REGRESSION

Introduction

• “Data version” of best linear prediction.
• Very widely used.

Available data

• $Y_i =$ value of response variable for $i$th observation
• $X_{i1}, \ldots, X_{ip} =$ values of predictor variables 1 through $p$ for the $i$th observation
Goals:

• to understand how $Y$ is related to $X_1, \ldots, X_p$.
• to model the conditional expectation of $Y$ given $X_1, \ldots, X_p$
• to predict future $Y$ values from $X_1, \ldots, X_p$
Straight Line Regression

- only one predictor variable
- model is

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

- \( \beta_0 \) and \( \beta_1 \) are the unknown intercept and slope of the line
- \( \epsilon_1, \ldots, \epsilon_n \) are iid with mean 0 and constant variance \( \sigma^2 \)
- often the \( \epsilon_i \)'s are assumed to be normally distributed
Least-squares estimation
• least-squares estimate finds $b_0$ and $b_1$ to minimize

$$\sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_i)\right)^2$$
• using calculus, one can show that

\[ b_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \].

and

\[ b_0 = \bar{Y} - b_1 \bar{X} \]
• the least-squares line is

$$\hat{Y} = b_0 + b_1 X = \bar{Y} + b_1 (X - \bar{X})$$

$$= \bar{Y} + \left\{ \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\} (X - \bar{X})$$

$$= \bar{Y} + \frac{s_{xy}}{s_x^2} (X - \bar{X}),$$

where

$$s_{xy} = (n - 1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})$$

and $s_x^2$ is the sample variance of the $X_i$’s, that is,

$$s_x^2 = (n - 1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$
Exercise:

Show that if $\epsilon_1, \ldots, \epsilon_n$ are IID $N(0, \sigma^2)$ then the least-squares estimates of $\beta_0$ and $\beta_1$ are also the maximum likelihood estimates.

Example: Some data on weekly interest rates, from Jan 1, 1970 to Dec 31, 1993, were obtained from the Federal Reserve Bank of Chicago. The URL is:

http://www.chicagofed.org/economicresearchanddata/data/index.cfm
Change in "CM10 = 10-YEAR TREASURY CONSTANT MATURITY RATE (AVERAGE, NSA)" plotted against "AAA = MOODYS SEASONED CORPORATE AAA BOND YIELDS".
Regression

Regression Plot

aaa_diff = 0.0902266 + 0.572371 cm10_diff

S = 0.0633340  R-Sq = 69.5%  R-Sq(adj) = 68.5%

Fitted line plot from MINITAB.
Regression Analysis: aaa_diff versus cm10_diff

The regression equation is
aaa_diff = 0.0002266 + 0.572371 cm10_diff

S = 0.0633340  R-Sq = 68.5 %  R-Sq(adj) = 68.5%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>10.8950</td>
<td>10.8950</td>
<td>2716.14</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>1249</td>
<td>5.0100</td>
<td>0.0040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1250</td>
<td>15.9049</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fitted Line Plot: aaa_diff versus cm10_diff
Here is the same analysis using “regression” in MINITAB.

The first output is the estimated regression line:

The regression equation is

diff = 0.00023 + 0.572 diff

1251 cases used 1 cases contain missing values

Next comes the estimates, standard errors, T-statistics, and p-values:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.000227</td>
<td>0.001791</td>
<td>0.13</td>
<td>0.899</td>
</tr>
<tr>
<td>cm10_dif</td>
<td>0.57237</td>
<td>0.01098</td>
<td>52.12</td>
<td>0.000</td>
</tr>
</tbody>
</table>
The next output is $S = \text{estimate of } \sigma$, $R^2$ and adjusted $R^2$:

$S = 0.06333 \quad R\text{-Sq} = 68.5\% \quad R\text{-Sq(adj)} = 68.5\%$

Finally, the analysis of variance table is printed. This table decomposed the variability in $Y$ into several components:

<table>
<thead>
<tr>
<th>Analysis of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>DF</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual Error</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
• From the output we see that the least-squares estimates of the intercept and slope are 0.000227 and 0.572.

• The missing values created by differencing caused MINITAB to print a warning.
Standard errors, t-values, and p-values

Each of the coefficients in the MINITAB output has three other statistics associated with it:

- **SE = standard error**
  - This is estimated standard deviation of the least squares estimator and tells us the precision of that estimator.

- **t-value**
  - This is the t-statistic for testing that the coefficient is 0.
• p-value
  – This is the p-value for the test of the null hypothesis that the coefficient is 0 versus the alternative that it is not 0.
  – The p-value is 0.000 here.

• If the p-value is small as it is here, then this is evidence that the coefficient is not 0 which means that the predictor has some effect.
Analysis of variance, $R^2$, and F-tests

$$\text{total SS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$  

$$\text{regression SS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \text{ where } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\text{residual error SS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$  

$$\text{total SS} = \text{regression SS} + \text{residual error SS}.$$  

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = 1 - \frac{\text{residual error SS}}{\text{total SS}}$$
• The degrees of freedom for regression is $p = \text{number of predictor variables}$.
  – Note that $p$ is 1 for straight-line regression.

• The total degrees of freedom is $n - 1$.

• The residual error degrees of freedom is $n - p - 1$.

• The mean sum of squares (MS) for any source is its sum of squared divided by its degrees of freedom.

• The residual MS is an unbiased estimator of $\sigma^2$.

• The other means sum of squares are used for testing.
\[ F = \frac{\text{regression MS}}{\text{residual error MS}} \]

- The F-statistic tests null hypothesis that there is no linear relationship between any of the predictors and \( Y \).

- The entry in the column labeled “P” is the p-value of this test.

- In our example, the p-value is 0.000 which is very strong evidence against the null hypothesis.

- We conclude that there \textit{is} a relationship between changes in CM10 and changes in AAA.
Regression and best linear prediction

- Note the similarity between the best linear predictor
  \[ \hat{Y} = E(Y) + \frac{\sigma_{xy}}{\sigma_x^2} \{X - E(X)\}, \]
  and the least-squares line
  \[ \hat{Y} = \bar{Y} + \frac{s_{xy}}{s_x^2} (X - \bar{X}), \]
- The least-squares line is a sample version of the best linear predictor.
• $\rho_{XY}^2$, the squared correlation between $X$ and $Y$, is the fraction of variation in $Y$ that can be predicted using the linear predictor.

• The sample version of $\rho_{XY}^2$ is $R^2$. 
Multiple Linear Regression

• The multiple regression model is

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i. \]

• \( \beta_0 \) is the intercept.
  – It is the expected value of \( Y_i \) when all the \( X_{ij} \)'s are zero.

• The regression coefficients \( \beta_1, \ldots, \beta_p \) are the slopes.
  – \( \beta_j \) is the partial derivative of the expected response with respect to the \( j \)th predictor:

\[ \beta_j = \frac{\partial E(Y_i)}{\partial X_{ij}}. \]
• All coefficients are estimated by least-squares.

• **Example:** weekly interest rates data
  
  – now with the 30-year Treasury rates as a second predictor.

  – Thus \( p = 2 \).

• Here is the analysis using SAS.
options linesize = 72 ;
data WeeklyInterest ;
infile 'C:\book\SAS\WeeklyInterest.dat' ;
input month day year ff tb03 cm10 cm30 discount prime aaa ;
if lag(cm30) > 0 ;
aaa_dif = dif(aaa) ;
cm10_dif = dif(cm10) ;
cm30_dif = dif(cm30) ;
id = _N_ ;
run ;
title 'Weekly Interest Rates' ;
proc reg ;
model aaa_dif = cm10_dif cm30_dif / ss1 ss2 vif ;
output out=WeeklyInterest predicted=predicted rstudent=rstudent cookd=cookd h=leverage ;
run ;
proc gplot ;
plot rstudent*predicted ;
plot (rstudent cookd leverage cm10_dif cm30_dif)*id ;
plot cm10_dif*cm30_dif ;
run ;
The REG Procedure
Dependent Variable: aaa_dif

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>11.35366</td>
<td>5.67683</td>
<td>1357.95</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>876</td>
<td>3.66206</td>
<td>0.00418</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>878</td>
<td>15.01572</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.06466  R-Square 0.7561
Dependent Mean -0.00130  Adj R-Sq 0.7556
Coeff Var -4985.33904

Parameter Estimates

| Variable     | DF | Estimate   | Error     | t Value | Pr > |t| | Type I SS |
|--------------|----|------------|-----------|---------|------|---|---------|
| Intercept    | 1  | -0.00010686| 0.00218   | -0.05   | 0.9609| 0.00148|
| cm10_dif     | 1  | 0.36041    | 0.04456   | 8.09    | <.0001| 11.20585|
| cm30_dif     | 1  | 0.29655    | 0.04987   | 5.95    | <.0001| 0.14781|

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Type II SS</th>
<th>Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.00001004</td>
<td>0</td>
</tr>
<tr>
<td>cm10_dif</td>
<td>1</td>
<td>0.27353</td>
<td>14.03581</td>
</tr>
<tr>
<td>cm30_dif</td>
<td>1</td>
<td>0.14781</td>
<td>14.03581</td>
</tr>
</tbody>
</table>
Model Selection

- Model selection means selection of the predictor variables.

- Two principles to balance:
  - Larger models have less bias
    - they would give the best predictions if all coefficients could be estimated without error.
  - When unknown coefficients are replaced by estimates, the prediction become less accurate
    - this effect is worse when there are more coefficients to estimate.
• Thus, larger models have:
  – less bias (good)
  – more variability (bad)
• Caveat: do not use automatic model selection software blindly.

• $R^2$ not useful for comparing models of different sizes.
  – It always chooses the largest model

• The adjusted $R^2$ statistic can be used to select models.
• $C_p$ is a statistic that estimates how well a model will predict.
  – $C_p$ is closely related to the AIC statistic.
• Suppose there are $M$ predictors.
  – $\hat{\sigma}^2_M$ is the estimate of $\sigma^2$ using all of them.
  – $SSE(p)$ is the sum of squares for error for a model with only $p \leq M$ predictors.

• $n$ is the sample size

• Then $C_p$ is

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2_M} - n + 2(p + 1)$$
Now let’s add more potential predictor variables so there is a total of four. They are the changes in:

"FF = FEDERAL FUNDS RATE"

"CM10 = 10-YEAR TREASURY CONSTANT MATURITY RATE (AVERAGE, NSA)"

"CM30 = 30-YEAR TREASURY CONSTANT MATURITY RATE (AVERAGE, NSA)"

"PRIME = PRIME LENDING RATE CHARGED BY COMMERCIAL BANKS"
Variable Selection Using SAS

```sas
options linesize = 72 ;
data WeeklyInterest ;
infile 'C:\book\SAS\WeeklyInterest.dat' ;
input month day year ff tb03 cm10 cm30 discount prime aaa ;
if lag(cm30) > 0 ;
aaa_dif = dif(aaa) ;
cm10_dif = dif(cm10) ;
cm30_dif = dif(cm30) ;
ff_dif = dif(ff) ;
prime_dif = dif(prime) ;
run ;
title 'Weekly Interest Rates' ;
proc reg ;
model aaa_dif = cm10_dif cm30_dif ff_dif prime_dif / selection=rsquare adjrsq cp sbc aic ;
run ;
```

- At the beginning of the data set, all values of cm30 are zero.
- These data are actually missing, not zero.
- Treating them as zero, not missing, causes all sorts of bad things to happen. (Earlier results done in MINITAB had this problem.)
**Regression**

Dependent Variable: aaa_dif

R-Square Selection Method

<table>
<thead>
<tr>
<th>Number in Model</th>
<th>R-Square</th>
<th>Adjusted R-Square</th>
<th>C(p)</th>
<th>AIC</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7463</td>
<td>0.7460</td>
<td>35.4718</td>
<td>-4778.8055</td>
<td>-4769.24795</td>
</tr>
<tr>
<td>1</td>
<td>0.7379</td>
<td>0.7376</td>
<td>65.5166</td>
<td>-4750.2675</td>
<td>-4740.70994</td>
</tr>
<tr>
<td>1</td>
<td>0.0625</td>
<td>0.0615</td>
<td>2489.033</td>
<td>-3630.0113</td>
<td>-3620.45378</td>
</tr>
<tr>
<td>1</td>
<td>0.0320</td>
<td>0.0309</td>
<td>2598.720</td>
<td>-3601.8083</td>
<td>-3592.25074</td>
</tr>
<tr>
<td>2</td>
<td>0.7561</td>
<td>0.7556</td>
<td>2.1482</td>
<td>-4811.5872</td>
<td>-4797.25086</td>
</tr>
<tr>
<td>2</td>
<td>0.7463</td>
<td>0.7458</td>
<td>37.2491</td>
<td>-4777.0205</td>
<td>-4762.68417</td>
</tr>
<tr>
<td>2</td>
<td>0.7463</td>
<td>0.7457</td>
<td>37.4036</td>
<td>-4776.8714</td>
<td>-4762.53501</td>
</tr>
<tr>
<td>2</td>
<td>0.7391</td>
<td>0.7385</td>
<td>63.2227</td>
<td>-4752.2898</td>
<td>-4737.95341</td>
</tr>
<tr>
<td>2</td>
<td>0.7379</td>
<td>0.7373</td>
<td>67.5166</td>
<td>-4748.2675</td>
<td>-4733.93116</td>
</tr>
<tr>
<td>2</td>
<td>0.0727</td>
<td>0.0706</td>
<td>2454.497</td>
<td>-3637.6104</td>
<td>-3623.27404</td>
</tr>
<tr>
<td>3</td>
<td>0.7563</td>
<td>0.7555</td>
<td>3.5415</td>
<td>-4810.1968</td>
<td>-4791.08170</td>
</tr>
<tr>
<td>3</td>
<td>0.7562</td>
<td>0.7553</td>
<td>3.9224</td>
<td>-4809.8141</td>
<td>-4790.69896</td>
</tr>
<tr>
<td>3</td>
<td>0.7464</td>
<td>0.7455</td>
<td>39.0751</td>
<td>-4775.1885</td>
<td>-4756.07337</td>
</tr>
<tr>
<td>3</td>
<td>0.7392</td>
<td>0.7383</td>
<td>64.8002</td>
<td>-4750.6865</td>
<td>-4731.57137</td>
</tr>
<tr>
<td>4</td>
<td>0.7564</td>
<td>0.7553</td>
<td>5.0000</td>
<td>-4808.7412</td>
<td>-4784.84732</td>
</tr>
<tr>
<td>Number in Model</td>
<td>R-Square</td>
<td>Variables in Model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.7463</td>
<td>cm10_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.7379</td>
<td>cm30_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0625</td>
<td>ff_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0320</td>
<td>prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.7561</td>
<td>cm10_dif cm30_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.7463</td>
<td>cm10_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.7463</td>
<td>cm10_dif ff_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.7391</td>
<td>cm30_dif ff_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.7379</td>
<td>cm30_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0727</td>
<td>ff_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.7563</td>
<td>cm10_dif cm30_dif ff_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.7562</td>
<td>cm10_dif cm30_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.7464</td>
<td>cm10_dif ff_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.7392</td>
<td>cm30_dif ff_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.7564</td>
<td>cm10_dif cm30_dif ff_dif prime_dif</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Nonlinear Regression

Often we can derive a theoretical model but it is not linear.

**Example:** the price of par $1,000 zero-coupon bonds

- The owner of a bond will be paid $1,000 at maturity.
  - but no payments prior to maturity

- The price of a zero-coupon bond will always be less than par.

- Suppose that there are a variety of bonds with different maturities
  - the $i$th type of bond has maturity $T_i$. 
• Suppose all market participants agree that the bonds should pay interest at a continuously compounded rate \( r \).
  
  – We want to estimate \( r \) from the price data.

• Under this assumption, the present price of a bond with maturity \( T_i \) is

\[
P_i = 1,000 \exp(-rT_i).
\]

• There will be some random variation in the observed prices.

• Model:

\[
P_i = 1,000 \exp(-rT_i) + \epsilon_i.
\]
• An estimate of $r$ can be determined by minimizing

$$\sum_{i=1}^{n} \left\{ P_i - 1,000 \exp(-rT_i) \right\}^2.$$

• Finding the least-squares estimate requires solving nonlinear equations.

**Example:**

This example uses simulated data with $r = .06$. 
<table>
<thead>
<tr>
<th>Maturity</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>967.26</td>
</tr>
<tr>
<td>2.5</td>
<td>834.21</td>
</tr>
<tr>
<td>3</td>
<td>810.52</td>
</tr>
<tr>
<td>5</td>
<td>769.30</td>
</tr>
<tr>
<td>7.5</td>
<td>656.64</td>
</tr>
<tr>
<td>7.9</td>
<td>639.71</td>
</tr>
<tr>
<td>8</td>
<td>604.61</td>
</tr>
<tr>
<td>12</td>
<td>502.11</td>
</tr>
<tr>
<td>16</td>
<td>393.38</td>
</tr>
</tbody>
</table>
The data and the predicted price curve using nonlinear regression.
Nonlinear regression in SAS:

data bondprices ;
infile 'c:\courses\or473\data\bondprices.dat' ;
input maturity price ;
run ;
title 'Nonlinear regression using simulated zero-coupon bond data';
proc nlin ;
parm r=.02 to .09 by .005 ;
model price = 1000*exp(-r*maturity) ;
run ;
Here is the SAS output:

Grid Search

Dependent Variable price

<table>
<thead>
<tr>
<th>r</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0200</td>
<td>390066</td>
</tr>
<tr>
<td>0.0250</td>
<td>279853</td>
</tr>
<tr>
<td>0.0300</td>
<td>192505</td>
</tr>
<tr>
<td>0.0350</td>
<td>124990</td>
</tr>
<tr>
<td>0.0400</td>
<td>74665.1</td>
</tr>
<tr>
<td>0.0450</td>
<td>39230.4</td>
</tr>
<tr>
<td>0.0500</td>
<td>16679.7</td>
</tr>
<tr>
<td>0.0550</td>
<td>5263.9</td>
</tr>
<tr>
<td>0.0600</td>
<td>3456.9</td>
</tr>
<tr>
<td>0.0650</td>
<td>9926.6</td>
</tr>
<tr>
<td>0.0700</td>
<td>23509.9</td>
</tr>
<tr>
<td>0.0750</td>
<td>43191.1</td>
</tr>
</tbody>
</table>
The NLIN Procedure
Iterative Phase
Dependent Variable price
Method: Gauss-Newton

<table>
<thead>
<tr>
<th>Iter</th>
<th>$r$</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0600</td>
<td>3456.9</td>
</tr>
<tr>
<td>1</td>
<td>0.0585</td>
<td>3072.8</td>
</tr>
<tr>
<td>2</td>
<td>0.0585</td>
<td>3072.8</td>
</tr>
</tbody>
</table>

NOTE: Convergence criterion met
**Regression**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>4490011</td>
<td>4490011</td>
<td>11689.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>3072.8</td>
<td>384.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected Total</td>
<td>9</td>
<td>4493084</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>8</td>
<td>252587</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>0.0585</td>
<td>0.00149</td>
<td>0.0551</td>
</tr>
</tbody>
</table>
Residual Plotting

Goals: find problems with model or data

Problems to look for:

- nonnormality
  - outliers can be a problem since they have a large influence on the estimation results
  - a common solution is transformation of the response
• non-constant variance
  – causes inefficient (= too variable) estimates
  – transformation of the response and weighting are common solutions
• nonlinearity (if the model is linear) or, more generally, model misspecification

• **model misspecification** means $E(Y|X_1, \ldots, X_p)$ has a functional form different from the model
  – causes biased estimates
  – response transformation, polynomial regression, and nonparametric regression (splines, loess) are common solutions
We assume a general form of the regression model:

\[ Y_i = f(X_i; \beta) + \epsilon_i \]

**Predicted values:**

\[ \hat{Y}_i = f(X_i; \hat{\beta}) \]

**Residuals:**

\[ e_i = Y_i - \hat{Y}_i \]

**Studentized residuals:**

\[
\frac{e_i}{\text{se}(e_i)}
\]

Approximately \( N(0, 1) \), if modelling assumptions are true.
**Example:** Simulated data: $Y$, $X_1$, $X_2$
Indication of right skewness, but not severe.
Indication of model misspecification

Model misspecification may cause major problems - fix!
No problems evident
Variance appears nonconstant
Problem of nonconstant variance:

Example: Assume \( X_1, X_2, X_3 \) are independent, all with mean \( \mu \). \( \text{Var}(X_1) = \text{Var}(X_2) = 1 \) and \( \text{Var}(X_3) = 10 \).

\[
\text{Var}\{\left(\frac{X_1 + X_2 + X_3}{3}\right)\} = \frac{1 + 1 + 10}{9} = \frac{12}{9} = \frac{4}{3}.
\]

\[
\text{Var}\{\left(\frac{X_1 + X_2}{2}\right)\} = \frac{1 + 1}{4} = \frac{1}{2}.
\]

It seems that dropping \( X_3 \) is a good idea.

But this does not seem quite right. Why throw any data?
Most efficient weighted average uses the **inverse variances** as weights.

\[
\text{Var}\{(X_1 + X_2 + .1X_3)/2.1\}
= \{1 + 1 + (.1)^210\}/(2.1)^2 = \frac{1}{2.1} < \frac{1}{2}.
\]
Example: Estimation of Default Probabilities

Data:

- ratings: 1=Aaa (best), ..., 16=B3 (worse)
- default frequency (estimate of default probability)

Bluhm, C., Overbeck, L., and Wagner, C. (2003), *An Introduction to Credit Risk Modeling* (denoted BOW here)
1. **linear model (poor):**

   \[ \Pr(\text{default}|\text{rating}) = \beta_0 + \beta_1 \text{rating} \]

2. **nonlinear model (good):**

   \[ \Pr(\text{default}|\text{rating}) = \exp\{\beta_0 + \beta_1 \text{rating}\} \]

3. **loglinear model (transform of (2) – also good):**

   \[ \log\{\Pr(\text{default}|\text{rating})\} = \beta_0 + \beta_1 \text{rating} \] (used by BOW)
The default probability is given as a percentage.
Regression

![Graph showing the relationship between rating and default probability. The graph includes data points and lines for BOW and nonlinear models.]
Nonlinear regression residuals
Nonlinear regression residuals
• **nonlinear model:**
  \[ Pr(\text{default}|\text{rating}) = \exp\{\beta_0 + \beta_1 \text{rating}\} \]

• **linear/transformation model (BOW’s model):**
  \[ \log\{Pr(\text{default}|\text{rating})\} = \beta_0 + \beta_1 \text{rating} \]
  – **Problem:** cannot take logs of default frequencies that are 0

• **Transform-both-sides (TBS) model:**
  \[ Pr(\text{default}|\text{rating})^\alpha = \exp[\alpha\{\beta_0 + \beta_1 \text{rating}\}] \]
  – \( \alpha \) chosen by looking for good residual plots
  – \( \alpha = 1/2 \) works well
TBS fit compared to others
TBS residuals
TBS residuals
| method  | $\hat{Pr}\{\text{default} | \text{Aaa}\}$ | as percent of BOW |
|---------|---------------------------------|------------------|
| BOW     | .005%                           | 100%             |
| nonlinear | .002%                           | 40%              |
| TBS     | .0008%                          | 16%              |
How TBS effectively weights data
Taylor series (tangent line) linearization:

\[ H(\alpha) \approx H(\alpha_0) + H'(\alpha_0)(\alpha - \alpha_0) \]

for any differentiable \( H \) and any fixed \( \alpha_0 \).

Apply to

\[
\sum_{i=1}^{n} \left[ Y_i^\alpha - \{ f(X_i; \hat{\beta}) \}^\alpha \right]^2
\]

\[
\approx (\alpha)^2 \sum_{i=1}^{n} \left[ \frac{Y_i - f(X_i; \hat{\beta})}{\{ f(X_i; \hat{\beta}) \}^{-\alpha + 1}} \right]^2
\]

Most appropriate if

\[ \text{Var}(Y_i|X_i) \propto \{ f(X_i; \hat{\beta}) \}^{2(-\alpha + 1)}. \]
From previous slide: most appropriate if

$$\text{Var}(Y_i | X_i) \propto \{ f(X_i; \hat{\beta}) \}^{2(-\alpha+1)}.$$ 

**Example: Poisson**

- Assume $Y_i | X_i$ is Poisson with mean $f(X_i; \hat{\beta})$.
- Variance equals the mean for the Poisson distribution, so
  $$\text{Var}(Y_i | X_i) = f(X_i; \hat{\beta}).$$
- Thus, $\alpha = 1/2$ (use square root transformation).
TBS Regression in SAS

data DefaultProb ;
infile ’C:\or473\data\DefaultData.txt’ ;
input rating prob;
alpha = .5 ;
transprob = prob**alpha ;
run ;
proc nlin ;
parm beta0=-8 to -1 by 1 beta1=0 to 3 by .5 ;
model transprob = exp(alpha*(beta0 + beta1*rating)) ;
output out = outdata p=PredValue r=Residual ;
run ;
data outdata;
set outdata;
absresid = abs(Residual);
run;
proc gplot;
plot absresid*PredValue;
run;
proc univariate normal;
var Residual;
probplot;
run;
The NLIN Procedure
Dependent Variable SqrtProb
Method: Gauss-Newton

Iterative Phase

<table>
<thead>
<tr>
<th>Iter</th>
<th>beta0</th>
<th>beta1</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-6.0000</td>
<td>0.5000</td>
<td>1.7369</td>
</tr>
<tr>
<td>1</td>
<td>-8.0208</td>
<td>0.6710</td>
<td>0.3230</td>
</tr>
<tr>
<td>2</td>
<td>-7.7480</td>
<td>0.6480</td>
<td>0.2748</td>
</tr>
<tr>
<td>3</td>
<td>-7.7690</td>
<td>0.6493</td>
<td>0.2747</td>
</tr>
<tr>
<td>4</td>
<td>-7.7676</td>
<td>0.6492</td>
<td>0.2747</td>
</tr>
<tr>
<td>5</td>
<td>-7.7677</td>
<td>0.6492</td>
<td>0.2747</td>
</tr>
</tbody>
</table>

NOTE: Convergence criterion met.
## Regression

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>28.7353</td>
<td>14.3676</td>
<td>732.21</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>14</td>
<td>0.2747</td>
<td>0.0196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected Total</td>
<td>16</td>
<td>29.0100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>15</td>
<td>18.4804</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The NLIN Procedure

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Approximate 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>-7.7677</td>
<td>0.5182</td>
<td>-8.8791</td>
</tr>
<tr>
<td>beta1</td>
<td>0.6492</td>
<td>0.0346</td>
<td>0.5750</td>
</tr>
</tbody>
</table>
The UNIVARIATE Procedure

Variable: RESIDUAL

Moments

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>16</td>
<td>Sum Weights</td>
<td>16</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.0194237</td>
<td>Sum Observations</td>
<td>-0.3107791</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>0.1338344</td>
<td>Variance</td>
<td>0.01791165</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.27064364</td>
<td>Kurtosis</td>
<td>-1.1282057</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>0.27471116</td>
<td>Corrected SS</td>
<td>0.26867468</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>-689.02644</td>
<td>Std Error Mean</td>
<td>0.0334586</td>
</tr>
</tbody>
</table>
### Basic Statistical Measures

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.01942</td>
</tr>
<tr>
<td>Median</td>
<td>-0.03524</td>
</tr>
<tr>
<td>Mode</td>
<td>.</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>0.13383</td>
</tr>
<tr>
<td>Variance</td>
<td>0.01791</td>
</tr>
<tr>
<td>Range</td>
<td>0.42499</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>0.24725</td>
</tr>
</tbody>
</table>
Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>--Statistic---</th>
<th>-----p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>$W$ 0.937735</td>
<td>$Pr &lt; W$ 0.3222</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>$D$ 0.151914</td>
<td>$Pr &gt; D$ &gt;0.1500</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>$W-Sq$ 0.063363</td>
<td>$Pr &gt; W-Sq$ &gt;0.2500</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>$A-Sq$ 0.402689</td>
<td>$Pr &gt; A-Sq$ &gt;0.2500</td>
</tr>
</tbody>
</table>
VIF’s

```
proc reg;
model aaa_dif=cm10_dif cm30_dif ff_dif prime_dif/vif;
run;
```

\[ VIF = \text{Variance Inflation Factor} \]
### Parameter Estimates

| Variable   | DF | Estimate | Standard Error | t Value | Pr > |t| | Type I SS |
|------------|----|----------|----------------|---------|-------|-----|--------------|
| Intercept  | 1  | -0.00010103 | 0.00218 | -0.05 | 0.9631 | 0.00148 |
| cm10_dif   | 1  | 0.35510   | 0.04517 | 7.86 | <.0001 | 11.20585 |
| cm30_dif   | 1  | 0.30093   | 0.05010 | 6.01 | <.0001 | 0.14781 |
| ff_dif     | 1  | 0.00531   | 0.00553 | 0.96 | 0.3371 | 0.00254 |
| prime_dif  | 1  | -0.00788  | 0.01071 | -0.74 | 0.4620 | 0.00227 |

### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Type II SS</th>
<th>Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0.00000897</td>
<td>0</td>
</tr>
<tr>
<td>cm10_dif</td>
<td>1</td>
<td>0.25860</td>
<td>14.41205</td>
</tr>
<tr>
<td>cm30_dif</td>
<td>1</td>
<td>0.15096</td>
<td>14.15236</td>
</tr>
<tr>
<td>ff_dif</td>
<td>1</td>
<td>0.00386</td>
<td>1.19941</td>
</tr>
<tr>
<td>prime_dif</td>
<td>1</td>
<td>0.00227</td>
<td>1.14743</td>
</tr>
</tbody>
</table>
Influence Diagnostics and Leverage

Linear Regression Example
Three tools for diagnosing problems due to leverage points:

• leverages

• RSTUDENT

• Cook’s D
\[ \hat{Y}_i = \sum_{j=1}^{n} H_{ij} Y_j. \]

\( H_{ij} \) is a somewhat complex function of the \( X \)-variables

- If there is a single \( X \)-variable, then

\[ H_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n - 1)s_x^2} \]

where \( s_x^2 \) is the sample variance of the \( X \)'s.

\( H_{11}, \ldots, H_{nn} \) are the leverages

The standard error of \( \hat{\epsilon}_i \) is \( s\sqrt{1 - H_{ii}} \)
\[
\text{RSTUDENT} = \frac{\hat{\epsilon}_i}{\{s_{(-i)} \sqrt{1 - H_{ii}}\}}
\]
Cook’s $D = \sum_{j=1}^{n} \left\{ \hat{Y}_j - \hat{Y}_j(-i) \right\}^2 \frac{1}{(p + 1)s^2}$. 
Linear Regression Example
Weekly Interest Rates Example
Example of leverage in two dimensions
Example of leverage in two dimensions
Modeling Interest Rate Volatility

Source: FRED
Regression

Daily change in rate

Year

Regression
Regression

Squared change in rate vs. price
Power-of-mean model:

- $R_t = \text{rate at time } t$
- $\Delta(R_t) = R_t - R_{t-1}$

- From plots:

\[ E\{\Delta(R_t)\} \approx 0 \]

and

\[ \text{Var}\{\Delta(R_t)\} \approx \beta_0 R_{t-1}^{\beta_1} \]

- Suggest model

\[ \{\Delta(R_t)\}^2 = \beta_0 R_{t-1}^{\beta_1} + \text{noise} \]
options linesize 72;
data OneYrTreasury;
infile 'C:\Documents and Settings\David Ruppert\My Documents\talks\TempleStatFin\price_data.txt';
input prices delta_prices;
delta_prices2 = delta_prices**2;
run;
proc nlin;
parm beta0 = .0001 to .1 by .01 beta1 = 0 to 4 by .25;
model delta_prices2 = beta0*prices**beta1;
run;

Note: prices = $R_{t-1}$ and delta_prices = $R_t - R_{t-1}$
The NLIN Procedure  
Dependent Variable delta_prices2  
Method: Gauss-Newton

Iterative Phase

<table>
<thead>
<tr>
<th>Iter</th>
<th>beta0</th>
<th>beta1</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000100</td>
<td>2.2500</td>
<td>16.2995</td>
</tr>
<tr>
<td>1</td>
<td>0.000069</td>
<td>2.3929</td>
<td>16.2746</td>
</tr>
<tr>
<td>2</td>
<td>0.000053</td>
<td>2.4994</td>
<td>16.2373</td>
</tr>
<tr>
<td>3</td>
<td>0.000035</td>
<td>2.6647</td>
<td>16.2181</td>
</tr>
<tr>
<td>4</td>
<td>0.000022</td>
<td>2.8533</td>
<td>16.1830</td>
</tr>
<tr>
<td>5</td>
<td>0.000011</td>
<td>3.1487</td>
<td>16.1165</td>
</tr>
<tr>
<td>6</td>
<td>0.000013</td>
<td>3.1657</td>
<td>15.9323</td>
</tr>
<tr>
<td>7</td>
<td>0.000014</td>
<td>3.1602</td>
<td>15.9316</td>
</tr>
<tr>
<td>8</td>
<td>0.000013</td>
<td>3.1611</td>
<td>15.9316</td>
</tr>
<tr>
<td>9</td>
<td>0.000013</td>
<td>3.1610</td>
<td>15.9316</td>
</tr>
</tbody>
</table>

NOTE: Convergence criterion met.
Approximate Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>beta0</th>
<th>beta1</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>1.0000000</td>
<td>-0.9960570</td>
</tr>
<tr>
<td>beta1</td>
<td>-0.9960570</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>
How do we define a “residual”?

Recall model:

$$\text{Var}\{\Delta(R_t)\} = E\{(\Delta(R_t))^2\} = \beta_0 R_t^{\beta_1}$$

(Assuming $E\{\Delta(R_t)\} = 0$.)

First attempt:

$$\{\Delta(R_t)\}^2 - \beta_0 R_t^{\beta_1}$$

- This does NOT work.
- Has mean zero but not a constant variance
Second attempt:

\[
\frac{\{\Delta(R_t)\}^2}{\beta_0 R_t^{\beta_1}}
\]

- This DOES work.
- Has constant mean and variance (if model is correctly specified)
- We can check model by plotting this variable against rate and year – we want to see NO pattern
Regression

Squared change in rate / $\left( \beta_0 \text{ rate}^{\beta} \right)$

Year
Regression

Squared change in rate / \((\beta_0 \cdot \text{rate} \cdot \beta_1)\)
The model has a misspecification problem.

- It does not fit the data well for the period around 2003 when interest rates were at their lows values during the 1962 to 2006 period.
Statistics and Finance
An Introduction
Ruppert, D.
2004, XXII, 474 p., Hardcover