X)^{-1}X^T Y \), the vector of predicted fitted values \( \hat{Y} = X\hat{\beta}_{OLS} = HY \) where the hat matrix \( H = X(X^T X)^{-1}X^T \) provided the inverse exists, and the vector of residuals is \( r = Y - \hat{Y} \). The response plot and residual plot are made in the same way and should be made before performing inference.

The main difference in the output is the ANOVA table. The ANOVA \( F \) test in Section 2.4 tests \( H_0 : \beta_2 = \cdots = \beta_p = 0 \). The test in this section tests \( H_0 : \beta_1 = \cdots = \beta_p = 0 \equiv H_0 : \beta = 0 \). The following definition and test follows Guttman (1982, p. 147) closely.

**Definition 2.25.** Assume that \( Y = X\beta + e \) where the \( e_i \) are iid. Assume that it is desired to test \( H_0 : \beta = 0 \) versus \( H_A : \beta \neq 0 \).

a) The uncorrected total sum of squares

\[
SST = \sum_{i=1}^{n} Y_i^2. \tag{2.25}
\]

b) The model sum of squares

\[
SSM = \sum_{i=1}^{n} \hat{Y}_i^2. \tag{2.26}
\]

c) The residual sum of squares or error sum of squares is

\[
SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{2.27}
\]

d) The degrees of freedom (df) for SSM is \( p \), the df for SSE is \( n - p \) and the df for SST is \( n \). The mean squares are \( \text{MSE} = SSE/(n - p) \) and \( \text{MSM} = SSM/p \).

The ANOVA table given for the “no intercept” or “intercept = F” option is below.
Summary Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>p</td>
<td>SSM</td>
<td>MSM</td>
<td>Fo=MSM/MSE for Ho: β = 0</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 4 step no intercept ANOVA F test for $\beta = 0$ is below.

i) State the hypotheses Ho: $\beta = 0$, Ha: $\beta \neq 0$.

ii) Find the test statistic $F_o = \text{MSM}/\text{MSE}$ or obtain it from output.

iii) Find the p-val from output or use the $F$–table: $\text{pval} = P(F_{p,n-p} > F_o)$.

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is an MLR relationship between $Y$ and the predictors $x_1, \ldots, x_p$. If you fail to reject Ho, conclude that there is not an MLR relationship between $Y$ and the predictors $x_1, \ldots, x_p$. (Or there is not enough evidence to conclude that there is an MLR relationship between $Y$ and the predictors.)

**Warning:** Several important models can be cast in the no intercept MLR form, but often a different test than Ho : $\beta = 0$ is desired. For example, when the generalized or weighted least squares models of Chapter 4 are transformed into no intercept MLR form, the test of interest is Ho: $\beta_2 = \cdots = \beta_p = 0$. The one way ANOVA model of Chapter 5 is equivalent to the cell means model, which is in no intercept MLR form, but the test of interest is Ho : $\beta_1 = \cdots = \beta_p$.

**Proposition 2.11.** Suppose $Y = X\beta + e$ where $X$ may or may not contain a column of ones. Then the partial $F$ test of Section 2.6 can be used for inference.

**Example 2.14.** Consider the Gladstone (1905) data described in Example 2.5. If the file of data sets bregdata is downloaded into $R$, then the ANOVA $F$ statistic for testing $\beta_2 = \cdots = \beta_4 = 0$ can be found with the following commands. The command `lsfit` adds a column of ones to $x$ which contains the variables size, sex, breadth, and circumference. Three of these predictor variables are head measurements. Then the response $Y$ is brain weight, and the model contains a constant (intercept).

```r
> y <- cbrainy
> x <- cbrainx[,c(11,10,3,6)]
> ls.print(lsfit(x,y))
F-statistic (df=4, 262)=196.2433
```

The ANOVA $F$ test can also be found with the no intercept model by adding a column of ones to the $R$ matrix $x$ and then performing the partial $F$ test with the full model and the reduced model that only uses the column of ones. Notice that the “intercept=F” option needs to be used to fit both models. The residual standard error = RSE = $\sqrt{MSE}$. Thus $\text{SSE} = (n - k)(\text{RSE})^2$ where $n - k$ is the denominator degrees of freedom for the $F$ test.
and \( k \) is the numerator degrees of freedom = number of variables in the model. The column of ones \( x_{\text{one}} \) is counted as a variable. The last line of output computes the partial F statistic and is again \( \approx 196.24 \).

\[
> \text{xone} <- 1 + 0*1:267 \\
> \text{x} <- \text{cbind(xone,x)} \\
> \text{ls.print(lsfit(x,y,intercept=F))}
\]

Residual Standard Error=82.9175  
F-statistic (df=5, 262)=12551.02

| Estimate | Std.Err | t-value | Pr(>|t|) |
|----------|---------|---------|----------|
| xone  | 99.8495  | 171.6189 | 0.5818 | 0.5612 |
| size   | 0.2209   | 0.0358   | 6.1733 | 0.0000 |
| sex    | 22.5491  | 11.2372  | 2.0066 | 0.0458 |
| breadth| -1.2464  | 1.5139   | -0.8233| 0.4111 |
| circum | 1.0255   | 0.4719   | 2.1733 | 0.0307 |

\[
> \text{ls.print(lsfit(x[,1],y,intercept=F))}
\]

Residual Standard Error=164.5028  
F-statistic (df=1, 266)=15744.48

| Estimate | Std.Err | t-value | Pr(>|t|) |
|----------|---------|---------|----------|
| X       | 1263.228 | 10.0674 | 125.477  | 0 |

\[
((266*(164.5028)^2 - 262*(82.9175)^2)/4)/(82.9175)^2
\]

[1] 196.2435

2.11 Summary

1) The response variable is the variable that you want to predict. The predictor variables are the variables used to predict the response variable.

2) Regression is the study of the conditional distribution \( Y|x \).

3) The MLR model is

\[
Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = x_i^T\beta + e_i
\]

for \( i = 1, \ldots, n \). Here \( n \) is the sample size and the random variable \( e_i \) is the \( i \)th error. Assume that the errors are iid with \( E(e_i) = 0 \) and \( \text{VAR}(e_i) = \sigma^2 < \infty \). Assume that the errors are independent of the predictor variables \( x_i \). The unimodal MLR model assumes that the \( e_i \) are iid from a unimodal distribution that is not highly skewed. Usually \( x_{i,1} \equiv 1 \).

4) In matrix notation, these \( n \) equations become
\[ Y = X\beta + e, \]

where \( Y \) is an \( n \times 1 \) vector of dependent variables, \( X \) is an \( n \times p \) matrix of predictors, \( \beta \) is a \( p \times 1 \) vector of unknown coefficients, and \( e \) is an \( n \times 1 \) vector of unknown errors.

5) The OLS estimators are \( \hat{\beta}_{OLS} = (X^T X)^{-1}X^T Y \) and \( \hat{\sigma}^2 = MSE = \sum_{i=1}^{n} r_i^2 / (n - p) \). Thus \( \hat{\sigma} = \sqrt{MSE} \). The vector of predicted or fitted values \( \hat{Y}_{OLS} = X\hat{\beta}_{OLS} = HY \) where the hat matrix \( H = X(X^T X)^{-1}X^T \). The \( i \)th fitted value \( \hat{Y}_i = x_i^T \hat{\beta} \). The \( i \)th residual \( r_i = Y_i - \hat{Y}_i \) and the vector of residuals \( r = Y - \hat{Y} = (I - H)Y \). The least squares regression equation for a model containing a constant is \( \hat{Y} = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p \).

6) Always make the response plot of \( \hat{Y} \) versus \( Y \) and residual plot of \( \hat{Y} \) versus \( r \) for any MLR analysis. The response plot is used to visualize the MLR model, that is, to visualize the conditional distribution of \( Y|x^T \beta \). If the unimodal MLR model of 3) is useful, then i) the plotted points in the response plot should scatter about the identity line with no other pattern, and ii) the plotted points in the residual plot should scatter about the \( r = 0 \) line with no other pattern. If either i) or ii) is violated, then the unimodal MLR model is not sustained. In other words, if the plotted points in the residual plot show some type of dependency, e.g. increasing variance or a curved pattern, then the multiple linear regression model may be inadequate.

7) Use \( x_f \leq \max h_i \) for valid predictions.

8) If the MLR model contains a constant, then SSTO = SSE + SSR where \( SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \), \( SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \) and \( SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2 \).

9) If the MLR model contains a constant, then \( R^2 = [\text{corr}(Y_i, \hat{Y}_i)]^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \).

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>p-1</td>
<td>SSR</td>
<td>MSR</td>
<td>Fo=MSR/MSE</td>
<td>for Ho: ( \beta_2 = \cdots = \beta_p = 0 )</td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>SSE</td>
<td>MSE</td>
<td>p-value</td>
<td></td>
</tr>
</tbody>
</table>

10) Be able to perform the 4 step ANOVA \( F \) test of hypotheses.
   i) State the hypotheses \( H_0: \beta_2 = \cdots = \beta_p = 0 \) \( H_a: \) not \( H_0 \).
   ii) Find the test statistic \( F_o = MSR/MSE \) or obtain it from output.
   iii) Find pval, the estimated pvalue, from output or use the \( F \)-table: \( pval = P(F_{p-1,n-p} > F_o) \).
   iv) State whether you reject \( H_0 \) or fail to reject \( H_0 \). If \( H_0 \) is rejected, conclude that there is an MLR relationship between \( Y \) and the predictors \( x_2, \ldots, x_p \). If you fail to reject \( H_0 \), conclude that there is a not an MLR relationship between \( Y \) and the predictors \( x_2, \ldots, x_p \).
11) The large sample $(1 - \delta)\%$ CI for $E(Y_f|x_f) = x_f^T\beta = E(\hat{Y}_f)$ is \(\hat{Y}_f \pm t_{n-p,1-\delta/2}se(\hat{Y}_f)\) where $P(T \leq t_{n-p,\delta}) = \delta$ if $T$ has a $t$ distribution with $n - p$ degrees of freedom.

12) The classical $(1 - \delta)\%$ PI for $Y_f$ is $\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred)$, but should be replaced with the asymptotically optimal PI (2.20).

Full model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Fo and p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$p-1$</td>
<td>SSR</td>
<td>MSR</td>
<td>$F_0=MSR/MSE$</td>
</tr>
<tr>
<td>Residual</td>
<td>$df_F = n-p$</td>
<td>SSE(F)</td>
<td>MSE(F)</td>
<td>for $Ho: \beta_2 = \cdots = \beta_p = 0$</td>
</tr>
</tbody>
</table>

Reduced model

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Fo and p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$q-1$</td>
<td>SSR</td>
<td>MSR</td>
<td>$F_0=MSR/MSE$</td>
</tr>
<tr>
<td>Residual</td>
<td>$df_R = n-q$</td>
<td>SSE(R)</td>
<td>MSE(R)</td>
<td>for $Ho: \beta_2 = \cdots = \beta_q = 0$</td>
</tr>
</tbody>
</table>

13) Be able to perform the 4 step partial $F$ test of hypotheses. i) State the hypotheses $Ho$: the reduced model is good $Ha$: use the full model.

ii) Find the test statistic $F_R = \frac{[SSE(R) - SSE(F)]}{df_R - df_F} / MSE(F)$

iii) Find the pval = $P(F_{df_R-df_F, df_F} > F_R)$. (On exams typically an $F$ table is used. Here $df_R - df_F = p-q$ = number of parameters set to 0, and $df_F = n-p$, while pval is the estimated p-value.)

iv) State whether you reject $Ho$ or fail to reject $Ho$. Reject $Ho$ if the pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject $Ho$ and conclude that the reduced model is good.

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$\hat{\beta}_1$</td>
<td>$se(\hat{\beta}_1)$</td>
<td>$t_{o,1}$</td>
<td>for $Ho: \beta_1 = 0$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\hat{\beta}_2$</td>
<td>$se(\hat{\beta}_2)$</td>
<td>$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$</td>
<td>for $Ho: \beta_2 = 0$</td>
</tr>
<tr>
<td>...</td>
<td>$\hat{\beta}_p$</td>
<td>$se(\hat{\beta}_p)$</td>
<td>$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$</td>
<td>for $Ho: \beta_p = 0$</td>
</tr>
</tbody>
</table>

14) The $100 (1 - \delta)\%$ CI for $\beta_k$ is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p \geq 30$, the $N(0,1)$ cutoff $z_{1-\delta/2}$ may be used.

15) The corresponding 4 step $t$-test of hypotheses has the following steps.

i) State the hypotheses $Ho: \beta_k = 0$ $Ha: \beta_k \neq 0$.

ii) Find the test statistic $t_{o,k} = \hat{\beta}_k/se(\hat{\beta}_k)$ or obtain it from output.

iii) Find the pval from output or use the $t$-table: pval =
2P(t_{n-p} < -|t_{o,k}|) = 2P(t_{n-p} > |t_{o,k}|).

Use the normal table or the $d = Z$ line in the $t$–table if the degrees of freedom $d = n - p \geq 30$.

iv) State whether you reject $H_0$ or fail to reject $H_0$ and give a nontechnical sentence restating your conclusion in terms of the story problem. If $H_0$ is rejected, then conclude that $x_k$ is needed in the MLR model for $Y$ given that the other predictors are in the model. If you fail to reject $H_0$, then conclude that $x_k$ is not needed in the MLR model for $Y$ given that the other predictors are in the model.

16) Given $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$, $\sum_{i=1}^{n} (X_i - \bar{X})^2$, $\bar{X}$, and $\bar{Y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

and $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$.

17) Given $\hat{\rho}$, $s_X$, $s_Y$, $\bar{X}$, and $\bar{Y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where $\hat{\beta}_2 = \hat{\rho} s_Y / s_X$ and $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$.

2.12 Complements

The Least Squares Central Limit Theorem 2.8 is often a good approximation if $n \geq 10p$ and the error distribution has “light tails,” i.e. the probability of an outlier is nearly 0 and the tails go to zero at an exponential rate or faster. For error distributions with heavier tails, much larger samples are needed, and the assumption that the variance $\sigma^2$ exists is crucial, e.g. Cauchy errors are not allowed. Norman and Streiner (1986, p. 63) recommend $n \geq 5p$.

The classical MLR prediction interval does not work well and should be replaced by the Olive (2007) asymptotically optimal PI (2.20). Lei and Wasserman (2014) provide an alternative: use the Lei et al. (2013) PI $[\hat{r}_L, \hat{r}_U]$ on the residuals, then the PI for $Y_f$ is

$$[\hat{Y}_f + \hat{r}_L, \hat{Y}_f + \hat{r}_U].$$

(2.28)

Bootstrap PIs need more theory and instead of using $B = 1000$ samples, use $B = \max(1000, n)$. See Olive (2014, pp. 279–285).

For the additive error regression model $Y = m(x) + e$, the response plot of $\hat{Y} = \hat{m}(x)$ vs. $Y$, with the identity line added as a visual aid, is used like the MLR response plot. We want $n \geq 10 \, df$ where $df$ is the degrees of freedom from fitting $\hat{m}$. Olive (2013a) provides PIs for this model, including the location model. These PIs are large sample PIs provided that the sample quantiles of the residuals are consistent estimators of the population quantiles.
of the errors. The response plot and PIs could also be used for methods described in James et al. (2013) such as ridge regression, lasso, principal components regression, and partial least squares. See Pelawa Watagoda and Olive (2017) if \( n \) is not large compared to \( p \).

In addition to large sample theory, we want the PIs to work well on a single data set as future observations are gathered, but only have the training data \((x_1, Y_1), \ldots, (x_n, Y_n)\). Much like \( k \)-fold cross validation for discriminant analysis, randomly divide the data set into \( k = 5 \) groups of approximately equal size. Compute the model from 4 groups and use the 5th group as a validation set: compute the PI for \( x_f = x_j \) for each \( j \) in the 5th group. Repeat so each of the 5 groups is used as a validation set. Compute the proportion of times \( Y_i \) was in its PI for \( i = 1, \ldots, n \) as well as the average length of the \( n \) PIs. We want the proportion near the nominal proportion and short average length if two or more models or PIs are being considered.

Following Chapter 11, under the regularity conditions, much of the inference that is valid for the normal MLR model is approximately valid for the unimodal MLR model when the sample size is large. For example, confidence intervals for \( \beta_i \) are asymptotically correct, as are \( t \) tests for \( \beta_i = 0 \) (see Li and Duan (1989, p. 1035)), the MSE is an estimator of \( \sigma^2 \) by Theorems 2.6 and 2.7, and variable selection procedures perform well (see Chapter 3 and Olive and Hawkins 2005).


Cook and Weisberg (1997, 1999a: ch. 17) call a plot that emphasizes model agreement a model checking plot. Anscombe (1961) and Anscombe and Tukey (1963) suggested graphical methods for checking multiple linear regression and experimental design methods that were the “state of the art” at the time.

The rules of thumb given in this chapter for residual plots are not perfect. Cook (1998, pp. 4–6) gives an example of a residual plot that looks like a right opening megaphone, but the MLR assumption that was violated was linearity, not constant variance. Ghosh (1987) gives an example where the residual plot shows no pattern even though the constant variance assumption is violated. Searle (1988) shows that residual plots will have parallel lines if several cases take on each of the possible values of the response variable, e.g. if the response is a count.

Several authors have suggested using the response plot to visualize the coefficient of determination \( R^2 \) in multiple linear regression. See, for example, Chambers et al. (1983, p. 280). Anderson-Sprecher (1994) provides an excellent discussion about \( R^2 \). Kachigan (1982, pp. 174–177) also gives a good explanation of \( R^2 \). Also see Kvålseth (1985), and Freedman (1983).

Hoaglin and Welsh (1978) discuss the hat matrix \( H \), and Brooks et al. (1988) recommend using \( x_f < \max h_i \) for valid predictions. Simultaneous
prediction intervals are given by Sadooghi-Alvandi (1990). Olive (2007) suggests three large sample prediction intervals for MLR that are valid under the unimodal MLR model. Also see Schoemoyer (1992).

Sall (1990) discusses the history of added variable plots while Darlington (1969) provides an interesting proof that $\hat{\beta}$ minimizes the OLS criterion.

### 2.12.1 Lack of Fit Tests

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>$\hat{\beta}_1$</td>
<td>$se(\hat{\beta}_1)$</td>
<td>$t_{o,1}$</td>
<td>for Ho: $\beta_1 = 0$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$\hat{\beta}_2$</td>
<td>$se(\hat{\beta}_2)$</td>
<td>$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$</td>
<td>for Ho: $\beta_2 = 0$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_p$</td>
<td>$\hat{\beta}_p$</td>
<td>$se(\hat{\beta}_p)$</td>
<td>$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$</td>
<td>for Ho: $\beta_p = 0$</td>
</tr>
</tbody>
</table>

R Squared: $R^2$
Sigma hat: $\sqrt{MSE}$
Number of cases: $n$
Degrees of Freedom : $n - p$

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>p-1</td>
<td>SSR</td>
<td>MSR</td>
<td>$F_0$=MSR/MSE</td>
<td>for Ho: $\beta_2 = \cdots = \beta_p = 0$</td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The typical “relevant OLS output” has the form given above, but occasionally software also includes output for a lack of fit test as shown below.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$F_0$=MSR/MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>p-1</td>
<td>SSR</td>
<td>MSR</td>
<td>$F_0$=MSR/MSE</td>
</tr>
<tr>
<td>Residual</td>
<td>n-p</td>
<td>SSE</td>
<td>MSE</td>
<td></td>
</tr>
<tr>
<td>lack of fit</td>
<td>c-p</td>
<td>SSLF</td>
<td>MSLF</td>
<td>$F_{LF} = MSLF/MSPE$</td>
</tr>
<tr>
<td>pure error</td>
<td>n-c</td>
<td>SSPE</td>
<td>MSPE</td>
<td></td>
</tr>
</tbody>
</table>

The lack of fit test assumes that

$$Y_i = m(x_i) + e_i$$ (2.29)

where $E(Y_i|x_i) = m(x_i)$, $m$ is some possibly nonlinear function, and that the $e_i$ are iid $N(0, \sigma^2)$. Notice that the MLR model is the special case with $m(x_i) = x_i^T \beta$. The lack of fit test needs at least one repicate: 2 or more $Y$s with the same value of predictors $x$. Then there a $c$ “replicate groups” with $n_j$ observations in the $j$th group. Each group has the vector of predictors $x_j$. 


say, and at least one \( n_j > 1 \). Also, \( \sum_{j=1}^{c} n_j = n \). Denote the Ys in the \( j \)th group by \( Y_{ij} \), and let the sample mean of the Ys in the \( j \)th group be \( \bar{Y}_j \). Then

\[
\frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2
\]

is an estimator of \( \sigma^2 \) for each group with \( n_j > 1 \). Let

\[
SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2.
\]

Then \( MSPE = SSPE/(n - c) \) is an unbiased estimator of \( \sigma^2 \) when model (2.29) holds, regardless of the form of \( m \). The PE in SSPE stands for “pure error.”

Now \( SSLF = SSE - SSPE = \sum_{j=1}^{c} n_j (\bar{Y}_j - \bar{Y})^2 \). Notice that \( \bar{Y}_j \) is an unbiased estimator of \( m(x_j) \) while \( \bar{Y} \) is an estimator of \( m \) if the MLR model is appropriate: \( m(x_j) = x_j^T \beta \). Hence SSLF and MSLF can be very large if the MLR model is not appropriate.

The 4 step lack of fit test is i) \( H_0: \) no evidence of MLR lack of fit, \( H_A: \) there is lack of fit for the MLR model.

ii) \( F_{LF} = MSLF/MSPE \).

iii) The pval = \( P(F_{c-p,n-c} > F_{LF}) \).

iv) Reject \( H_0 \) if pval \( \leq \delta \) and state the \( H_A \) claim that there is lack of fit. Otherwise, fail to reject \( H_0 \) and state that there is not enough evidence to conclude that there is MLR lack of fit.

Although the lack of fit test seems clever, examining the response plot and residual plot is a much more effective method for examining whether or not the MLR model fits the data well provided that \( n \geq 10p \). A graphical version of the lack of fit test would compute the \( \bar{Y}_j \) and see whether they scatter about the identity line in the response plot. When there are no replicates, the range of \( \bar{Y} \) could be divided into several narrow nonoverlapping intervals called slices. Then the mean \( \bar{Y}_j \) of each slice could be computed and a step function with step height \( \bar{Y}_j \) at the \( j \)th slice could be plotted. If the step function follows the identity line, then there is no evidence of lack of fit. However, it is easier to check whether the \( Y_i \) are scattered about the identity line. Examining the residual plot is useful because it magnifies deviations from the identity line that may be difficult to see until the linear trend is removed. The lack of fit test may be sensitive to the assumption that the errors are iid \( N(0, \sigma^2) \).

When \( Y \perp x | x^T \beta \), then the response plot of the estimated sufficient predictor (ESP) \( x^T \hat{\beta} \) versus \( Y \) is used to visualize the conditional distribution of \( Y | x^T \beta \), and will often greatly outperform the corresponding lack of fit test. When the response plot can be combined with a good lack of fit plot such as
a residual plot, using a one number summary of lack of fit such as the test statistic $F_{LF}$ makes little sense.

Nevertheless, the literature for lack of fit tests for various statistical methods is enormous. See Joglekar et al. (1989), Peña and Slate (2006), and Su and Yang (2006) for references.

For the following homework problems, Cody and Smith (2006) is useful for SAS, while Cook and Weisberg (1999a) is useful for Arc. Becker et al. (1988) and Crawley (2013) are useful for $R$.

2.13 Problems

Problems with an asterisk * are especially important.

Output for Problem 2.1
Full Model Summary Analysis of Variance Table
Source df SS MS F p-value
Regression 6 265784.44297.4172.14 0.0000
Residual 67 17240.9 257.327

Reduced Model Summary Analysis of Variance Table
Source df SS MS F p-value
Regression 1 264621.264621.1035.26 0.0000
Residual 72 18403.8 255.608

2.1. Assume that the response variable $Y$ is height, and the explanatory variables are $X_2 =$ sternal height, $X_3 =$ cephalic index, $X_4 =$ finger to ground, $X_5 =$ head length, $X_6 =$ nasal height, and $X_7 =$ bigonal breadth. Suppose that the full model uses all 6 predictors plus a constant ($= X_1$) while the reduced model uses the constant and sternal height. Test whether the reduced model can be used instead of the full model using the output above. The data set had 74 cases.

Output for Problem 2.2
Full Model Summary Analysis of Variance Table
Source df SS MS F p-value
Regression 9 16771.71863.521479148.9 0.0000
Residual 235 0.29607 0.00126

Reduced Model Summary Analysis of Variance Table
Source df SS MS F p-value
Regression 2 16771.7 8385.85 6734072.0 0.0000
Residual 242 0.301359 0.0012453
Coefficient Estimates, Response = y, Terms = (x2 x2^2)

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>958.470</td>
<td>5.88584</td>
<td>162.843</td>
<td>0.0000</td>
</tr>
<tr>
<td>x2</td>
<td>-1335.39</td>
<td>11.1656</td>
<td>-119.599</td>
<td>0.0000</td>
</tr>
<tr>
<td>x2^2</td>
<td>421.881</td>
<td>5.29434</td>
<td>79.685</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

2.2. The above output, starting on the previous page, comes from the Johnson (1996) STATLIB data set bodyfat after several outliers are deleted. It is believed that \( Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e \) where \( Y \) is the person’s bodyfat and \( X_2 \) is the person’s density. Measurements on 245 people were taken. In addition to \( X_2 \) and \( X_2^2 \), 7 additional measurements \( X_4, \ldots, X_{10} \) were taken. Both the full and reduced models contain a constant \( X_1 \equiv 1 \).

a) Predict \( Y \) if \( X_2 = 1.04 \). (Use the reduced model \( Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e \).)

b) Test whether the reduced model can be used instead of the full model.

Output for Problem 2.3

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-5.07459</td>
<td>1.85124</td>
<td>-2.741</td>
<td>0.0076</td>
</tr>
<tr>
<td>log[H]</td>
<td>1.12399</td>
<td>0.498937</td>
<td>2.253</td>
<td>0.0270</td>
</tr>
<tr>
<td>log[S]</td>
<td>0.573167</td>
<td>0.116455</td>
<td>4.922</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R Squared: 0.895655 Sigma hat: 0.223658, n = 82
(\( \log[H] \ \log[S] \)) (4 5)
Prediction = 2.2872, s(pred) = 0.467664,
Estimated population mean value = 2.287, s = 0.410715

2.3. The above output was produced from the file mussels.lsp in Arc. See Cook and Weisberg (1999a). Let \( Y = \log(M) \) where \( M \) is the muscle mass of a mussel. Let \( X_1 \equiv 1, X_2 = \log(H) \) where \( H \) is the height of the shell, and let \( X_3 = \log(S) \) where \( S \) is the shell mass. Suppose that it is desired to predict \( Y_f \) if \( \log(H) = 4 \) and \( \log(S) = 5 \), so that \( x_f^T = (1, 4, 5) \). Assume that \( \text{se}(\hat{Y}_f) = 0.410715 \) and that \( \text{se}(\text{pred}) = 0.467664 \).

a) If \( x_f^T = (1, 4, 5) \) find a 99% confidence interval for \( E(Y_f) \).

b) If \( x_f^T = (1, 4, 5) \) find a 99% prediction interval for \( Y_f \).

Problem 2.4 Output, Coef. Estimates Response = height

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>227.351</td>
<td>65.1732</td>
<td>3.488</td>
<td>0.0008</td>
</tr>
<tr>
<td>sternal height</td>
<td>0.955973</td>
<td>0.0515390</td>
<td>18.549</td>
<td>0.0000</td>
</tr>
<tr>
<td>finger to ground</td>
<td>0.197429</td>
<td>0.0889004</td>
<td>2.221</td>
<td>0.0295</td>
</tr>
</tbody>
</table>
R Squared: 0.879324  Sigma hat: 22.0731

Summary Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>259167</td>
<td>129583</td>
<td>265.96</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>73</td>
<td>35567.2</td>
<td>487.222</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.4. The above output, starting on the previous page, is from the multiple linear regression of the response $Y = \text{height}$ on the two nontrivial predictors 

- **sternal height** = height at shoulder,
- **finger to ground** = distance from the tip of a person’s middle finger to the ground.

a) Consider the plot with $Y_i$ on the vertical axis and the least squares fitted values $\hat{Y}_i$ on the horizontal axis. Sketch how this plot should look if the multiple linear regression model is appropriate.

b) Sketch how the residual plot should look if the residuals $r_i$ are on the vertical axis and the fitted values $\hat{Y}_i$ are on the horizontal axis.

c) From the output, are **sternal height** and **finger to ground** useful for predicting **height**? (Perform the ANOVA $F$ test.)

2.5. Suppose that it is desired to predict the weight of the brain (in grams) from the cephalic index measurement. The output below uses data from 267 people.

<table>
<thead>
<tr>
<th>predictor</th>
<th>coef</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>865.001</td>
<td>274.252</td>
<td>3.154</td>
<td>0.0018</td>
</tr>
<tr>
<td>cephalic</td>
<td>5.05961</td>
<td>3.48212</td>
<td>1.453</td>
<td>0.1474</td>
</tr>
</tbody>
</table>

Do a 4 step test for $\beta_2 \neq 0$.

2.6. Suppose that the scatterplot of $X$ versus $Y$ is strongly curved rather than ellipsoidal. Should you use simple linear regression to predict $Y$ from $X$? Explain.

2.7. Suppose that the 95% confidence interval for $\beta_2$ is $[-17.457, 15.832]$. In the simple linear regression model, is $X$ a useful linear predictor for $Y$? If your answer is no, could $X$ be a useful predictor for $Y$? Explain.

2.8. Suppose it is desired to predict the yearly return from the stock market from the return in January. Assume that the correlation $\hat{\rho} = 0.496$. Using the table below, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$.

<table>
<thead>
<tr>
<th>variable</th>
<th>mean $\overline{X}$ or $\overline{Y}$</th>
<th>standard deviation $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>January return</td>
<td>1.75</td>
<td>5.36</td>
</tr>
<tr>
<td>yearly return</td>
<td>9.07</td>
<td>15.35</td>
</tr>
</tbody>
</table>
2.9. Suppose that \[ \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 70690.0, \]
\[ \sum (X_i - \bar{X})^2 = 19800.0, \bar{X} = 70.0, \text{ and } \bar{Y} = 312.28. \]
a) Find the least squares slope \( \hat{\beta}_2 \).
b) Find the least squares intercept \( \hat{\beta}_1 \).
c) Predict \( Y \) if \( X = 80 \).

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i - \bar{x} )</th>
<th>( y_i - \bar{y} )</th>
<th>( (x_i - \bar{x})(y_i - \bar{y}) )</th>
<th>( (x_i - \bar{x})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>41</td>
<td>5</td>
<td>14</td>
<td>-70</td>
<td>25</td>
</tr>
<tr>
<td>56</td>
<td>63</td>
<td>6</td>
<td>23</td>
<td>141</td>
<td>36</td>
</tr>
<tr>
<td>59</td>
<td>70</td>
<td>8</td>
<td>20</td>
<td>160</td>
<td>64</td>
</tr>
<tr>
<td>64</td>
<td>72</td>
<td>5</td>
<td>12</td>
<td>-60</td>
<td>25</td>
</tr>
<tr>
<td>74</td>
<td>84</td>
<td>11</td>
<td>14</td>
<td>154</td>
<td>121</td>
</tr>
</tbody>
</table>

2.10. In the above table, \( x_i \) is the length of the femur and \( y_i \) is the length of the humerus taken from five dinosaur fossils (Archaeopteryx) that preserved both bones. See Moore (2000, p. 99).

a) Complete the table and find the least squares estimators \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \).
b) Predict the humerus length if the femur length is 60.

2.11. Suppose that the regression model is \( Y_i = 7 + \beta X_i + e_i \) for \( i = 1, \ldots, n \) where the \( e_i \) are iid \( N(0, \sigma^2) \) random variables. The least squares criterion is \( Q(\eta) = \sum_{i=1}^{n} (Y_i - 7 - \eta X_i)^2 \).

a) What is \( E(Y_i) \)?
b) Find the least squares estimator \( \hat{\beta} \) of \( \beta \) by setting the first derivative \( \frac{d}{d\eta} Q(\eta) \) equal to zero.
c) Show that your \( \hat{\beta} \) is the global minimizer of the least squares criterion \( Q \) by showing that the second derivative \( \frac{d^2}{d\eta^2} Q(\eta) > 0 \) for all values of \( \eta \).

2.12. The location model is \( Y_i = \mu + e_i \) for \( i = 1, \ldots, n \) where the \( e_i \) are iid with mean \( E(e_i) = 0 \) and constant variance \( \text{VAR}(e_i) = \sigma^2 \). The least squares estimator \( \hat{\mu} \) of \( \mu \) minimizes the least squares criterion \( Q(\eta) = \sum_{i=1}^{n} (Y_i - \eta)^2 \).

To find the least squares estimator, perform the following steps.
a) Find the derivative \( \frac{d}{d\eta} Q \), set the derivative equal to zero and solve for \( \eta \). Call the solution \( \hat{\mu} \).

b) To show that the solution was indeed the global minimizer of \( Q \), show that \( \frac{d^2}{d\eta^2} Q > 0 \) for all real \( \eta \). (Then the solution \( \hat{\mu} \) is a local min and \( Q \) is convex, so \( \hat{\mu} \) is the global min.)

2.13. The normal error model for simple linear regression through the origin is

\[
Y_i = \beta X_i + e_i
\]

for \( i = 1, \ldots, n \) where \( e_1, \ldots, e_n \) are iid \( N(0, \sigma^2) \) random variables.

a) Show that the least squares estimator for \( \beta \) is

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.
\]

b) Find \( E(\hat{\beta}) \).

c) Find \( \text{VAR}(\hat{\beta}) \).

(Hint: Note that \( \hat{\beta} = \sum_{i=1}^{n} k_i Y_i \) where the \( k_i \) depend on the \( X_i \) which are treated as constants.)

2.14. Suppose that the regression model is \( Y_i = 10 + 2X_{i2} + \beta_3 X_{i3} + e_i \) for \( i = 1, \ldots, n \) where the \( e_i \) are iid \( N(0, \sigma^2) \) random variables. The least squares criterion is \( Q(\eta_3) = \sum_{i=1}^{n} (Y_i - 10 - 2X_{i2} - \eta_3 X_{i3})^2 \). Find the least squares estimator \( \hat{\beta}_3 \) of \( \beta_3 \) by setting the first derivative \( \frac{d}{d\eta_3} Q(\eta_3) \) equal to zero. Show that your \( \hat{\beta}_3 \) is the global minimizer of the least squares criterion \( Q \) by showing that the second derivative \( \frac{d^2}{d\eta_3^2} Q(\eta_3) > 0 \) for all values of \( \eta_3 \).

Minitab Problems

“Double click” means press the rightmost “mouse” button twice in rapid succession. “Drag” means hold the mouse button down. This technique is used to select “menu” options.

After your computer is on, get into Minitab, often by searching programs and then double clicking on the icon marked “Student Minitab.”

i) In a few seconds, the Minitab session and worksheet windows fill the screen. At the top of the screen there is a menu. The upper left corner has the menu option “File.” Move your cursor to “File” and drag down the option “Open Worksheet.” A window will appear. Double click on the icon “Student.” This will display a large number of data sets.
2.13 Problems

ii) In the middle of the screen there is a “scroll bar,” a gray line with left and right arrow keys. Use the right arrow key to make the data file “Prof.mtw” appear. Double click on “Prof.mtw.” A window will appear. Click on “OK.”

iii) The worksheet window will now be filled with data. The top of the screen has a menu. Go to “Stat” and drag down “Regression.” Another window will appear: drag down Regression (write this as Stat>Regression>Regression).

iv) A window will appear with variables to the left and the response variable and predictors (explanatory variables) to the right. Double click on “instructions” to make it the response. Double click on “manner” to make it the (predictor) explanatory variable. Then click on “OK.”

v) The required output will appear in the session window. You can view the output by using the vertical scroll bar on the right of the screen.

vi) Copy and paste the output into Word, or to print your single page of output, go to “File,” and drag down the option “Print Session Window.” A window will appear. Click on “ok.” Then get your output from the printer.

Use the F3 key to clear entries from a dialog window if you make a mistake or want a new plot.

To get out of Minitab, move your cursor to the “x” in the upper right corner of the screen. When asked whether to save changes, click on “no.”

2.15. (Minitab problem.) See the above instructions for using Minitab. Get the data set prof.mtw. Assign the response variable to be instrucr (the instructor rating from course evaluations) and the explanatory variable (predictor) to be manner (the manner of the instructor). Run a regression on these variables.

a) Place the computer output into Word.

b) Write the regression equation.

c) Predict instrucr if manner = 2.47.

d) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands “Stat>Regression>Regression” to get the regression window. Put instrucr in the Response and manner in the Predictors boxes. The click on Storage. From the resulting window click on Fits and Residuals. Then click on OK twice.

To get a response plot, use the commands “Graph>Plot,” (double click) place instrucr in the Y box, and Fits1 in the X box. Then click on OK. Print the plot by clicking on the graph and then clicking on the printer icon.

e) To make a residual plot, use the menu commands “Graph>Plot” to get a window. Place “Resi1” in the Y box and “Fits1” in the X box. Then click on OK. Print the plot by clicking on the graph and then clicking on the printer icon.
2.16. a) Enter the following data on the *Minitab* worksheet:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>73</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>60</td>
<td>128</td>
</tr>
<tr>
<td>80</td>
<td>170</td>
</tr>
<tr>
<td>40</td>
<td>87</td>
</tr>
<tr>
<td>50</td>
<td>108</td>
</tr>
<tr>
<td>60</td>
<td>135</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>70</td>
<td>148</td>
</tr>
<tr>
<td>60</td>
<td>132</td>
</tr>
</tbody>
</table>

To enter the data click on the **C1** column header and enter *x*. Then click on the **C2** header and enter *y*. Then enter the data. Or copy the data from Problem 2.17 obtained from [http://lagrange.math.siu.edu/Olive/lreghw.txt](http://lagrange.math.siu.edu/Olive/lreghw.txt).

Then in *Minitab*, use the menu commands “Edit>Paste Cells” and click on “OK.” Obtain the regression output from *Minitab* with the menu commands “Stat>Regression>Regression.”

b) Place the output into *Word*.

c) Write down the least squares equation.

d) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands “Stat>Regression>Regression” to get the regression window. Put Y in the **Response** and X in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

To make a response plot, use the menu commands “Graph>Plot” to get a window. Place “Y” in the **Y** box and “Fits1” in the **X** box. Then click on **OK**. Print the plot by clicking on the graph and then clicking on the printer icon.

e) To make a residual plot of the fitted values versus the residuals, use the menu commands “Graph>Plot” to get a window. Place “Resi1” in the **Y** box and “Fits1” in the **X** box. Then click on **OK**. Print the plot by clicking on the graph and then clicking on the printer icon.

f) To save your *Minitab* data on your flash drive, use the menu commands “File>Save Current Worksheet as.” In the resulting dialog window, the top box says **Save in** and there is an arrow icon to the right of the top box. Click several times on the arrow icon until the **Save in** box reads “My com-
SAS Problems

Copy and paste the SAS programs for problems 2.17 and 2.18 from (http://lagrange.math.siu.edu/Olive/lreghw.txt), or enter the SAS program in Notepad or Word.

SAS is a statistical software package widely used in industry. You will need a flash drive. Referring to the program for Problem 2.17, the semicolon “;” is used to end SAS commands and the “options ls = 70;” command makes the output readable. (An “*” can be used to insert comments into the SAS program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into SAS. The command “data wcdata;” gives the name “wcdata” to the data set. The command “input x y;” says the first entry is variable x and the 2nd variable y. The command “cards;” means that the data is entered below. Then the data is entered and the isolated semicolon indicates that the last case has been entered. The command “proc print;” prints out the data. The command “proc corr;” will give the correlation between x and y. The commands “proc plot; plot y*x;” makes a scatterplot of x and y. The commands “proc reg; model y=x; output out = a p=pred r=resid;” tells SAS to perform a simple linear regression with y as the response variable. The output data set is called “a” and contains the fitted values and residuals. The command “proc plot data = a;” tells SAS to make plots from data set “a” rather than data set “wcdata.” The command “plot resid*(pred x);” will make a residual plot of the fitted values versus the residuals and a residual plot of x versus the residuals. The next plot command makes a response plot.

To use SAS on windows (PC), use the following steps.

i) Get into SAS, often by double clicking on an icon for programs such as a “Math Progs” icon and then double clicking on a SAS icon. If your computer does not have SAS, go to another computer.

ii) A window should appear with 3 icons. Double click on The SAS System for . . . .

iii) Like Minitab, a window with a split screen will open. The top screen says Log-(Untitled) while the bottom screen says Editor-Untitled1. Press the spacebar and an asterisk appears: Editor-Untitled1*.

2.17. a) Copy and paste the program for this problem from (http://lagrange.math.siu.edu/Olive/lreghw.txt), or enter the SAS program in Notepad or Word. The ls stands for linesize so l is a lowercase L, not the number one.

When you are done entering the program, you may want to save the program as h2d17.sas on your flash drive (J: drive, say). (On the top menu of the editor, use the commands “File > Save as.” A window will appear. Use
the upper right arrow to locate “Removable Disk (J:)” and then type the file name in the bottom box. Click on OK.)

b) Get back into SAS, and from the top menu, use the “File > Open” command. A window will open. Use the arrow in the upper right corner of the window to navigate to “Removable Disk (J:).” (As you click on the arrow, you should see My Documents, C: etc, then Removable Disk (J:).) Double click on h2d17.sas. (Alternatively cut and paste the program into the SAS editor window.) To execute the program, use the top menu commands “Run > Submit.” An output window will appear if successful.

If you were not successful, look at the log window for hints on errors. A single typo can cause failure. Reopen your file in Word or Notepad and make corrections. Occasionally you cannot find your error. Then find your instructor or wait a few hours and reenter the program.

c) To copy and paste relevant output into Word or Notepad, click on the output window and use the top menu commands “Edit > Select All” and then the menu commands “Edit > Copy.”

In Notepad use the commands “Edit > Paste.” Then use the mouse to highlight the relevant output. Then use the commands “Edit > Copy.”

Finally, in Word, use the command “Paste.” You can also cut output from Word and paste it into Notepad.

You may want to save your SAS output as the file HW2d17.doc on your flash drive.

d) To save your output on your flash drive, use the Word menu commands “File > Save as.” In the Save in box select “Removable Disk (J:)” and in the “File name box” enter HW2d17.doc. To get a Word printout, click on the printer icon or use the menu commands “File > Print.”

Save the output giving the least squares coefficients in Word.

e) Predict $Y$ if $X = 40$.

f) What is the residual when $X = 40$?

2.18. This problem shows how to use SAS for MLR. The data are from Kutner et al. (2005, problem 6.5). The response is “brand liking,” a measurement for whether the consumer liked the brand. The variable $X_1$ is “moisture content” and the variable $X_2$ is “sweetness.” Copy and paste the program for this problem from (http://lagrange.math.siu.edu/Olive/lreghw.txt).

a) Execute the SAS program and copy the output file into Notepad. Scroll down the output that is now in Notepad until you find the regression coefficients and ANOVA table. Then cut and paste this output into Word.

b) Do the 4 step ANOVA $F$ test.

You should scroll through your SAS output to see how it made the response plot and various residual plots, but cutting and pasting these plots is
Problems

2.13 Problems
tedious. So we will use Minitab to get these plots. Find the program for this problem from (http://lagrange.math.siu.edu/Olive/lreghw.txt). Then copy and paste the numbers (between "cards;" and the semicolon ":") into Minitab. Use the mouse commands "Edit>Paste Cells." This should enter the data in the Worksheet (bottom part of Minitab). Under C1 enter Y and under C2 enter X1 under C3 enter X2. Use the menu commands “Stat>Regression>Regression” to get a dialog window. Enter Y as the response variable and X1 and X2 as the predictor variable. Click on Storage then on Fits, Residuals, and OK OK.

c) To make a response plot, enter the menu commands “Graph>Plot” and place “Y” in the Y–box and “FITS1” in the X–box. Click on OK. Then use the commands “Edit>Copy Graph” to copy the plot. Include the plot in Word with the commands “Edit>Paste.” If these commands fail, click on the graph and then click on the printer icon.

d) Based on the response plot, does a linear model seem reasonable?

e) To make a residual plot, enter the menu commands “Graph>Plot” and place “RESI 1” in the Y–box and “FITS1” in the X–box. Click on OK. Then use the commands “Edit>Copy Graph” to copy the plot. Include the plot in Word with the commands “Edit>Paste.” If these commands fail, click on the graph and then click on the printer icon.

f) Based on the residual plot does a linear model seem reasonable?

Problems using ARC

To quit Arc, move the cursor to the x in the upper right corner and click.

Warning: Some of the following problems uses data from the book’s webpage (http://lagrange.math.siu.edu/Olive/lregbk.htm). Save the data files on a flash drive G, say. Get in Arc and use the menu commands “File>Load” and a window with a Look in box will appear. Click on the black triangle and then on Removable Disk (G:). Then click twice on the data set name.

2.19*. (Scatterplot in Arc.) Get cbrain.lsp as described above. (Activate the cbrain.lsp dataset with the menu commands “File>Load>Removable Disk (G:) > cbrain.lsp.”) Scroll up the screen to read the data description.

a) Make a plot of age versus brain weight brnweight. The commands “Graph&Fit > Plot of” will bring down a menu. Put age in the H box and brnweight in the V box. Put sex in the Mark by box. Click OK. Make the lowess bar on the plot read .1. Open Word.

In Arc, use the menu commands “Edit>Copy.” In Word, use the menu command “Paste.” This should copy the graph into the Word document.

b) For a given age, which gender tends to have larger brains?

c) At what age does the brain weight appear to be decreasing?
2.20. (SLR in Arc.) Activate `cbrain.lsp` as in Problem 2.19. Brain weight and the cube root of size should be linearly related. To add the cube root of size to the data set, use the menu commands “cbrain > Transform.” From the window, select `size` and enter $1/3$ in the `p:` box. Then click `OK`. Get some output with commands “Graph&Fit > Fit linear LS.” In the dialog window, put `brnweight` in `Response`, and $(size)^{1/3}$ in `terms`.

a) Cut and paste the output (from `Coefficient Estimates` to `Sigma hat`) into `Word`. Write down the least squares equation $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$.

b) If $(size)^{1/3} = 15$, what is the estimated `brnweight`?

c) Make a residual plot of the fitted values versus the residuals. Use the commands “Graph&Fit > Plot of” and put “L1:Fit-values” in `H` and “L1:Residuals” in `V`. Put sex in the `Mark by` box. Move the OLS bar to 1. Put the plot into `Word`. Does the plot look ellipsoidal with zero mean?

d) Make a response plot of the fitted values versus $Y = \text{brnweight}$. Use the commands “Graph&Fit > Plot of” and put “L1:Fit-values in `H` and `brnweight` in `V`. Put `sex` in `Mark by`. Move the OLS bar to 1. Put the plot into `Word`. Does the plot look linear?

2.21. In Arc enter the menu commands “File>Load>Data” and open the file `mussels.lsp`. This data set is from Cook and Weisberg (1999a).

The response variable $Y$ is the mussel muscle mass $M$, and the explanatory variables are $X_2 = S =$ shell mass, $X_3 = H =$ shell height, $X_4 = L =$ shell length, and $X_5 = W =$ shell width.

Enter the menu commands “Graph&Fit>Fit linear LS” and fit the model: enter $S, H, L, W$ in the “Terms/Predictors” box, $M$ in the “Response” box and click on `OK`.

a) To get a response plot, enter the menu commands “Graph&Fit>Plot of” and place `L1:Fit-Values` in the `H`-box and $M$ in the `V`-box. Copy the plot into `Word`.

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands “Graph&Fit>Plot of” and place `L1:Fit-Values` in the `H`-box and `L1:Residuals` in the `V`-box. Copy the plot into `Word`.

d) Based on the residual plot, what MLR assumption seems to be violated?

e) Include the regression output in `Word`. 
f) Ignoring the fact that an important MLR assumption seems to have been violated, do any of predictors seem to be needed given that the other predictors are in the model?


g) Ignoring the fact that an important MLR assumption seems to have been violated, perform the ANOVA $F$ test.

2.22. Get *cyp.lsp* as described above Problem 2.19. You can open the file in *Notepad* and then save it on a flash drive G, say, using the *Notepad* menu commands “File>Save As” and clicking the top checklist then click “Removable Disk (G:).” You could also save the file on the desktop, load it in *Arc* from the desktop, and then delete the file (sending it to the Recycle Bin).

a) In *Arc* enter the menu commands “File>Load>Removable Disk (G:)” and open the file *cyp.lsp*. This data set consists of various measurements taken on men from Cyprus around 1920. Let the response $Y = \text{height}$ and $X = \text{cephalic index} = 100(\text{head breadth})/(\text{head length})$. Use *Arc* to get the least squares output and include the relevant output in *Word*.

b) Intuitively, the cephalic index should not be a good predictor for a person’s height. Perform a 4 step test of hypotheses with $H_0: \beta_2 = 0$.

2.23. a) In *Arc* open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable $Y$ is \textit{height}, and the explanatory variables are a constant, $X_2 = \text{sternal height}$ (probably height at shoulder), and $X_3 = \text{finger to ground}$.

Enter the menu commands “Graph&Fit>Fit linear LS” and fit the model: enter \textit{sternal height} and \textit{finger to ground} in the “Terms/Predictors” box, \textit{height} in the “Response” box and click on OK.

Include the output in *Word*. Your output should certainly include the lines from “Response = height” to the ANOVA table.

b) Predict $Y$ if $X_2 = 1400$ and $X_3 = 650$.

c) Perform a 4 step ANOVA $F$ test of the hypotheses with $H_0: \beta_2 = \beta_3 = 0$.

d) Find a 99% CI for $\beta_2$.

e) Find a 99% CI for $\beta_3$.

f) Perform a 4 step test for $\beta_2 = 0$.

g) Perform a 4 step test for $\beta_3 = 0$.

h) What happens to the conclusion in g) if $\delta = 0.01$?
i) The Arc menu “L1” should have been created for the regression. Use the menu commands “L1>Prediction” to open a dialog window. Enter 1400 650 in the box and click on OK. Include the resulting output in Word.

j) Let $X_{f,2} = 1400$ and $X_{f,3} = 650$ and use the output from i) to find a 95% CI for $E(Y_f)$. Use the last line of the output, that is, $se = S(\hat{Y}_f)$.

k) Use the output from i) to find a 95% PI for $Y_f$. Now $se(\text{pred}) = s(\text{pred})$.

2.24. In Arc enter the menu commands “File>Load>Removable Disk (G:)” and open the file cyp.lsp (obtained as in Problem 2.22).

The response variable $Y$ is height, and the explanatory variables are $X_2 = \text{sternal height}$ (probably height at shoulder), and $X_3 = \text{finger to ground}$.

Enter the menu commands “Graph&Fit>Fit linear LS” and fit the model: enter sternal height and finger to ground in the “Terms/Predictors” box, height in the “Response” box and click on OK.

a) To get a response plot, enter the menu commands “Graph&Fit>Plot of” and place L1:Fit-Values in the H–box and height in the V–box. Copy the plot into Word.

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands “Graph&Fit>Plot of” and place L1:Fit-Values in the H–box and L1:Residuals in the V–box. Copy the plot into Word.

d) Based on the residual plot, does a linear model seem reasonable?

2.25. In Arc enter the menu commands “File>Load>Removable Disk (G:)” and open the file cyp.lsp (obtained as in Problem 2.22).

The response variable $Y$ is height, and the explanatory variables are $X_2 = \text{sternal height}$, $X_3 = \text{finger to ground}$, $X_4 = \text{bigonal breadth}$, $X_5 = \text{cephalic index}$, $X_6 = \text{head length}$, and $X_7 = \text{nasal height}$. Enter the menu commands “Graph&Fit>Fit linear LS” and fit the model: enter the 6 predictors (in order: $X_2$ 1st and $X_7$ last) in the “Terms/Predictors” box, height in the “Response” box and click on OK. This gives the full model. For the reduced model, only use predictors 2 and 3.

a) Include the ANOVA tables for the full and reduced models in Word.

b) Use the menu commands “Graph&Fit>Plot of…” to get a dialog window. Place L2:Fit-Values in the H–box and L1:Fit-Values in the V–box. Place the resulting plot in Word.
c) Use the menu commands “Graph&Fit>Plot of...” to get a dialog window. Place \textit{L2:Residuals} in the H–box and \textit{L1:Residuals} in the V–box. Place the resulting plot in \textit{Word}.

d) Both plots should cluster tightly about the identity line if the reduced model is about as good as the full model. Is the reduced model good?

e) Perform the 4 step partial \( F \) test (of Ho: the reduced model is good) using the 2 ANOVA tables from part a).

\textbf{2.26.} a) Activate the \texttt{cbrain.lsp} data set in \textit{ARC}. Fit least squares with \textit{age}, \textit{sex}, \textit{size}^{1/3}, and \textit{headht} as terms and \textit{brnweight} as the response. See Problem 2.20. Assume that the multiple linear regression model is appropriate. (This may be a reasonable assumption, 5 infants appear as outliers but the data set has hardly any cases that are babies. If \textit{age} was uniformly represented, the babies might not be outliers anymore.) Assuming that \textit{ARC} makes the menu “L1” for this regression, select “AVP-All 2D.” A window will appear. Move the OLS slider bar to 1 and click on the “zero line box.” The window will show the added variable plots for \textit{age}, \textit{sex}, \textit{size}^{1/3}, and \textit{headht} as you move along the slider bar that is below “case deletions.” Include all 4 added variable plots in \textit{Word}.

b) What information do the 4 plots give? For example, which variables do not seem to be needed?

(If it is clear that the zero and OLS lines intersect at the origin, then the variable is probably needed, and the point cloud should be tilted away from the zero line. If it is difficult to see where the two lines intersect since they nearly coincide near the origin, then the variable may not be needed, and the point cloud may not tilt away from the zero line.)

\textbf{R Problems}

Use the command \texttt{source(“G:/lregpack.txt“)} to download the functions and the command \texttt{source(“G:/lregdata.txt“)} to download the data. See Preface or Section 14.1. Typing the name of the \texttt{lregpack} function, e.g. \texttt{piplot}, will display the code for the function. Use the \texttt{args} command, e.g. \texttt{args(pisim)}, to display the needed arguments for the function. For some of the following problems, the \texttt{R} commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into \texttt{R}.

\textbf{2.27.} a) Download the data into \texttt{R} as described above.

For the Buxton (1920) data suppose that the response \( Y = \text{height} \) and the predictors were a constant, \textit{head length}, \textit{nasal height}, \textit{bigonal breadth}, and \textit{cephalic index}. There are 87 cases.

Type the following commands

\begin{verbatim}
zbux <- cbind(buxx,buxy)
zbux <- as.data.frame(zbux)
zfull <- lm(buxy~len+nasal+bigonal+cephalic,data=zbux)
\end{verbatim}
Multiple Linear Regression

```r
zred <- lm(buxy~len+nasal, data=zbux)
anova(zred, zfull)
```

b) Include the output in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu commands “Paste” in *Word* (or copy and paste the output: hit the *Ctrl* and *v* keys at the same time).

c) Use the output to perform the partial $F$ test where the full model is described in a) and the reduced model uses a constant, *head length*, and *nasal height*. The output from the `anova(zred, zfull)` command produces the correct partial $F$ statistic.

d) Use the following commands to make the response plot for the reduced model. Include the plot in *Word*.

```r
plot(zred$fit, buxy)
abline(0, 1)
```

e) Use the following command to make the residual plot for the reduced model. Include the plot in *Word*.

```r
plot(zred$fit, zred$resid)
```

f) The plots look bad because of 5 massive outliers. The following commands remove the outliers. Include the output in *Word*.

```r
zbux <- zbux[-c(60, 61, 62, 63, 64, 65),]
zfull <- lm(buxy~len+nasal+bigonal+cephalic, data=zbux)
zred <- lm(buxy~len+nasal, data=zbux)
anova(zred, zfull)
```

g) Redo the partial $F$ test.

h) Use the following commands to make the response plot for the reduced model without the outliers. Include the plot in *Word*.

```r
plot(zred$fit, zbux[,5])
abline(0, 1)
```

i) Use the following command to make the residual plot for the reduced model without the outliers. Include the plot in *Word*.

```r
plot(zred$fit, zred$resid)
```

j) Do the plots look ok?

2.28. Get the *R* commands for this problem. The data is such that $Y = 2 + x_2 + x_3 + x_4 + \epsilon$ where the zero mean errors are iid [exponential(2) - 2]. Hence the residual and response plots should show high skew. Note that $
\beta = (2, 1, 1, 1)^T$. The *R* code uses 3 nontrivial predictors and a constant, and the sample size $n = 1000$. 

a) Copy and paste the commands for part a) of this problem into $R$. Include the response plot in $Word$. Is the lowess curve fairly close to the identity line?

b) Copy and paste the commands for part b) of this problem into $R$. Include the residual plot in $Word$: press the $Ctrl$ and $c$ keys as the same time. Then use the menu command “Paste” in $Word$. Is the lowess curve fairly close to the $r = 0$ line?

c) The output out$\$coef gives $\hat{\beta}$. Write down $\hat{\beta}$. Is $\hat{\beta}$ close to $\beta$?

2.29. a) Download the $R$ functions $\text{piplot}$ and $\text{pisim}$ from $\text{lregpack}$.

b) The command $\text{pisim}(n=100, \text{type} = 1)$ will produce the mean length of the classical, semiparametric, asymptotically conservative, and asymptotically optimal PIs when the errors are normal, as well as the coverage proportions. Give the simulated lengths and coverages. Note: $\text{pimenlen}$ gives the four lengths (classical, semi, ac, aopt). Make table with headers classical, semi, ac, and aopt.

c) Repeat b) using the command $\text{pisim}(n=100, \text{type} = 3)$. Now the errors are $\text{EXP}(1) - 1$.

d) Download $\text{lregdata.txt}$ and type the command $\text{piplot(cbrainx,cbrainy)}$. This command gives the semiparametric PI limits for the Gladstone data. Include the plot in $Word$.

e) The infants are in the lower left corner of the plot. Do the PIs seem to be better for the infants or the bulk of the data. Explain briefly.

2.30. Use the function $\text{MLRsIm}$ as described in Rule of Thumb 2.1 to generate 10 pairs of response and residual plots. Right click Stop twenty times, and include the last plot in $Word$. 

Linear Regression
Olive, D.
2017, XIV, 494 p. 57 illus., Hardcover
ISBN: 978-3-319-55250-7