Everything You Need To Know About the Linear Model

JEFF GILL
Professor of Political Science
Professor of Biostatistics
Professor of Surgery (Public Health Sciences)
Washington University, St. Louis
Why We Need to Start with the Linear Model For This Course

- Start with a standard linear model specification indexed by subjects and a first level of grouping, the *context* level.

- Now use a single explanatory variable that has the form:

\[ y_i = \beta_{j0[i]} + \beta_{j1[i]}X_i + \epsilon_i. \]

- Add a second level to the model that explicitly nests effects within groups and index these groups \( j = 1 \) to \( J \):

\[ \beta_{j0} = \gamma_{00} + \gamma_{10}Z_{j0} + u_{j0} \]
\[ \beta_{j1} = \gamma_{01} + \gamma_{11}Z_{j1} + u_{j1}, \]

where all individual level variation is assigned to groups producing department level residuals: \( u_{j0} \) and \( u_{j1} \).

- These \( Z_{j} \) are group-level variables in that their effect is assumed to be measured at the aggregated rather than at the individual level.
Why We Need to Start with the Linear Model For This Course

- The two-level model is produced by inserting the context level specifications into the original linear expression for the outcome variable of interest:

\[ y_i = \gamma_{00} + \gamma_{01}X_i + \gamma_{10}Z_{j0} + \gamma_{11}X_iZ_{j1} + u_{j1}X_i + u_{j0} + \epsilon_i. \]

- This equation shows that the composite error structure, \( u_{j1}X_i + u_{j0} + \epsilon_i \), is now clearly heteroscedastic since it is conditioned on levels of the explanatory variable, causing additional estimation complexity.

- Notice that there is an “automatic” interaction component: \( \gamma_{11}X_iZ_{j1} \).

- Now we are going model distributions for \( y, \beta_{j0}, \) and \( \beta_{j1} \).

- Thus it is important to review the linear regression model in some detail.
Precursor To Linear Models: Following Trends

- Sometimes trends are obvious and easy to follow in data, but often they are not.
- Two standard tools: smoothing and linear regression.
- Usually one or the other is appropriate.
- Smoothers simply follow the trends in the data, with a given smoothing parameter.
- Main smoother: lowess, “locally weighted running line smoother.”

- Is it possible for a linear model result to look like it fits when it is the wrong specification?
Running Lowess

```r
x <- seq(1, 25, length=600)
y <- (2/(pi*x))^(0.5)*(1-cos(x)) + rnorm(100, 0, 1/10)
summary(lm(y~x))$coef

     Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.538054  0.019850 27.105 3.8061e-106
          x  -0.018702  0.001347 -13.884 3.4425e-38

postscript("Class.Multilevel/trends1.ps")
par(mar=c(3,3,2,2), bg="white")
plot(x,y,pch="+")
ols.object <- lm(y~x)
abline(ols.object,col="blue")
lo.object <- lowess(y~x,f=2/3)
lines(lo.object$x,lo.object$y,lwd=2,col="red")
lo.object <- lowess(y~x,f=1/5)
lines(lo.object$x,lo.object$y,lwd=2,col="purple")
dev.off()
```
Running Lowess
What Does the Linear Model Get You? An Example

▸ Consider a study of anaemia in women in a given clinic, perhaps in St. Louis, where 20 cases are chosen at random from the full study to get the data here.

▸ From a blood sample we get:
  ▶ haemoglobin level (Hb) in grams per deciliter (12–15 g/dl is normal in adult females)
  ▶ packed cell volume (hematocrit) in percent (38% to 46% is normal in adult females)

▸ We also have:
  ▶ age in years
  ▶ menopausal status (0 = no, 1 = yes)

▸ There is an obvious endogeneity problem in modeling Hb(g/dl) versus PCV(%).
## What Does the Linear Model Get You? An Example

<table>
<thead>
<tr>
<th>Subject</th>
<th>Hb(g/dl)</th>
<th>PCV(%)</th>
<th>Age</th>
<th>Menopausal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.1</td>
<td>35</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>10.7</td>
<td>45</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>12.4</td>
<td>47</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>14.0</td>
<td>50</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>13.1</td>
<td>31</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>10.5</td>
<td>30</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>9.6</td>
<td>25</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>12.5</td>
<td>33</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>13.5</td>
<td>35</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>13.9</td>
<td>40</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>15.1</td>
<td>45</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>13.9</td>
<td>47</td>
<td>49</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>16.2</td>
<td>49</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>16.3</td>
<td>42</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>16.8</td>
<td>40</td>
<td>57</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>17.1</td>
<td>50</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>16.6</td>
<td>46</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>16.9</td>
<td>55</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>15.7</td>
<td>42</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>16.5</td>
<td>46</td>
<td>67</td>
<td>1</td>
</tr>
</tbody>
</table>
Scatterplot of the Anaemia Data
Scatterplot of the Anaemia Data
Scatterplot of the Anaemia Data

postscript("Class.PreMed.Stats/Images/anaemia2.fig.ps")
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray", cex.lab=1.3)
plot(anaemia$PCV[anaemia$Menapause==0],anaemia$Hb[anaemia$Menapause==0],
    pch=19,col="yellow", xlim=range(anaemia$Age),ylim=range(anaemia$Hb),
    xlab="PCV (Menapausal in Red)",ylab="Hb(g/dl)"
points(anaemia$PCV[anaemia$Menapause==1],anaemia$Hb[anaemia$Menapause==1],
    pch=19,col="red")
dev.off()
What Does the Linear Model Get You? An Example

anaemia <- read.table("http://jgill.wustl.edu/data/anaemia.dat",
header=TRUE,row.names=1)
a.lm.out <- lm(Hb ~ Age + PCV, data=anaemia)
summary(a.lm.out)

Residuals:
    Min  1Q Median   3Q  Max
-1.600 -0.676  0.216  0.558  1.759

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.2388   1.2064  4.34  0.00044
   Age         0.1104   0.0164  6.74 3.5e-06
   PCV         0.0971   0.0326  2.98  0.00847

Residual standard error: 0.979 on 17 degrees of freedom
Multiple R-squared: 0.851,  Adjusted R-squared: 0.834
F-statistic: 48.6 on 2 and 17 DF,  p-value: 9.26e-08
What Happens When it Doesn’t Work?


- 24 countries: average survey review of restaurant service quality and a tipping index from three travel etiquette web sites.

The data:

<table>
<thead>
<tr>
<th>Country</th>
<th>Quality</th>
<th>Tip</th>
<th>Country</th>
<th>Quality</th>
<th>Tip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan</td>
<td>4.4</td>
<td>0.00</td>
<td>Thailand</td>
<td>3.9</td>
<td>0.03</td>
</tr>
<tr>
<td>Canada</td>
<td>3.7</td>
<td>0.16</td>
<td>New_Zealand</td>
<td>3.7</td>
<td>0.07</td>
</tr>
<tr>
<td>UAE</td>
<td>3.6</td>
<td>0.10</td>
<td>Germany</td>
<td>3.6</td>
<td>0.08</td>
</tr>
<tr>
<td>USA</td>
<td>3.6</td>
<td>0.18</td>
<td>South_Africa</td>
<td>3.5</td>
<td>0.11</td>
</tr>
<tr>
<td>Australia</td>
<td>3.4</td>
<td>0.08</td>
<td>Argentina</td>
<td>3.4</td>
<td>0.10</td>
</tr>
<tr>
<td>Morocco</td>
<td>3.4</td>
<td>0.07</td>
<td>Turkey</td>
<td>3.4</td>
<td>0.08</td>
</tr>
<tr>
<td>India</td>
<td>3.3</td>
<td>0.10</td>
<td>Brazil</td>
<td>3.3</td>
<td>0.07</td>
</tr>
<tr>
<td>Vietnam</td>
<td>3.2</td>
<td>0.05</td>
<td>England</td>
<td>3.2</td>
<td>0.10</td>
</tr>
<tr>
<td>Greece</td>
<td>3.2</td>
<td>0.08</td>
<td>Spain</td>
<td>3.1</td>
<td>0.08</td>
</tr>
<tr>
<td>France</td>
<td>3.1</td>
<td>0.08</td>
<td>Italy</td>
<td>3.0</td>
<td>0.07</td>
</tr>
<tr>
<td>Egypt</td>
<td>3.0</td>
<td>0.08</td>
<td>Mexico</td>
<td>3.0</td>
<td>0.13</td>
</tr>
<tr>
<td>China</td>
<td>2.9</td>
<td>0.03</td>
<td>Russia</td>
<td>1.7</td>
<td>0.10</td>
</tr>
</tbody>
</table>
What Happens When it Doesn’t Work?

```r
service <- read.table("http://jgill.wustl.edu/data/service.dat", 
header=TRUE,row.names=1)
service.lm <- lm(Quality ~ Tip, data=service)
source("./Class.MLE/graph.summary.R")
graph.summary(service.lm)

Family: gaussian
Link function: identity

<table>
<thead>
<tr>
<th>Coef</th>
<th>Std.Err.</th>
<th>0.95 Lower</th>
<th>0.95 Upper</th>
<th>CIs: ZE+RO</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.495</td>
<td>0.244</td>
<td>3.018</td>
<td>3.973</td>
</tr>
<tr>
<td>Tip</td>
<td>-2.113</td>
<td>2.632</td>
<td>-7.272</td>
<td>3.046</td>
</tr>
</tbody>
</table>

N: 24       Estimate of Sigma: 0.485
```
What Happens When it Doesn’t Work?

postscript("Class.Multilevel/Images/tipping.ps",height=5,width=7)
par(mfrow=c(1,1),mar=c(5,5,2,2),lwd=2,col.axis="white",col.lab="white",col.sub="white",col="white",bg="slategray", cex.lab=1.3)
# PLOT POINTS AND REGRESSION LINES
plot(service$Tip, service$Quality, pch="+",xlab="Tip",ylab="Quality")
abline(service.lm,col="gold3",lwd=5)
abline(h=service.lm$coef[1],col="gold3",lty=3,lwd=2)
# ADD CONFIDENCE BOUNDS AT THREE LEVELS
ruler.df <- data.frame(Tip = seq(-0.1, 2,length=200))
for (k in c(0.99,0.95,0.90)) {
  confidence.interval <- predict(service.lm, ruler.df, interval="confidence", level=k)
  lines(ruler.df[,1],confidence.interval[,2],col="peachpuff",lwd=0.75)
  lines(ruler.df[,1],confidence.interval[,3],col="peachpuff",lwd=0.75)
}
# IDENTIFY POTENTIAL OUTLIERS
text(0.113,1.7,"Russia")
text(0.011,4.38,"Japan")
dev.off()
What Happens When it Doesn’t Work?
How Influential is Japan?

```r
barplot(cooks.distance(service.lm))
```
Correlation and Regression

- Also, regression is correlation, since:

\[
\text{cor}(X, Y) = \frac{s_X}{s_Y} \beta
\]

- From our anaemia example picking Age:

```r
coef(a.lm.out)[2]
0.1342515
apply(anaemia,2,sd)
Hb       PCV        Age       Menopause
2.4018852 7.8958683 15.7366485 0.5129892

cor(anaemia[,1],anaemia[,3])
[1] 0.8795875

(15.7366485/2.4018852) * 0.1342515
[1] 0.8795877
```
When Not To Use Correlation
Correlation: Tests of Significance

- Hypotheses: $H_0: \rho = 0$ versus $H_1: \rho \neq 0$.

- Test statistic:

  $$ t = \frac{r}{SE(r)}, \quad SE(r) = \sqrt{\frac{1 - r^2}{n - 2}}. $$

- HB and PCV from the anaemia data:

  $$ r = 0.6733745 \quad SE(r) = \sqrt{\frac{1 - 0.6733745^2}{20 - 2}} = 0.1742551 \quad t = 3.864304 \quad p \approx 0.001. $$

- HB and Age from the anaemia data:

  $$ r = 0.8795875 \quad SE(r) = \sqrt{\frac{1 - 0.6733745^2}{20 - 2}} = 0.1121323 \quad t = 7.844191 \quad p \approx 0.0001. $$
Data Structures in Matrix/Vector Form

► The vector $Y$ contains values of the outcome variable in a column vector:

$$Y = [y_1, y_2, \ldots, y_n]'$$

► The matrix $X$ contains the explanatory variables down the columns:

$$X = \begin{bmatrix}
1 & x_{11} & x_{12} & \ldots & x_{1k} \\
1 & x_{21} & x_{22} & \ldots & x_{2k} \\
1 & x_{31} & x_{32} & \ldots & x_{3k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \ldots & x_{nk} 
\end{bmatrix},$$

with a leading column of 1s (e.g. there are $k-1$ explanatory variables).

► The vector $e$ contains values of the observed residuals (disturbances, errors) in a column vector:

$$e = [e_1, e_2, \ldots, e_n]'$$
Gauss-Markov Assumptions for Classical Linear Regression

- **Functional Form:** \( \mathbf{Y}_{(n \times 1)} = \mathbf{X}_{(n \times k)} \mathbf{\beta}_{(k \times 1)} + \mathbf{\epsilon}_{n \times 1} \) (recall \( \mathbf{X} \) has a leading column of 1’s)

- **Mean Zero Errors:** \( \mathbf{E}[\mathbf{\epsilon}] = \mathbf{0} \)

- **Homoscedasticity:** \( \text{Var}[\mathbf{\epsilon}] = \sigma^2 \mathbf{I} \)

- **Non-Correlated Errors:** \( \text{Cov}[\epsilon_i, \epsilon_j] = 0, \quad \forall i \neq j \)

- **Exogeneity of Explanatory Variables:** \( \text{Cov}[\epsilon_i, \mathbf{X}] = 0, \quad \forall i \)

Note that every one of these lines has \( \mathbf{\epsilon} \) in it, meaning that these are assumptions about the underlying population values.
Other Considerations

**Requirements:**

▷ conformability of matrix/vector objects

▷ $\mathbf{X}$ has full rank $k$, so $\mathbf{X}'\mathbf{X}$ is invertible (non-zero determinant, nonsingular)

▷ identification condition: not all points lie on a vertical line.

**Freebee:** eventual normality... $\mathbf{\epsilon}|\mathbf{X} \sim N(0, \sigma^2 \mathbf{I})$.

**Toughness:** the linear model is both *robust* to minor violations of the Gauss-Markov assumptions and *resistant* to outlying values.
Estimation With OLS:

Define the following function:

\[
S(\beta) = \epsilon' \epsilon \\
= (Y - X\beta)'(Y - X\beta) \\
= \begin{bmatrix} Y'Y \\ X'Y \end{bmatrix} - \begin{bmatrix} 2Y'X \beta \\ \beta'X'X\beta \end{bmatrix} \\
(1 \times n)(n \times 1) - (1 \times n)(n \times k)(k \times 1) + (1 \times k)(k \times n)(n \times k)(k \times 1)
\]

Take the derivative of \( S(\beta) \) with respect to \( \beta \):

\[
\frac{\partial}{\partial \beta} S(\beta) = 0 - 2 \begin{bmatrix} X'Y \\ \beta'X'X\beta \end{bmatrix} \equiv 0, \\
(1 \times n)(n \times 1) - (1 \times n)(n \times k)(k \times 1) + (1 \times k)(k \times n)(n \times k)(k \times 1)
\]

(think about what sign you would get by taking another derivative).

So there exists a (minimizing) solution at some value \( \hat{\beta} \) (or notationally \( b \)) of \( \beta \): \( X'Xb = X'Y \) which is the Normal Equation.

Premultiplying the Normal Equation by \( (X'X)^{-1} \), gives: \( \hat{\beta} = (X'X)^{-1}X'Y \), where we can call \( \hat{\beta} \) as \( b \) for notational convenience (this is where the requirement for \( X'X \) to be nonsingular comes in).
Estimation With MLE:

- Assume: $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(0, \Sigma)$, $\Sigma = \sigma^2 I$
  (notice that normality is an added assumption here, since MLE calculations require a distribution to work with).

- The likelihood function for iid $\boldsymbol{\epsilon}$:
  \[
  L(\boldsymbol{\epsilon}) = (2\pi \sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma^2} \boldsymbol{\epsilon}' \boldsymbol{\epsilon} \right].
  \]

- Plug-in: $\epsilon_i = y_i - \mathbf{X}_i\beta$ (where $\beta$ to be estimated):
  \[
  L(\beta) = (2\pi \sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \right].
  \]

- Which in log-likelihood form is:
  \[
  \ell(\beta) = -\frac{n}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta).
  \]
Estimation With MLE:

Now take the first derivative with respect to $\beta$ and set it equal to zero:

$$
\frac{\partial}{\partial \beta} \ell(\beta) = -\frac{1}{2\sigma^2}(-X)'(2)(Y - X\beta) \equiv 0 \\
X'(Y - X\beta) = 0 \\
X'Y - X'X\beta = 0 \\
X'X\beta = XY \quad \text{the "normal" equation} \\
\hat{\beta} = (X'X)^{-1}X'Y
$$

We can also take the first derivative with respect to $(\sigma^2)$ and set it equal to zero:

$$
\frac{\partial}{\partial \sigma^2} \ell(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) \\
= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}e'e \equiv 0 \\
0 = -n\sigma^2 + e'e \\
\hat{\sigma}^2 = \frac{e'e}{n}
$$

which is slightly biased in finite samples for $e'e/(n - k)$, more on this later.
Implications

- Normal Equation: \( X'Xb - X'Y = -X'(Y - Xb) = -X'e \equiv 0 \) (by assumption)
- Summation of errors: \( \sum e_i \approx 0 \)
- The regression hyperplane passes through the mean vectors: \( \bar{Y} = \bar{X}b \)
- Equivalence of means: \( \text{mean}(\hat{Y}) = \text{mean}(Y) \)
- The hat matrix with rank and trace \( k \) (\( H, P \), or \( I - M \)) starts with:

\[
e = Y - Xb
= Y - X((X'X)^{-1}X'Y)
= Y - (X(X'X)^{-1}X')Y
= Y - HY
= (I - H)Y
= MY
\]

where \( M \) and \( H \) are symmetric and idempotent.

- For example: \( H\cdot H = (X(X'X)^{-1}X')(X(X'X)^{-1}X') = X[(X'X)^{-1}XX'(X'X)^{-1}X'] = X(X'X)^{-1}X' \).
The HAT Matrix

The name is because $\hat{Y} = Xb = X((X'X)^{-1}X'Y) = (X(X'X)^{-1}X')Y = HY$, but “projection matrix” ($P$) is better for geometric reasons.
The HAT Matrix

▶ Related properties of interest:

▷ $\mathbf{I} - \mathbf{M} = \mathbf{P}$, $\mathbf{I} - \mathbf{P} = \mathbf{M}$

▷ $\mathbf{PX} = \mathbf{X}$ (an orthogonal projection onto $\mathbf{X}$)

▷ $\mathbf{PM} = \mathbf{MP} = \mathbf{0}$ and $\mathbf{P(} \mathbf{I} - \mathbf{P}) = \mathbf{0}$ (orthogonality)

▷ $\mathbf{e'e} = \mathbf{Y'M'MY} = \mathbf{Y'MY} = \mathbf{Ye}$ (sum of squares)

▷ $\mathbf{Y'Y} = \mathbf{Y'\hat{Y}} + \mathbf{e'e}$ (the Pythagorean Theorem!)
Fit & Decomposition, Illustration

![Graph showing linear regression with Age on the x-axis and Haemoglobin Level on the y-axis. The graph includes a line of best fit, and markers for Case Number 7, Regression Improvement, and Regression Error. The mean of Haemoglobin Level is also indicated.](image-url)
Fit & Decomposition, Variability Definitions

- **Sum of Squares Total**, all the variability to obtain over the mean estimate,

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

- **Sum of Squares Regression**, the variability accounted for by the regression,

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

- **Sum of Squares Error**, the remaining variability not accounted for by the regression,

\[
SSE = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2
\]
Fit & Decomposition, Total Variability

- Interesting manipulations of the sum of squares total:

\[
\text{SST} = \sum_{i=1}^{n} (Y_i^2 - 2Y_i \bar{Y} + \bar{Y}^2) \\
= \sum_{i=1}^{n} Y_i^2 - 2 \sum_{i=1}^{n} Y_i \bar{Y} + n \bar{Y}^2 \\
= \sum_{i=1}^{n} Y_i^2 - 2n \bar{Y}^2 + n \bar{Y}^2 \\
= \sum_{i=1}^{n} Y_i^2 - n \bar{Y}^2 \quad \text{(scalar description)} \\
= \left[ Y'Y - \frac{1}{n} Y'JY \right] \quad \text{(matrix algebra description)}
\]

where \( J \) is a \( n \times n \) matrix of all 1’s.

- Note that pre-multiplying by \( J \) produces a same-sized matrix where the values are the sum by column, and post-multiplying by \( J \) produces a same-sized matrix where the values are the sum by row.
A Small Demonstration of the $J$ Matrix

Y <- c(1,3,5)
J <- matrix(rep(1,9),ncol=3)
J

[,1] [,2] [,3]
[1,] 1 1 1
[2,] 1 1 1
[3,] 1 1 1

3*(mean(Y))^2
[1] 27

t(Y) %*% J %*% Y/3
[1,] 27

Demonstrating the last line from scalar to matrix form

$$
\sum_{i=1}^{n} Y_i^2 - n \bar{Y}^2 = Y'Y - \frac{1}{n} Y'JY.
$$
Fit & Decomposition, Regression Variability

► Sum of Squares Regression:

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

\[ = \hat{Y}'\hat{Y} - 2\bar{Y} \sum_{i=1}^{n} \hat{Y}_i + n\bar{Y}^2 \]

\[ = (b'X')(\hat{Y}) - 2n\bar{Y}^2 + n\bar{Y}^2 \]

\[ = b'X'\hat{Y} - n\bar{Y}^2 \]

\[ = b'X'\hat{Y} - \frac{1}{n} Y'JY \]
Fit & Decomposition, Remaining Variability

- Sum of Squares Error (using the Normal Equation, \( X'Xb = X'Y \)):

\[
SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = e'e
\]
\[
= (Y - Xb)'(Y - Xb)
\]
\[
= Y'Y - Y'Xb - b'X'Y + b'X'Xb
\]

now do the Normal Equation substitution…

\[
= Y'Y - (X'Xb)'b - b'(X'Xb) + b'X'Xb
\]
\[
= Y'Y - (b'X')(Xb)
\]
\[
= Y'Y - b'X'\hat{Y}
\]
Total %*^#%ing Magic!

Adding Total Sum of Squares Regression to Total Sum of Squares Error:

\[
SSR + SSE = (b'X'\hat{Y} - nY'JY) + (Y'Y - b'X'\hat{Y})
= Y'Y - nY'JY \\
= SST
\]

Because in general sums of squares do not equal squares of sums, for example:

\[
7^2 + 3^2 = 58, \quad (7+3)^2 = 100
\]

(except in unusual or pathological circumstances).
A Measure of Fit

- The “R-Square” or “R-Squared” measure:

\[ R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{e'e}{Y'M^oY} = \frac{b'X'M^oXb}{Y'M^oY} \]

where \( M^o = I - \frac{1}{n}ii' \), \( i = c(1, 1, \ldots, 1) \).

- Note: \( M^o \) is idempotent and transforms means to deviances for the explanatory variables:

\[
M.0 \leftarrow \text{diag}(3) - \frac{1}{3}c(1,1,1)\%\%t(c(1,1,1))
\]

\[
M.0
\begin{array}{c c c}
[1,] & 0.66667 & -0.33333 & -0.33333 \\
[2,] & -0.33333 & 0.66667 & -0.33333 \\
[3,] & -0.33333 & -0.33333 & 0.66667
\end{array}
\]

- Also, there is another version that accounts for sample size and the number of explanatory variables \( k \):

\[ R^2_{adj} = 1 - \frac{e'e/(n-k)}{Y'M^oY/(n-1)} \]
Warnings about $R^2$

- it is not actually a statistic, it is a measure
- therefore it does not have a distribution
- there is not population analog
- it can never be reduced by adding more explanatory variables
- it is a quadratic form in $[0 : 1]$ space
- therefore it does not have quite the meaning that people expect.
Properties of the Estimator, Unbiasedness

- $\mathbf{b}$ is an estimator for $\mathbf{\beta}$, which can be rewritten:
  $$
  \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}
  = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{\beta} + \mathbf{\epsilon})
  = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon}
  = \mathbf{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon}
  $$

  which also implies $\mathbf{b} - \mathbf{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon}$.

- Taking expectations:
  $$
  E[\mathbf{b}] = E[\mathbf{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon}]
  = E[\mathbf{\beta}] + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon}]
  = \mathbf{\beta} + E[\mathbf{K}\mathbf{\epsilon}]
  = \mathbf{\beta} + \mathbf{K}E[\mathbf{\epsilon}] = \mathbf{\beta}
  $$

  shows that it is unbiased.
Properties of the Estimator, Variance

- By definition (using an inner product):

\[
\text{Var}[\mathbf{b} | \mathbf{X}] = E[(\mathbf{b} - \mathbf{\beta})(\mathbf{b} - \mathbf{\beta})' | \mathbf{X}] - E[(\mathbf{b} - \mathbf{\beta} | \mathbf{X})]^2 \\
= E[(\mathbf{b} - \mathbf{\beta})(\mathbf{b} - \mathbf{\beta})' | \mathbf{X}] - E[0]^2.
\]

(using the elementary property \( \text{Var}[A] = E[A^2] - (E[A])^2 \)).

- Now using \( \mathbf{b} - \mathbf{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon} \) from the previous slide,

\[
\text{Var}[\mathbf{b} | \mathbf{X}] = E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon})' | \mathbf{X}] \\
= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\epsilon}\mathbf{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}] \\
= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{\epsilon}\mathbf{\epsilon}' | \mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\text{Var}[\mathbf{\epsilon} | \mathbf{X}] + E[\mathbf{\epsilon} | \mathbf{X}]^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\]
Properties of the Estimator, General

- The OLS estimator is consistent (converges in probability):
  \[
  \lim_{n \to \infty} \mathbb{P}(n^{-1/2} (b - \beta)) = 0,
  \]
  with Gauss-Markov assumption #5: \( \text{Cov} [\epsilon_i, X] = 0, \forall i \), and there is no perfect multicollinearity.

- The OLS estimate is optimal:
  \[
  \text{Var}(b_{\text{OLS}}) \leq \text{Var}(b_{\text{All Other}})
  \]
  with Gauss-Markov assumptions #3 \( \text{Var} [\epsilon] = \sigma^2 I \), and #4 \( \text{Cov} [\epsilon_i, \epsilon_j] = 0, \forall i \neq j \).

- Given all of the Gauss-Markov assumptions and \( \sigma^2 < \infty \), we say that \( b \) is BLUE for \( \beta \) if calculated from OLS or MLE.

- Given sufficient sample size \( b|X \sim N(\beta, \sigma^2 (X'X)^{-1}) \).
Dealing With the Variance

- We want to use some \( \text{Est.} \text{Var}[b] \), since \( \sigma^2 \) is generally not known.

- What about calculating the standard error of the regression?

- We will use the sample quantities for estimation:

\[
E[e'e|X] = \text{Var}[e|X] + E[e|X]^2 = \text{Var}[e|X] + 0 = \sigma^2 I
\]

(Using the elementary property \( \text{Var}[A] = E[A^2] - (E[A])^2 \)) meaning that we only need to manipulate \( E[e'e|X] \) to get an estimate of \( \sigma^2 \).

- Perspective here: \( X \) is fixed once observed and \( e \) is the random variable (to be estimated with \( e \)):
  - since \( \text{Var}[e] = \sigma^2 I \), then no single term dominates and we get the Lindberg-Feller CLT result,
  - so \( e \) (the sample quantity) is IID normal and we write the joint PDF as:

\[
f(e) = \prod_{i=1}^{n} f(e_i) = (2\pi \sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{e'e}{2\sigma^2}\right]
\]

based on sample quantities.
Estimating From Sample Quantities

- Population derived variance/covariance matrix: $\text{Var}[b] = \sigma^2(X'X)^{-1}$.

- We also know: $E[e_i] = \epsilon_i$, although in practice we always have finite samples.

- And by assumption: $E[e_i^2] = \text{Var}[e_i] + (E[e_i])^2 = \sigma^2$
  (again using the elementary property $\text{Var}[A] = E[A^2] - (E[A])^2$).

- Therefore the sum is $\sum E[e_i^2] = \text{tr}(\sigma^2 I) = n\sigma^2$.

- So why not use: $\hat{\sigma}^2 \approx \frac{1}{n} \sum e_i^2$.

- But:
  
  $e_i = Y_i - X_i b$  
  (now insert population values for $Y_i$)

  $= (X_i'\beta + \epsilon_i) - X_i b$

  $= \epsilon_i - X_i(b - \beta)$

  meaning that $\text{plim}[e_i] = \epsilon_i$ since $b \xrightarrow{n \to \infty} \beta$.

- So asymptotically this substitution is fine, but is it okay in finite samples?
Some Needed Relations

Recall that:

\[ M = I - H = I - X(X'X)^{-1}X'. \]

So that we can derive this in the opposite direction from the way we did before:

\[ MY = (I - X(X'X)^{-1}X')Y \]
\[ = Y - X(X'X)^{-1}X'Y \]
\[ = Y - Xb \]
\[ = e \]

From similar calculations we get:

\[ Me = (I - X(X'X)^{-1}X')e \]
\[ = e - X(X'X)^{-1}X'e \]
\[ = e - X(X'X)^{-1}X'(Y - Xb) \]
\[ = e - X(X'X)^{-1}X'Y + X(X'X)^{-1}X'Xb \]
\[ = e - Xb + Xb \]
\[ = e \]
Some Needed Relations

- Equating $MY = e$ and $Me = e$ from the last slide, we get $MY = Me$

- We could also get this from the corresponding population values:

$$
MY = M[X\beta + \epsilon] \\
= MX\beta + Me \\
= (I - X(X'X)^{-1}X')X\beta + Me \\
= X\beta - X(X'X)^{-1}X'X\beta + Me \\
= Me
$$

- So $e'e = (Me)'Me = e'M'Me = e'Me$. 
Estimating From Sample Quantities

This means that we can use $e'e = e'MM'e = e'Me$ accordingly:

$$E[e'e|X] = E[e'Me|X]$$

$$= E[tr(e'Me)|X]$$

$$= tr(ME[e'e|X])$$

$$= tr(M)I\sigma^2$$

$$= tr(I - H)I\sigma^2$$

$$= [tr(I_{n\times n}) - tr((X'X)^{-1}X')\sigma^2$$

$$= [tr(I_{n\times n}) - tr(X(X'X)^{-1}X')\sigma^2$$

$$= [tr(I_{n\times n}) - k]\sigma^2$$

$$= [n - k]\sigma^2$$

(Gauss-Markov assumption #4: $\text{Cov}[\epsilon_i, \epsilon_j] = 0, \forall i \neq j$)

(property of traces: $\text{tr}(ABC) = \text{tr}(BAC)$)

($M$ is fixed for observed $X$)

(Gauss-Markov assumption #3: $\text{Var}[\epsilon] = \sigma^2I$)

(property of traces: $\text{tr}(A - B) = \text{tr}(A) - \text{tr}(B)$)

(property of traces: $\text{tr}(AB) = \text{tr}(BA)$)

(for linear models the trace of the hat matrix is the of rank $X$)
Estimating From Sample Quantities

From $E[e'e|X] = (n - k)\sigma^2$, we algebraically get an unbiased estimator:

$$\hat{\sigma}^2 = \frac{e'e}{n - k} = s^2,$$

so that a finite sample estimator of $\text{Var}[\hat{b}] = \sigma^2(X'X)^{-1}$ is:

$$\hat{\text{Var}}[\hat{b}] = s^2(X'X)^{-1}.$$

The Wald-style traditional linear inference, for the $k$th coefficient is:

$$z_k = \frac{b_k - \beta_{\text{null}}}{\sqrt{\sigma^2(X'X)^{-1}_k}} \sim N(0, 1),$$

with the assumption that we know $\sigma^2$ (which we usually do not).

But we can use a well-known distributional relation to modify the above form:

$$\text{if } X^2 = \frac{(n - k)s^2}{\sigma^2} \sim \chi_{n-k}^2 \text{ then } \frac{z_k}{\sqrt{X^2/df}} \sim t_{(n-k)}(0)$$

provided the random variables $z_k$ and $X^2$ are independent.
Estimating From Sample Quantities

- Making the obvious substitution gives:

\[ t_{(n-k)} = \frac{b_k - \beta_k^{null}}{\sqrt{\frac{\sigma^2}{\mathbf{X}'\mathbf{X}}}} \times \sqrt{\frac{\frac{(n-k)s^2}{\sigma^2}}{(n-k)}} = \frac{b_k - \beta_k^{null}}{\sqrt{\frac{s^2}{\mathbf{X}'\mathbf{X}}}} \]

- Typical (Wald) regression test:

\[ H_0: \beta_k = 0 \quad \quad H_1: \beta_k \neq 0 \]

making:

\[ t_{(n-k)} = \frac{b_k - \beta_k^{null}}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}}} = \frac{b_k}{SE(\beta_k)} \]

- Alternatives usually look like:

\[ H_0: \beta_k < 7 \quad \quad H_1: \beta_k \geq 7 \]

making:

\[ t_{(n-k)} = \frac{b_k - 7}{SE(\beta_k)} \]
Summary Statistics

\( (1 - \alpha) \) Confidence Interval for \( b_k \):

\[
[b_k - SE(b)t_{\alpha/2,df} : b_k + SE(b)t_{\alpha/2,df}]
\]

\( (1 - \alpha) \) Confidence Interval for \( \sigma^2 \):

\[
\left[ \frac{(n - k)s^2}{\chi^2_{1-\alpha/2}} : \frac{(n - k)s^2}{\chi^2_{\alpha/2}} \right]
\]

F-statistic test for all but \( b_0 \) equal to zero:

\[
F = \frac{SSR/(k - 1)}{SSE/(n - k)} \sim F_{k-1,n-k} \text{ under the null.}
\]
Linear Model Confidence Bands

We want the predicted value of the outcome variable for \( x_i \) in the sample:

\[
y_i = x_i \beta + \epsilon_i \\
\hat{y}_i = x_i b
\]

The variance at this point on the regression line, bivariate, is:

\[
\text{Var}_b|x_i = s^2 \left( \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum(x_j - \bar{x})} \right).
\]

The variance at this point on the regression line, multivariate, is:

\[
\text{Var}[ e_i | X, x_i ] = \text{Var}[ x_i(\beta - b) ] \\
= \text{Var}[ x_i \beta ] + \text{Var}[ x_i b ] \\
= s^2 + x_i \text{Var}[ b ] x_i' \\
= s^2 + x_i (s^2 (XX)^{-1}) x_i' \\
= s^2 [1 + x_i (X'X)^{-1} x_i']
\]
Linear Regression Basics, Jeff Gill [50]

**Linear Model Predictions/Forecasts**

- We want the predicted value for $\mathbf{x}^0$ *not in the sample*:
  
  $$
  y^0 = x^0 \beta + \epsilon^0 \quad \hat{y}^0 = x^0 \mathbf{b}
  $$

  since $\hat{y}^0$ is the LMVUE of $E[\hat{y}^0|\mathbf{x}^0]$.

- The *prediction error* is:
  
  $$
  e^0 = y^0 - \hat{y}^0 = x^0(\beta - \mathbf{b}) = \epsilon^0.
  $$

  (notationally suppressing the conditionality on $\mathbf{X}$ here).

- The Prediction variance is:
  
  $$
  \text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = s^2 + \text{Var}[x^0(\beta - \mathbf{b})|\mathbf{X}, \mathbf{x}^0] = s^2 + s^2(x^0)(\mathbf{X}'\mathbf{X})^{-1}(x^0)'
  $$

  and if we have a constant term in the regression, this is:

  $$
  \text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = s^2 \left[ 1 + \frac{1}{n} + \sum_{j=1}^{K-1} \sum_{k=1}^{K-1} (x^0_j - \bar{x}_j)(x^0_k - \bar{x}_k)(\mathbf{X}_{-1}\mathbf{M}^0\mathbf{X}_{-1})^{jk} \right],
  $$

  where $\mathbf{X}_{-1}$ is $\mathbf{X}$ omitting the first column, $K$ is the number of explanatory variables (including the constant), and $\mathbf{M}^0 = \mathbf{I} - \frac{1}{n} \mathbf{i}\mathbf{i}'$.  

Linear Model Predictions/Forecasts

- The prediction interval (in the vertical direction) is created from

\[ CI[\hat{y}^0] = \hat{y}^0 \pm t_{\alpha/2} \sqrt{\text{Var}[e^0|X, x^0]} \] .

- Note that the value of \( x^0 \) is buried in there, and like the CI for \( \beta \), it is smallest around \( \bar{x} \).

- It is important to also distinguish between two interval estimates around the regression line: the CI for \( \hat{y} = X\beta \) and the CI for \( \hat{y}^0 \).

- Where the prediction interval is always wider than the regression confidence interval.
Linear Model Predictions/Forecasts
Linear Model Predictions/Forecasts

The R code for these intervals can be produced by:

```r
postscript("Class.Multilevel/linear.prediction.ps")
X <- rnorm(25,3,1); Y <- X + rnorm(25,2,2)
ruler <- data.frame(X = seq(-3, 8,length=200))
confidence.interval <- predict(lm(Y ~ X), ruler, interval="confidence")
predict.interval <- predict(lm(Y ~ X), ruler, interval="prediction")
par(mar=c(1,1,1,1),oma=c(3,3,1,1),mfrow=c(1,1),col.axis="white",col.lab="white",
    col.sub="white",col="white",bg="slategray")
# REGRESSION LINE
plot(ruler[,1], confidence.interval[,1], type="l",lwd=4,ylim=c(-9,16),col="black")
# UPPER AND LOWER CONFIDENCE INTERVALS
lines(ruler[,1],confidence.interval[,2], lwd=2, lty=2, col="aquamarine3")
lines(ruler[,1],confidence.interval[,3], lwd=2, lty=2, col="aquamarine3")
# UPPER AND LOWER PREDICTION INTERVALS
lines(ruler[,1],predict.interval[,2], lwd=3, lty=3, col="rosybrown2")
lines(ruler[,1],predict.interval[,3], lwd=3, lty=3, col="rosybrown2")
segments(mean(X),-10,mean(X),mean(Y), lwd=0.5, col="peachpuff")
segments(-3,mean(Y),mean(X),mean(Y), lwd=0.5, col="peachpuff")
devo.off()
```
Multicollinearity Issues

- If one explanatory variable is a linear combination of another then \( \text{rank}(X) = k - 1 \).
- Therefore \( \text{rank}(X'X) = k - 1 \) (matrix size \( k \times k \)), and it is singular and non-invertible.
- Now no parameter estimates are possible, and model is now unidentified.

- More typically: 2 explanatory variables are highly but not perfectly correlated.

- Symptoms:
  - small changes in data give large changes in parameter estimates.
  - coefficients have large standard errors and poor \( t \)-statistics even if F-statistics and \( R^2 \) are okay.
  - coefficients seem illogical (obviously wrong sign, huge magnitude)
Multicollinearity Remedies

- Respecify model (if reasonable).
- Center explanatory variables, or standardize.
- Create a new variable that is a weighted combination of highly correlated variables and use it to replace both (two variables to one variable in the model).
- Ridge regression (add a little bias):

\[ \mathbf{b} = \left( \mathbf{X}'\mathbf{X} + \mathbf{RI} \right)^{-1}\mathbf{X}'\mathbf{Y} \]

such that the \([\ ]\) part barely inverts, and can involve a penalty function.

- R packages that do this: ridge, bigRR, genridge, parcor, and more.

- See also: Jeff Gill and Gary King (SMR 2004), “What to do When Your Hessian is Not Invertible: Alternatives to Model Respecification in Nonlinear Estimation.”
Simple Ridge Regression Example

```r
anaemia <- read.table("http://jgill.wustl.edu/data/anaemia.dat",
                   header=TRUE,row.names=1)
library(MASS)
a.lm3.out <- lm.ridge(Hb ~ Age + Menapause + I(Age+rnorm(nrow(anaemia))],
                      data=anaemia)
cbind(a.lm3.out$coef, sqrt(a.lm3.out$scales))
[,1] [,2]
Age  1.06565 3.91640
Menapause  0.29350 0.70711
I(Age + rnorm(nrow(anaemia)))  0.73817 3.89488

summary(lm(Hb ~ Age + Menapause,data=anaemia))$coef[2:3,1:2]
  Estimate Std. Error
Age   0.11716   0.035881
Menapause  0.60002  1.100703
```
Summary of Asymptotic Results, Meeting the “Grenander Conditions:”

G1: For each column of $\mathbf{X}$: $\mathbf{X}'_k \mathbf{X}_k \longrightarrow +\infty$: sums of squares grow as $n$ grows, no columns of all zeros.

G2: No single observation dominates each explanatory variable $k$ in the limit:

$$\lim_{n \to \infty} \frac{\mathbf{X}^2_{ik}}{\mathbf{X}'_k \mathbf{X}_k} = 0, \quad i = 1, \ldots, n, \ i \neq k$$

G3: $\mathbf{R}$ is the sample correlation matrix of the observed columns of $\mathbf{X}$, excluding the leading column of 1s. Then $\lim_{n \to \infty} \mathbf{R} = \mathbf{C}$, where $\mathbf{C}$ is a positive definite matrix ($\mathbf{q}' \mathbf{X} \mathbf{q} > 0$ for any conformable, non-null $\mathbf{q}$).

► Now G1: + G2: + G3: give:

$$\mathbf{b} \overset{\text{asym.}}{\sim} N \left[ \beta, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right]$$

where:

$$\mathbf{Q} = \lim_{n \to \infty} \left[ \frac{1}{n} \mathbf{X}' \mathbf{X} \right].$$

► See: Grenander and Rosenblatt (1957).
In the Limit What About $s^2$

Recall:

$$s^2 = \frac{1}{n-k} \mathbf{e}' \mathbf{M} \mathbf{e} = \frac{1}{n-k} \mathbf{e}' (\mathbf{I} - \mathbf{H}) \mathbf{e} = \frac{1}{n-k} \left[ \mathbf{e}' \mathbf{e} - \mathbf{e}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{e} \right]$$

$$= \frac{n}{n-k} \left[ \frac{\mathbf{e}' \mathbf{e}}{n} - \left( \frac{\mathbf{e}' \mathbf{X}}{n} \right) \left( \frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{X}' \mathbf{e}}{n} \right) \right]$$

where $n/(n-k)$ goes to 1 as $n$ goes to $\infty$.

Taking the limit now as $n \to \infty$:

$$\lim s^2 = \lim \frac{\mathbf{e}' \mathbf{e}}{n} - (0) \mathbf{Q}^{-1}(0) = \lim \frac{1}{n} \sum_{i=1}^{n} \mathbf{e}_i^2 = \frac{1}{n} \text{tr}(\sigma^2 \mathbf{I}) = \frac{1}{n} (n \sigma^2) = \sigma^2$$

Summarizing:

$$\lim s^2 \left[ \frac{\mathbf{X}' \mathbf{X}}{n} \right]^{-1} = \sigma^2 \mathbf{Q}^{-1}$$

$$\lim s^2 (\mathbf{X}' \mathbf{X})^{-1} = \frac{\sigma^2}{n} \mathbf{Q}^{-1} = \frac{\sigma^2}{n} \left[ \frac{1}{n} \mathbf{X}' \mathbf{X} \right]^{-1}$$

$\therefore \text{Est.Asy.Var}[\mathbf{b}] = \sigma^2 (\mathbf{X}' \mathbf{X})^{-1}$
Testing Linear Restrictions

- A theory has *testable implications* if it implies some testable restrictions on the model definition:

\[ H_0: \beta_k = 0 \quad \text{versus} \quad H_1: \beta_k \neq 0, \]

for example.

- Most restrictions involve *nested* parameter spaces:

  - unrestricted: \([\beta_0, \beta_1, \beta_2, \beta_3]\)
  - restricted: \([\beta_0, 0, \beta_2, \beta_3]\)

  although the restriction does not have to be \(\beta_1 = 0\).

- Note that *non*-nested comparisons cause problems for *non*-Bayesians.

- Likelihood-based non-nested comparisons require use of a “super model.”
Testing Linear Restrictions

Continuing with the simple example:

unrestricted: \([\beta_0, \beta_1, \beta_2, \beta_3]\) 
restricted: \([\beta_0, 0, \beta_2, \beta_3]\)

This example can be notated with \(R = [0, 1, 0, 0]\) to indicate the location of the restriction, and \(q = 0\) to indicate the value of the restriction, so that \(R\beta = q\) gives the full specification of the restriction in linear algebra terms:

\[
\begin{bmatrix}
0, 1, 0, 0
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix}
= 0 \times \beta_0 + 1 \times \beta_1 + 0 \times \beta_2 + 0 \times \beta_3 = \beta_1 = q
\]

which forces the restriction.

The test statistic that we will build here, after estimating the regression model, has tail-values that indicate that the restriction is not supported: “it modifies the model more than the data wants”.

Testing Linear Restrictions

More generally we express these restrictions as:

\[
\begin{align*}
    r_{11} \beta_1 + r_{12} \beta_2 + \ldots + r_{1k} \beta_k &= q_1 \\
    r_{21} \beta_1 + r_{22} \beta_2 + \ldots + r_{2k} \beta_k &= q_2 \\
    \vdots \\
    r_{j1} \beta_1 + r_{j2} \beta_2 + \ldots + r_{jk} \beta_k &= q_j
\end{align*}
\]

or in more succinct matrix algebra form:

\[
R \beta = q
\]

or in more succinct matrix algebra form: \( R \beta = q \).

Notes:

- Each row of \( R \) is one restriction.
- \( J < k \) for \( R \) to be full rank.
- This setup imposes \( J \) restrictions on \( k \) parameters, so there are \( k - J \) free parameters left.
- We are still assuming that \( \epsilon_i \sim N(0, \sigma^2) \).

General test where tail values reject the restriction set:

\[
H_0: R \beta - q = 0 \quad \quad H_1: R \beta - q \neq 0
\]
Testing Linear Restrictions, Examples

- One of the coefficients is zero, $\beta_j = 0$, $J = 1$:

$$R = [0, 0, 0, 1, \ldots, 0, 0, 0], \quad q = 0$$

- Two coefficients are equal, $\beta_j = \beta_k$, $J = 2$:

$$R = [0, 0, 0, 1, \ldots, -1, \ldots, 0, 0, 0], \quad q = 0$$

- First three coefficients are zero, $J = 3$:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \end{bmatrix} = [I_3 : 0], \quad q = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$
Testing Linear Restrictions, Examples

- A set of coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$:
  \[
  \mathbf{R} = [0, 1, 1, 1, 0, \ldots, 0], \quad q = 1
  \]

- Several restrictions, $\beta_2 + \beta_3 = 4$, $\beta_1 + \beta_6 = 0$, $\beta_5 = 9$:
  \[
  \mathbf{R} = \begin{bmatrix}
  0 & 1 & 1 & 0 & 0 & 0 \\
  0 & 0 & 0 & 1 & 0 & 1 \\
  0 & 0 & 0 & 0 & 1 & 0
  \end{bmatrix}, \quad q = \begin{bmatrix} 4 \\ 0 \\ 9 \end{bmatrix}
  \]

- All coefficients except the constant are zero:
  \[
  \mathbf{R} = [0:1], \quad q = [0]
  \]
Testing Linear Restrictions, Examples

Create the *discrepancy vector* dictated by the null hypothesis:

\[ \mathbf{Rb} - \mathbf{q} = \mathbf{m} \approx 0, \]

which asks whether \( \mathbf{m} \) is sufficiently different from zero. Note that \( \mathbf{m} \) is linear function of \( \mathbf{b} \) and therefore also normally distributed (\( \mathbf{R} \) and \( \mathbf{q} \) are full of constants we set).

This makes it straightforward to think about:

\[
E[\mathbf{m} | \mathbf{X}] = R E[\mathbf{b} | \mathbf{X}] - \mathbf{q} = R \mathbf{\beta} - \mathbf{q} = 0 \\
\text{Var}[\mathbf{m} | \mathbf{X}] = \text{Var}[\mathbf{Rb} - \mathbf{q} | \mathbf{X}] = R \text{Var}[\mathbf{b} | \mathbf{X}] R' = \sigma^2 R (X'X)^{-1} R'
\]

Wald Test:

\[
W = \mathbf{m}'[\text{Var}[\mathbf{m} | \mathbf{X}]]^{-1} \mathbf{m} = (\mathbf{Rb} - \mathbf{q})'\sigma^2 R (X'X)^{-1} R' (\mathbf{Rb} - \mathbf{q}) \sim \chi^2_J
\]

where \( J \) is the number of rows of \( \mathbf{R} \), i.e. the number of restrictions.
Testing Linear Restrictions, Examples

Unfortunately we do not have $\sigma^2$, so we use a test with $s^2$, by modifying $W$:

$$F = W \times \frac{1}{J} \frac{s^2}{\sigma^2} \left( \frac{n-k}{n-k} \right)$$

$$F = (Rb - q)'(\sigma^2 R (X'X)^{-1} R')(Rb - q) \times \frac{1}{J} \frac{s^2}{\sigma^2} \left( \frac{n-k}{n-k} \right)$$

$$= \frac{(Rb - q)'(\sigma^2 R (X'X)^{-1} R')^{-1}(Rb - q)/J}{[(n-k)\sigma^2/s^2]/(n-k)} = \frac{X_n/J}{X_d/(n-k)}$$

$$= \frac{1}{J}(Rb - q)'(s^2 R (X'X)^{-1} R')^{-1}(Rb - q)$$

Where:

$$X_{\text{numerator}} = \frac{X_n}{J} \sim \chi^2_J$$

$$X_{\text{denominator}} = \frac{X_d}{(n-k)} \sim \chi^2_{n-k}.$$
Testing Linear Restrictions, Examples

With only 1 linear restriction, this simplifies down to:

\[ H_0: r_1 \beta_1 + r_2 \beta_2 + \ldots + r_k \beta_k = r \beta = q \]

\[ F_{1,n-k} = \frac{\sum_j (r_j b_j - q)^2}{\sum_j \sum_k r_j r_k \text{Est.Cov.}[b_j, b_k]} \]

and suppose the restriction is on the \( \ell \)th coefficient:

\[ \beta_\ell = 0, \quad \text{so} \quad R = [0, 0, \ldots, 0, 1, 0, \ldots, 0, 0], \quad q = [0] \]

so that \( R(X'X)^{-1}R \) is just the \( \ell \) th diagonal value of \( (X'X)^{-1} \).

We use \( b \) to test for \( \beta \).

Giving:

\[ Rb - q = b_\ell - q, \quad F_{1,n-k} = \frac{(b_\ell - q)^2}{\text{Est.Var.}[b_\ell]^{-1}}. \]
Testing a Restriction with a Slightly Different Anaemia Model

```r
anaemia <- read.table("http://jgill.wustl.edu/data/anaemia.dat", 
                      header=TRUE,row.names=1)
a.lm2.out <- lm(Hb ~ Age + Menapause, data=anaemia)
summary(a.lm2.out)

Residuals:
     Min  1Q Median  3Q Max
-2.8375 -0.5839  0.1495  0.7821  2.0312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.68823    1.15466   7.525 8.32e-07
Age         0.11716    0.03588   3.265  0.00456
Menapause   0.60002    1.10070   0.545  0.59275

Residual standard error: 1.198 on 17 degrees of freedom
Multiple R-squared: 0.7776,    Adjusted R-squared: 0.7514
F-statistic: 29.71 on 2 and 17 DF,  p-value: 2.827e-06
```
Testing a Restriction with the Anaemia Data

Test the hypothesis that **Menopause** is equal to zero:

\[ \beta_2 = 0, \quad J = 1, \quad R = [0, 0, 1], \quad q = 0 \]

Pieces:

\[ (R\beta - q) = [0, 0, 1] \begin{bmatrix} 8.688 \\ 0.117 \\ 0.600 \end{bmatrix} - 0 = 0.6, \quad s^2 = 1.198, \quad X'X = \begin{bmatrix} 20 & 876 & 10 \\ 876 & 43074 & 572 \\ 10 & 572 & 10 \end{bmatrix} \]

Test Statistic:

\[
F[J, n - K] = \frac{1}{J} (Rb - q)' (s^2R(X'X)^{-1}R')^{-1} (Rb - q)
\]

\[
F[1, 20 - 3] = \frac{1}{1} (0.6)' \begin{bmatrix} 0.930 & -0.027 & 0.631 \\ -0.027 & 0.001 & -0.024 \\ 0.631 & -0.024 & 0.845 \end{bmatrix} [0, 0, 1]' \begin{bmatrix} 0.930 & -0.027 & 0.631 \\ -0.027 & 0.001 & -0.024 \\ 0.631 & -0.024 & 0.845 \end{bmatrix}^{-1} (0.6)
\]

\[ F[1, 17] = 0.356 \]

which is *not* in the tail, meaning that the restriction is *supported*. 
Testing a Restriction with the Anaemia Data

R <- c(0,0,1); q <- 0
b <- coef(a.lm2.out)
s.2 <- summary(a.lm2.out)$sigma
X <- cbind(rep(1,length=nrow(anaemia)),as.matrix(anaemia[,3:4]))

t(X) %*% X

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Menopause</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>876</td>
<td>10</td>
</tr>
<tr>
<td>Age</td>
<td>876</td>
<td>43074</td>
</tr>
<tr>
<td>Menopause</td>
<td>10</td>
<td>572</td>
</tr>
</tbody>
</table>

round(solve(t(X) %*% X),3)

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Menopause</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.930</td>
<td>-0.027</td>
</tr>
<tr>
<td>Age</td>
<td>-0.027</td>
<td>0.001</td>
</tr>
<tr>
<td>Menopause</td>
<td>0.631</td>
<td>-0.024</td>
</tr>
</tbody>
</table>

(F <- t(R%*%b-q) %*% solve(s.2*R %*% solve(t(X) %*% X) %*% R) %*% (R%*%b-q))
0.3558787

pf(F,1,17,lower.tail=FALSE)
0.5586653
Testing Nonlinear Restrictions

\( H_0: c(\beta) = q \) where \( c() \) is some nonlinear function.

Simple 1-restriction case:
\[
z = \frac{c(\hat{\beta})}{\text{est.} \, SE(c(\hat{\beta}))} \sim t_{n-k}
\]
(or equivalently \( z^2 \sim F_{1, n-k} \)).

But getting \( \text{est.} \, SE(c(\hat{\beta})) \) is hard, so start with a Taylor series expansion:
\[
f(b) = f(a) + f'(a) \frac{(b-a)}{1!} + f''(a) \frac{(b-a)^2}{2!} + f'''(a) \frac{(b-a)^3}{3!} + \ldots
\]
and then drop all but the first two terms to get an approximation:
\[
f(b) \approx f(a) + f'(a) \frac{(b-a)}{1!} = f(a) + f'(a)(b-a).
\]

Substituting \( \beta = a, \hat{\beta} = b \), and \( c() = f() \), gives:
\[
c(\hat{\beta}) \approx c(\beta) + \left( \frac{\partial c(\beta)}{\partial \beta} \right)^T (\hat{\beta} - \beta),
\]
and also
Testing Nonlinear Restrictions

Now we can calculate the needed variance term:

\[ \text{Var}(c(\hat{\beta})) = E\left[c(\hat{\beta})^2\right] - (E[c(\hat{\beta})])^2 \]

\[ = E\left[\left(c(\beta) + \left(\frac{\partial c(\beta)}{\partial \beta}\right)^T(\hat{\beta} - \beta)\right)^2 \right] - (E[c(\hat{\beta})])^2 \]

\[ \approx E\left[c(\beta)^2 - 2c(\beta)\left(\frac{\partial c(\beta)}{\partial \beta}\right)^T(\hat{\beta} - \beta) + \left(\frac{\partial c(\beta)}{\partial \beta}\right)^T(\hat{\beta} - \beta)^2 \left(\frac{\partial c(\beta)}{\partial \beta}\right)\right] - (E[c(\hat{\beta})])^2 \]

\[ = c(\beta)^2 - 2c(\beta)^2 \left(\frac{\partial c(\beta)}{\partial \beta}\right)^T(0) + E\left[\left(\frac{\partial c(\beta)}{\partial \beta}\right)^T(\hat{\beta} - \beta)^2 \left(\frac{\partial c(\beta)}{\partial \beta}\right)\right] - c(\beta)^2 \]

\[ = E\left[\left(\frac{\partial c(\beta)}{\partial \beta}\right)^T \text{Var}(\hat{\beta}) \left(\frac{\partial c(\beta)}{\partial \beta}\right)\right] \]

since \( E[(\hat{\beta} - \beta)^2] = E[\hat{\beta}^2 - 2\hat{\beta}\beta + \beta^2] = E\hat{\beta}^2 - \beta^2 = \left[\text{Var}[\hat{\beta}] - (E[\hat{\beta}])^2\right] - \beta^2 = \text{Var}(\hat{\beta}). \)

This means that we can use sample estimates for \(\frac{\partial c(\beta)}{\partial \beta}\) and plug in \(s^2(X'X)^{-1}\) for \(\text{Var}(\hat{\beta})\) and then test with a normal distribution, provided reasonable sample size.
Testing a Nonlinear Restriction with the Anaemia Data

Test: \( H_0: c(\beta) = q \) where \( c(\beta) = \beta^{1/2} \) and \( q = 0 \), versus \( H_1: c(\beta) \neq q \).

Calculate: \( \frac{\partial c(\beta)}{\partial \beta} = \frac{1}{2}(\beta)^{-\frac{1}{2}} \).

From the model fit we have:

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| Age      | 0.11716  | 0.03588    | 3.265   | 0.00456  |

Since this is scalar test, use (with \( c(\beta) = 0 \)):

\[
z = \frac{c(\hat{\beta}) - 0}{\text{est.}SE(c(\beta))} \approx \frac{c(\beta) + \left( \frac{\partial c(\beta)}{\partial \beta} \right)^T (\hat{\beta} - \beta)}{\sqrt{\left( \frac{\partial c(\beta)}{\partial \beta} \right)' \text{Var}(\beta) \left( \frac{\partial c(\beta)}{\partial \beta} \right)}}
\]

\[
= \frac{\frac{1}{2}(0.11716)^{-\frac{1}{2}}(0.11716)}{\sqrt{\frac{1}{2}(0.11716)^{-\frac{1}{2}}(0.03588)^2 + \frac{1}{2}(0.11716)^{-\frac{1}{2}}}} = 0.7651 \sim t_{n-k}
\]

We fail to reject the restriction since:

\[
\text{pt}(0.7651, 17, \text{lower.tail}=\text{FALSE})
\]

[1] 0.2273
Dealing with Heteroscedasticity: The *TWEED* Dataset

- **TWEED** = Terrorism in Western Europe: Events Data (Jan Oskar Engene)

- Contains information on events related to internal (domestic) terrorism in 18 countries in Western Europe.

- The time period covered is 1950 to 2004.

- By focusing on internal terrorism, the TWEED data set only includes events initiated by agents originating in the West European countries.

- Terrorism data is characterized by observable and latent groupings/clusters.

- So there is likely to be extra heteroscedasticity in the linear model from the presence of these groups.
Dealing With Heteroscedasticity From Group Effects

What if there is heterogeneity in the standard errors from a group definition.

This does not bias the coefficient estimates but will affect the estimated standard errors.

Suppose there are $M$ groups, with modified degrees of freedom for the model now equal to

$$df_{robust} = \frac{M}{(M - 1)(N - K)}$$

Cluster-Robust standard errors (White, Huber, etc.) adjust the variance-covariance matrix with a “sandwich estimation” approach:

$$VC^* = f_{robust} \left( X'X \right)^{-1} \underbrace{ (U'U) }_{bread} \underbrace{ (X'X)^{-1} }_{meat} \underbrace{ (X'X)^{-1} }_{bread}$$

where:

- $U$ is an $M \times k$ matrix,
- such that each row is produced by $X_m \ast e_m$ for group/cluster $m$, the element-wise product of the $N_m \times k$ matrix of observations in group $m$,
- and the $N_m$-length $e_m$ corresponding residuals vector.
Dealing With Heteroscedascity From Group Effects

▶ A non-clustered, non-robusted model:

```r
tweed <- read.table("http://jgill.wustl.edu/data/tweed2.dat",header=TRUE)
tweed.lm <- lm(I(killed+injured)~year+arrests+factor(attitude), data=tweed)
summary(tweed.lm)
```

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 252.070    | 213.799 | 1.18     | 0.239    |
| year      | -0.123     | 0.108   | -1.15    | 0.253    |
| arrests   | 5.359      | 2.188   | 2.45     | 0.015    |
| factor(attitude)Ethnic/regionalist Separatist | -6.373   | 7.137  | -0.89    | 0.373    |
| factor(attitude)Left wing extremist Other | -6.788    | 3.631   | -1.87    | 0.063    |
| factor(attitude)Right wing extremist Other | -4.710    | 3.632   | -1.30    | 0.196    |

Residual standard error: 12.4 on 283 degrees of freedom
Multiple R-squared: 0.0374, Adjusted R-squared: 0.0204
F-statistic: 2.2 on 5 and 283 DF, p-value: 0.0546

▶ The reference group for the factor is Ethnic/regionalist Irredentist.
Dealing With Heteroscedasticity From Group Effects

A model with robust standard errors:

```r
lapply(c("sandwich","lmtest","plm"),library, character.only=TRUE)
M <- length(table(tweed$attitude))
new.df <- tweed.lm$df / (tweed.lm$df - (M -1))
clx(tweed.lm, new.df, tweed$attitude)
```

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 252.0696 | 112.4921 | 2.24 | 0.026 |
| year | -0.1233 | 0.0566 | -2.18 | 0.030 |
| arrests | 5.3585 | 6.6622 | 0.80 | 0.422 |
| factor(attitude)Ethnic/regionalist Separatist | -6.3725 | 0.3234 | -19.70 | < 2e-16 |
| factor(attitude)Left wing extremist Other | -6.7882 | 0.7866 | -8.63 | 4.5e-16 |
| factor(attitude)Right wing extremist Other | -4.7096 | 0.0379 | -124.22 | < 2e-16 |
Dealing With Heteroscedasticity From Group Effects

Stata calculates Huber-White standard errors differently by using the same coefficient estimates as the regular linear model results, but scaling the variance covariance matrix by the degrees of freedom:

\[ V_{\text{Stata}} = \frac{M}{(M - 1)(N - K)} \times f_{\text{robust}}(X'X)^{-1}(U'U)(X'X)^{-1} \]

This is easily obtained in R with the following:

```r
library(sandwich)
hw.se <- sqrt(diag(vcovHC(tweed.lm,type="HC1")))
cbind(tweed.lm$coef,hw.se)
```

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>252.06960</td>
<td>219.06530</td>
</tr>
<tr>
<td>year</td>
<td>-0.12333</td>
<td>0.11027</td>
</tr>
<tr>
<td>arrests</td>
<td>5.35855</td>
<td>6.04840</td>
</tr>
<tr>
<td>factor(attitude)Ethnic/regionalist Separatist</td>
<td>-6.37254</td>
<td>3.05351</td>
</tr>
<tr>
<td>factor(attitude)Left wing extremist Other</td>
<td>-6.78818</td>
<td>3.14835</td>
</tr>
<tr>
<td>factor(attitude)Right wing extremist Other</td>
<td>-4.70957</td>
<td>3.41273</td>
</tr>
</tbody>
</table>