Everything You Need To Know About the Linear Model

JEFF GILL
Distinguished Professor
Departments of Government and Mathematics & Statistics
American University

Why We Need to Start with the Linear Model For This Course

- We are interested in Bayesian Hierarchical Models.
- 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 $Z_{j\cdot}$ 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, β_{j0}, and β_{j1}.
- Thus it is important to review the linear regression model in some detail.
- Even though this lecture proves or derives every quantity of interest, it is assumed that you've seen linear regression models before.

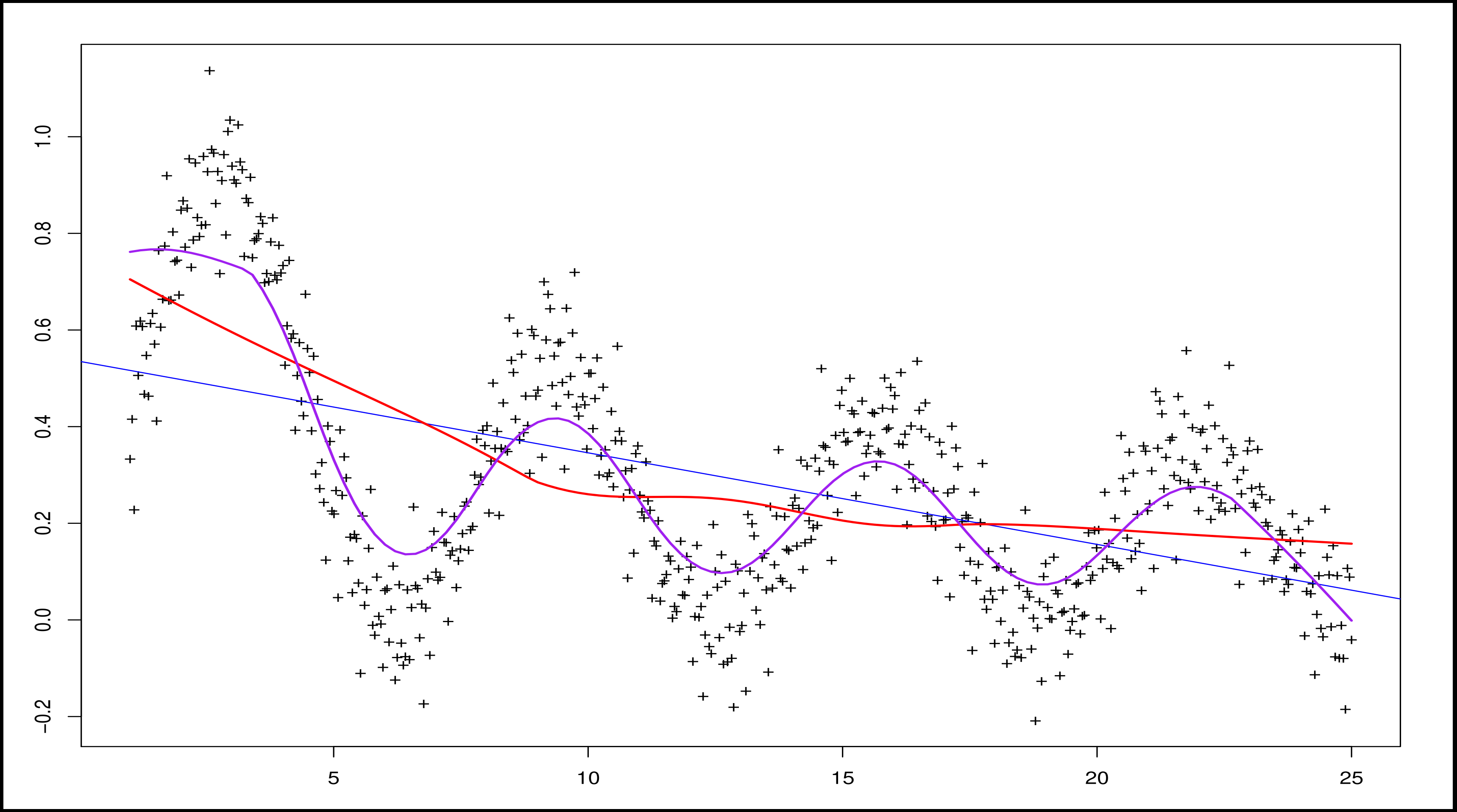
Precursor To Linear Models: Following Trends

- 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

```
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

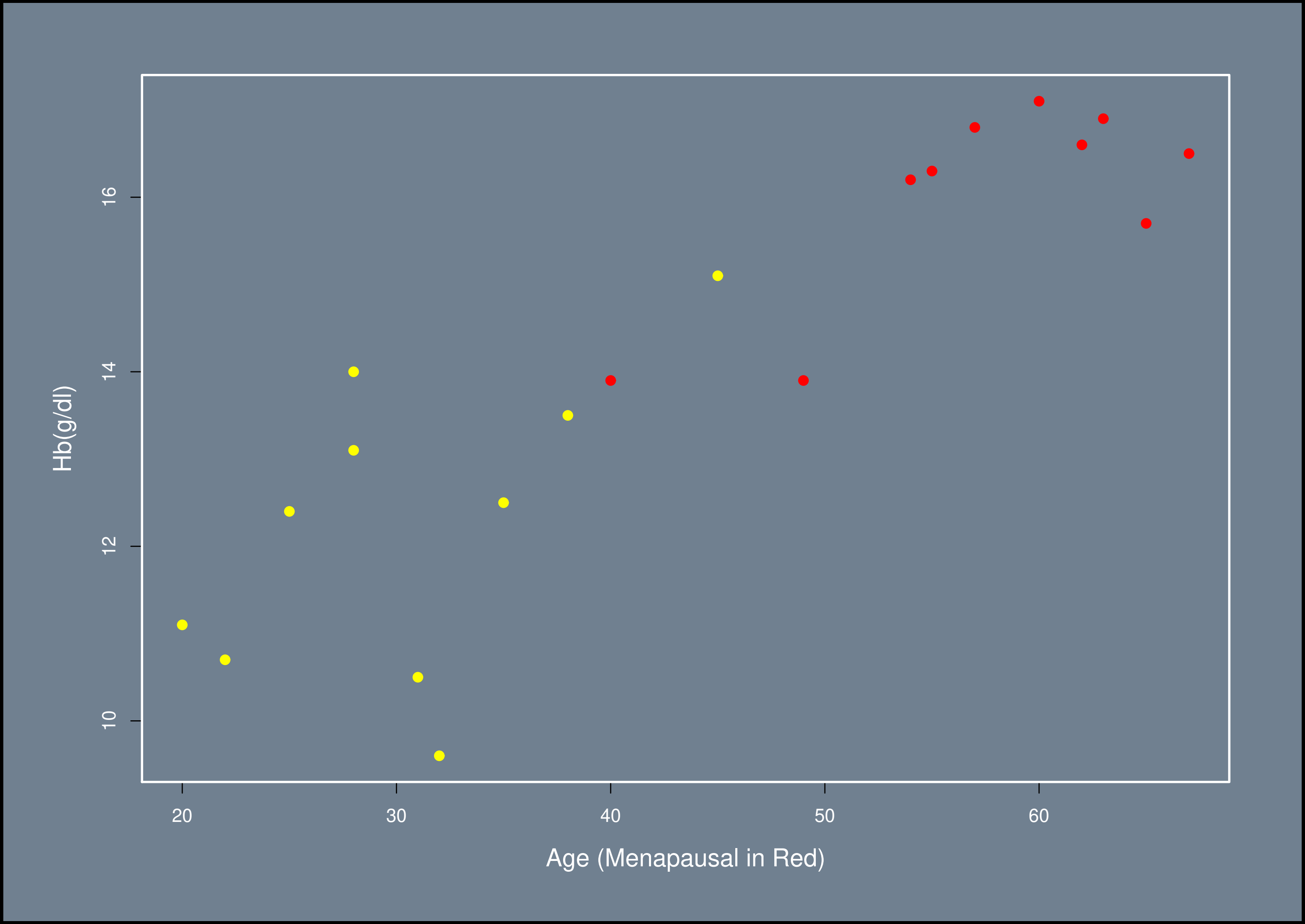
  - menopausal status ($0 = no, 1 = yes$)

- There is an obvious endogeneity problem in modeling Hb(g/dl) versus PCV(%).

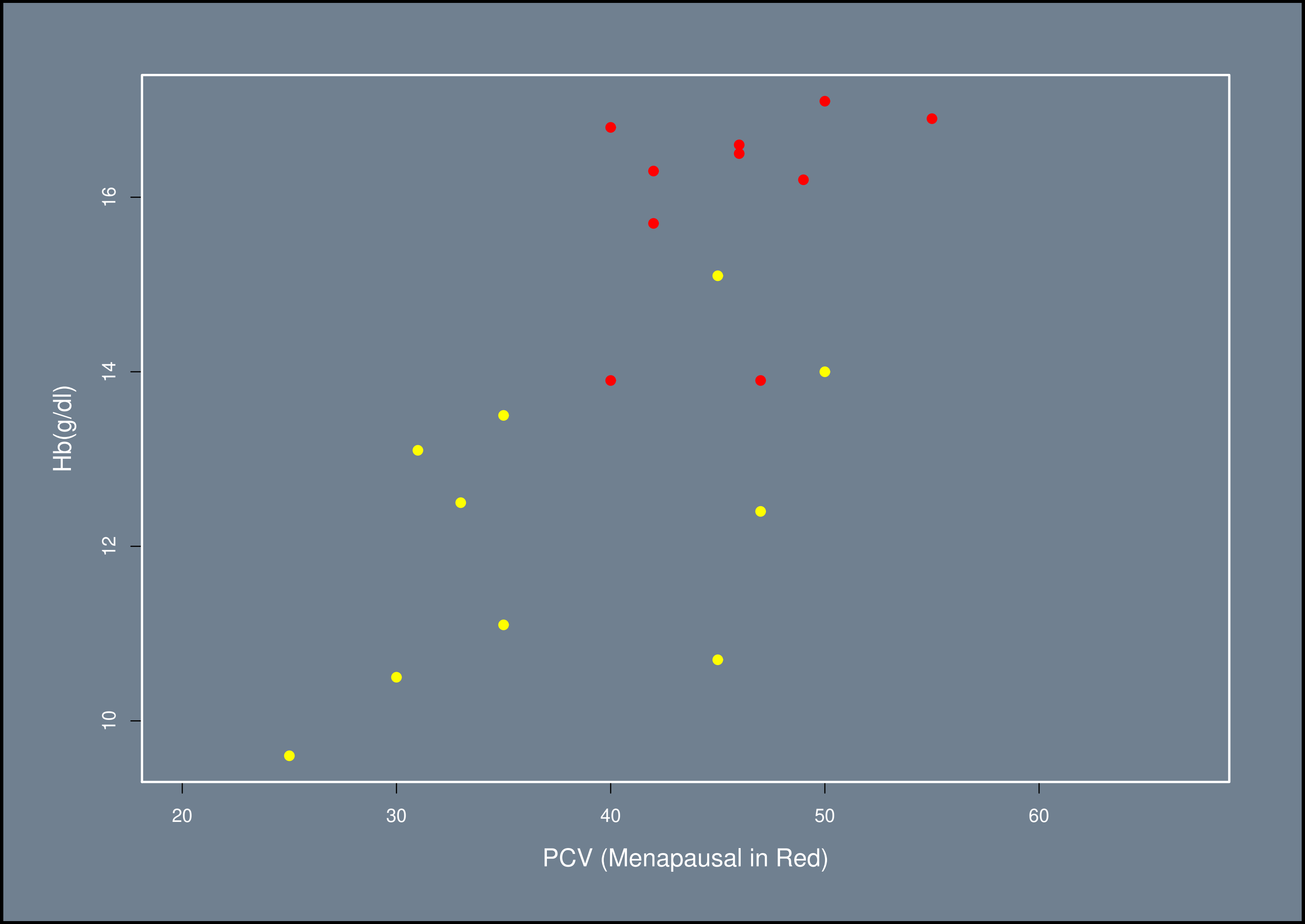
What Does the Linear Model Get You? An Example

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

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("https://jeffgill.org/files/jeffgill/files/anaemia.txt",
    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?

- The New York Times Magazine, August 7, 2011, page 13.
- 24 countries: average survey review of restaurant service quality and a tipping index from three travel etiquette web sites.
- The data

Country	Quality	Tip	Country	Quality	Tip
Japan	4.4	0.00	Thailand	3.9	0.03
Canada	3.7	0.16	New_Zealand	3.7	0.07
UAE	3.6	0.10	Germany	3.6	0.08
USA	3.6	0.18	South_Africa	3.5	0.11
Australia	3.4	0.08	Argentina	3.4	0.10
Morocco	3.4	0.07	Turkey	3.4	0.08
India	3.3	0.10	Brazil	3.3	0.07
Vietnam	3.2	0.05	England	3.2	0.10
Greece	3.2	0.08	Spain	3.1	0.08
France	3.1	0.08	Italy	3.0	0.07
Egypt	3.0	0.08	Mexico	3.0	0.13
China	2.9	0.03	Russia	1.7	0.10

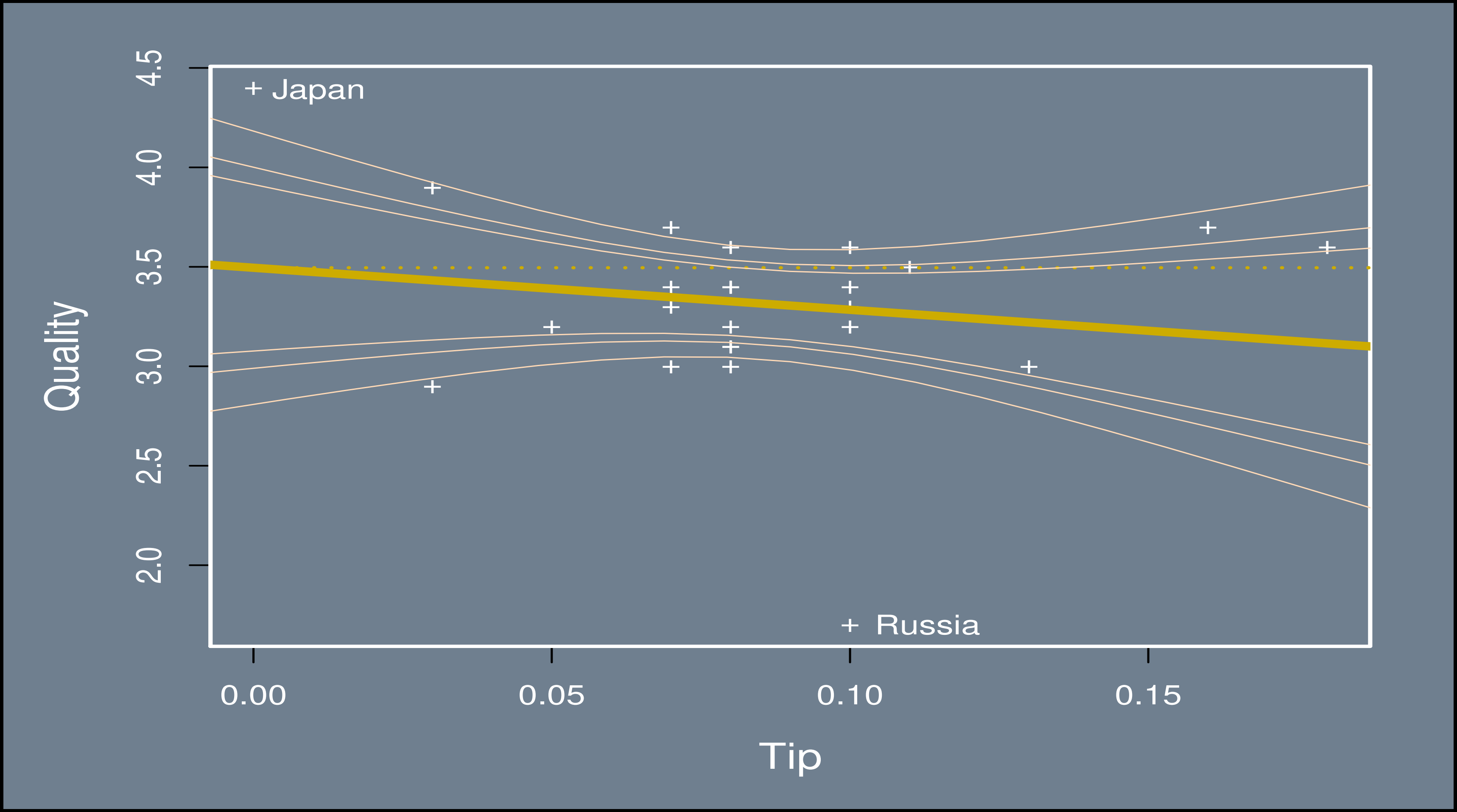
What Happens When it Doesn’t Work?

```
service <-
    read.table("https://jeffgill.org/files/jeffgill/files/service.dat_.txt",
    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

              Coef Std.Err. 0.95 Lower 0.95 Upper   CIs:  ZE+RO
(Intercept)  3.495    0.244      3.018      3.973             |o|
Tip         -2.113    2.632     -7.272      3.046  |-----o-----|

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?



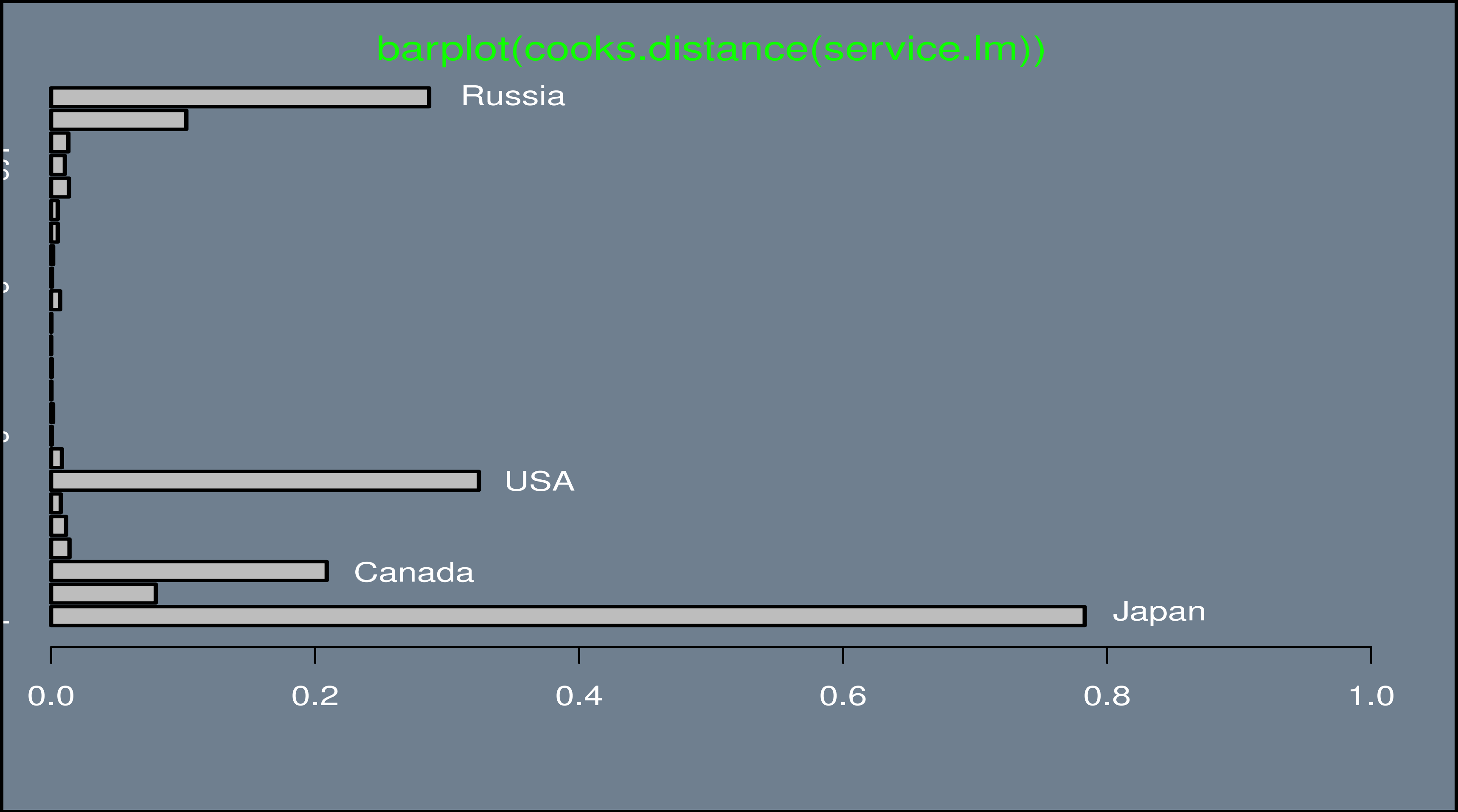
Are Some Cases More Influential in LM Results

- Leverage is defined by the ability to be influential and is related to the distance that a point is from $\bar{X}$ on the x-axis.
  - Define the Hat Matrix as $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (more on this later), which is square and symmetric, and $\mathbf{X}$ is the matrix containing the explanatory variables.
  - The the diagonal value h_{ii} is the potential of the ith point to to influential, which is its leverage.
- Influence is defined as realized leverage, meaning how much does a case with large leverage exert influence over the estimated coefficients relative to other cases.
  - The most common way to measure influence is *Cook's Distance*:

    $$D_i = \frac{e_i^2}{ks^2}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right)$$

    where: e_i is the ith residual, k is the number of columns in the $\mathbf{X}$ matrix, and $s^2 = \mathbf{e}'\mathbf{e}/(n-k)$ is the MSE of the regression.
  - The function in `R` is `cooks.distance()`.
- This is getting just a little ahead of ourselves here, so consider it a preview.

How Influential is Japan?



```
barplot(cooks.distance(service.lm))
```

Bi-Directional Symmetric Bivariate Relationships

- It is useful to start with bivariate regression where the quantities calculated are in scalar rather than matrix notation.

- Suppose the vectors **X** and **Y** have equal standard deviation, denoted s (obviously a special case).

- Then the regression of **Y** on **X** (`y ~ x`) has the same slope as the regression of **X** on **Y** (`x ~ y`) since:
$$\hat{\beta}_{y|x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Cov}(X,Y)}{s^2} = \hat{\beta}_{x|y}$$
which is interesting in its own right.

- But now recall that the correlation coefficient is:
$$cor(X,Y) = r_{x,y} = \frac{\text{Cov}(X,Y)}{s_X s_Y}$$
which is the equation above rescaled for different standard deviations and letting $(n-1)$ in the numerator and denominator cancel each other since:
$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Correlation *Is* Regression

- So the above point means that regression *is* correlation, since:

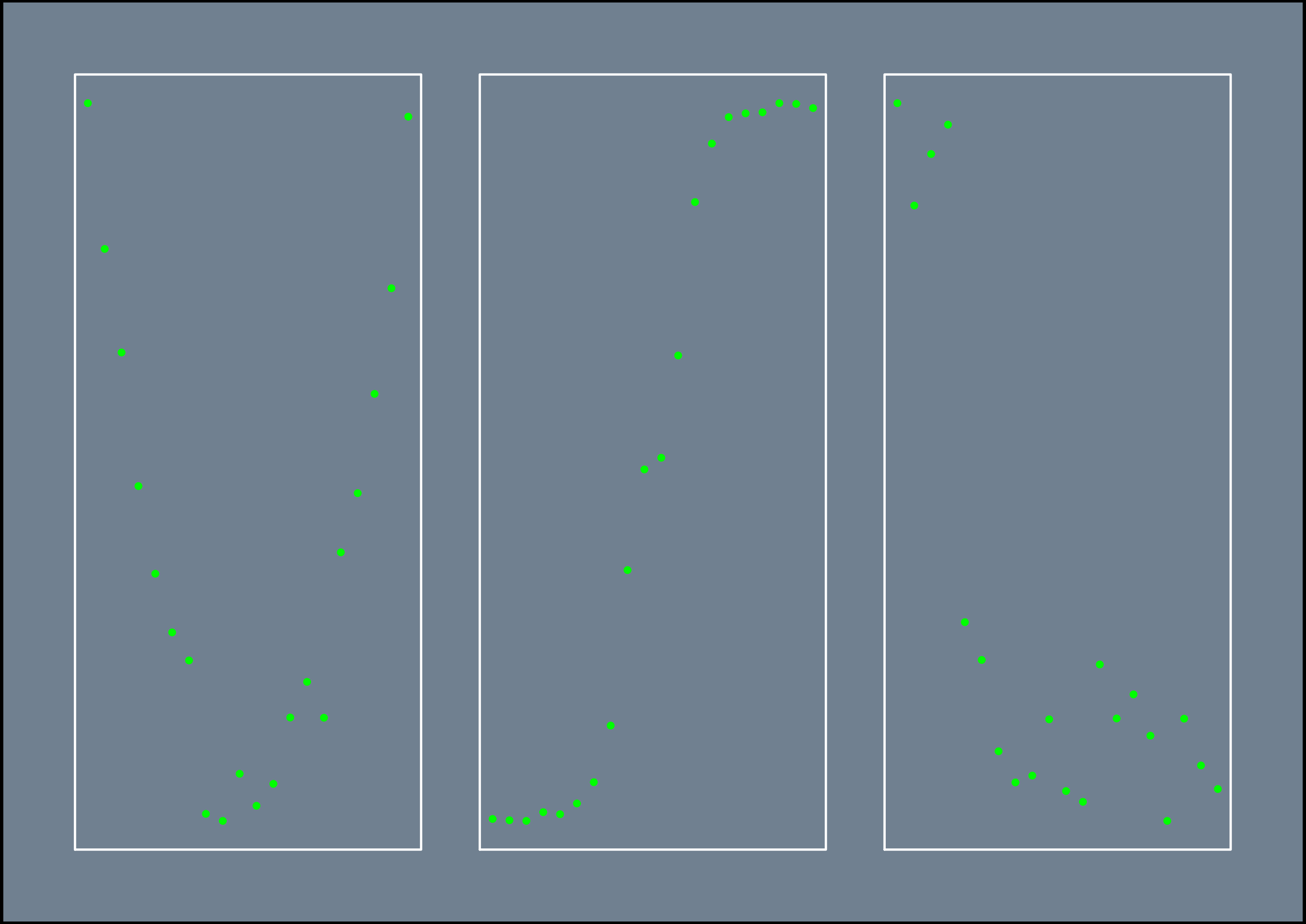
  $$cor(X,Y) = \frac{s_X}{s_Y}\hat{\beta}_{y|x}$$

  in the more general case where $s_X \neq s_Y$(and they are identical in the special case).

- From our anaemia example picking `Age`:

```
coef(a.lm.out)[2]
0.1342515
apply(anaemia,2,sd)
        Hb        PCV        Age  Menapause
 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:

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

- HB and PCV from the anaemia data:

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

- HB and Age from the anaemia data:

$$r = 0.8795875 \qquad SE(r) = \sqrt{\frac{1 - 0.6733745^2}{20 - 2}} = 0.1121323 \qquad t = 7.844191 \qquad p \approx 0.0001.$$

Data Structures in Matrix/Vector Form

- The vector $\mathbf{Y}$ contains values of the outcome variable in a column vector:

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

- The matrix $\mathbf{X}$ contains the explanatory variables down the columns:

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

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

- The vector $\mathbf{e}$ contains values of the observed residuals (disturbances, errors) in a column vector:

$$\mathbf{e} = [e_1, e_2, \ldots, e_n]'$$

Gauss-Markov Assumptions for Classical Linear Regression

- Functional Form: $\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times k)(k\times 1)}{\mathbf{X}\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$ (recall $\mathbf{X}$ has a leading column of 1's)

- Mean Zero Errors: $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$

- Homoscedasticity: $\text{Var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$

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

- Exogeneity of Explanatory Variables: $\text{Cov}[\epsilon_i, \mathbf{X}] = 0, \qquad \forall i$

▷ Note that every one of these lines has $\boldsymbol{\epsilon}$ 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. . . $\boldsymbol{\epsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.
- 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:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \underset{(1\times n)(n\times 1)}{\mathbf{Y}'\mathbf{Y}} - \underset{(1\times n)(n\times k)(k\times 1)}{2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}} + \underset{(1\times k)(k\times n)(n\times k)(k\times 1)}{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}} \end{aligned}$$

  which is quadratic in $\boldsymbol{\beta}$ space.

- Take the derivative of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = 0 - 2 \underset{(k\times n)(n\times 1)}{\mathbf{X}'\mathbf{Y}} + \underset{(k\times n)(n\times k)(k\times 1)}{2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}} \equiv 0,$$

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

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

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

OLS Estimator Notes

- Another way to express the OLS estimator is to say that we want the $\boldsymbol{\beta}$ that minimizes the *squared prediction error*:
$$S(\boldsymbol{\beta}) = \mathbb{E}[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^2],$$
which is a restatement of $S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}$.

- Sometimes the solution is called the *linear projection coefficient*:
$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\Re}{\text{argmin}} S(\mathbf{b}).$$

- And yet another expression for this quantity is:
$$\hat{\boldsymbol{\beta}} = \mathbb{E}(\mathbf{X}\mathbf{Y})\left(\mathbb{E}(\mathbf{X}\mathbf{X}')\right)^{-1}$$
meaning that:
$$\hat{\mathbf{Y}} = \mathbf{X}\mathbb{E}(\mathbf{X}\mathbf{Y})\left(\mathbb{E}(\mathbf{X}\mathbf{X}')\right)^{-1}.$$

Estimation With MLE:

- Assume: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \sigma^2 I$
(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: $\epsilon_i = y_i - \mathbf{X}_i\boldsymbol{\beta}$ (where $\boldsymbol{\beta}$ to be estimated):

$$L(\boldsymbol{\beta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right].$$

- Which in log-likelihood form is:

$$\ell(\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Estimation With MLE:

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

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}(-\mathbf{X})'(2)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \equiv 0 \\
&\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0 \\
&\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0 \\
&\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{Y} \qquad\qquad \text{the ``normal'' equation} \\
&\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\partial}{\partial \sigma^2} \ell(\sigma^2) &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \equiv 0 \\
0 &= -n\sigma^2 + \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\
\hat{\sigma}^2 &= \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{n}
\end{aligned}
$$

which is slightly biased in finite samples for $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/(n-k)$, more on this later.

Implications

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

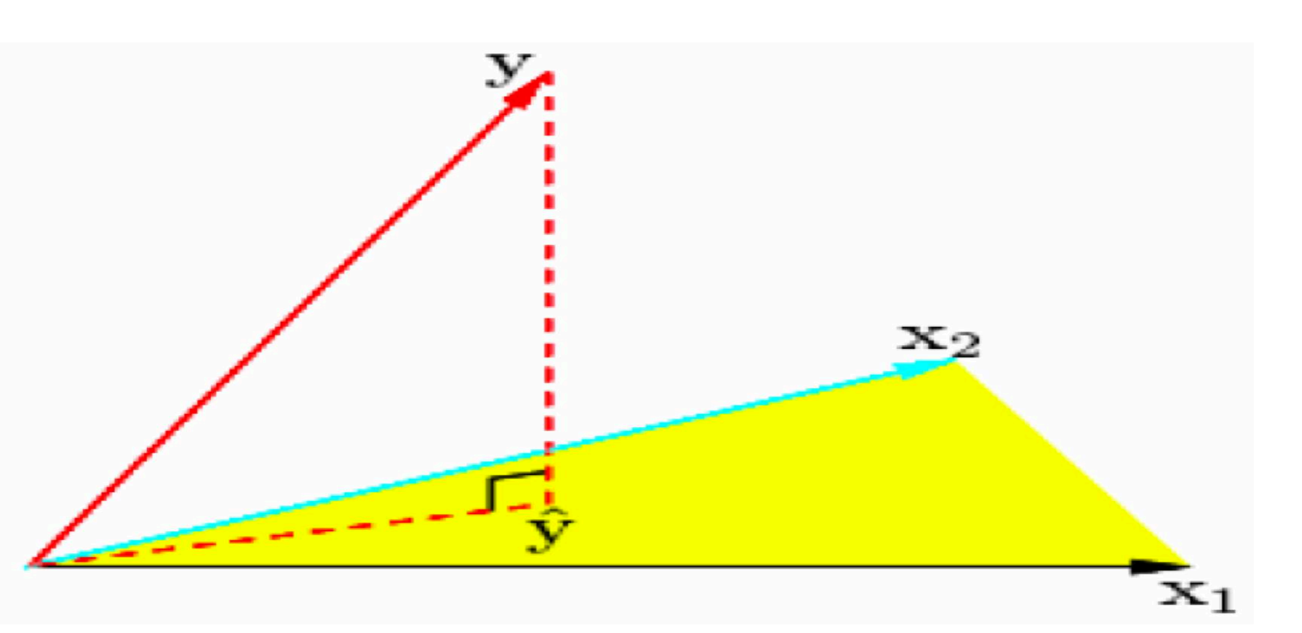
$$\begin{aligned}
\mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{Y} - \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= \mathbf{Y} - (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\
&= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\
&= \mathbf{M}\mathbf{Y}
\end{aligned}$$

  where $\mathbf{M}$ and $\mathbf{H}$ are symmetric and idempotent.

- For example: $\mathbf{H}{\cdot}\mathbf{H} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

The HAT Matrix

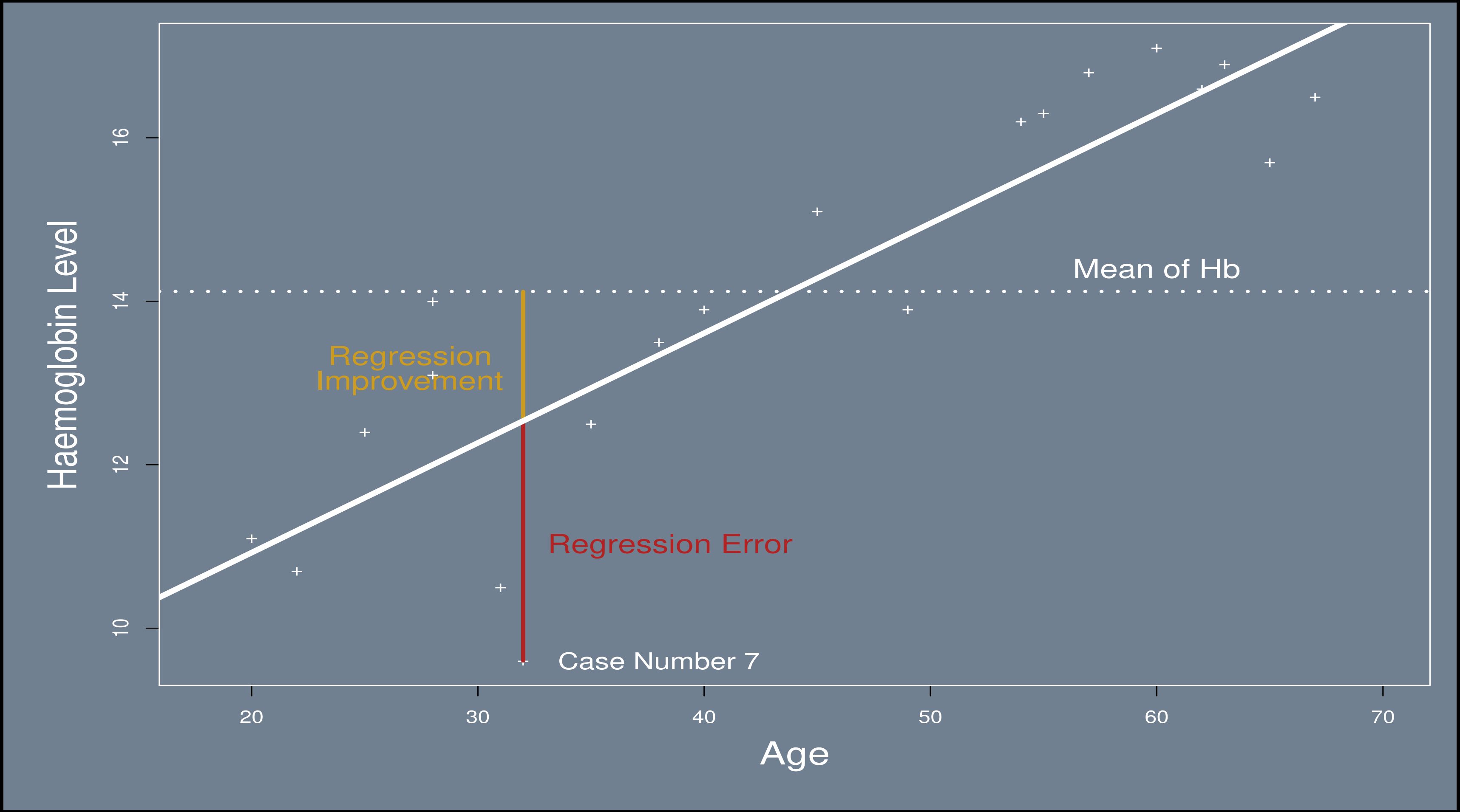
- The name is because $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} = \mathbf{HY}$, but "projection matrix" ($\mathbf{P}$) is better for geometric reasons.



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 $\mathbf{P}(\mathbf{I} - \mathbf{P}) = 0$ (orthogonality)
  - $\mathbf{e}'\mathbf{e} = \mathbf{Y}'\mathbf{M}'\mathbf{M}\mathbf{Y} = \mathbf{Y}'\mathbf{M}\mathbf{Y} = \mathbf{Y}\mathbf{e}$ (sum of squares)
  - $\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{Y} + \boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ (the Pythagorean Theorem!)
- Interestingly, the hat matrix also comes up in some nonlinear model theory.

Fit & Decomposition, Illustration



Fit & Decomposition, Variability Definitions

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

$$\text{SST} = \sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})^2$$

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

$$\text{SSR} = \sum_{i=1}^{n} (\hat{\mathbf{Y}}_i - \bar{\mathbf{Y}})^2$$

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

$$\text{SSE} = \sum_{i=1}^{n} (\hat{\mathbf{Y}}_i - \mathbf{Y}_i)^2$$

Fit & Decomposition, Total Variability

- Interesting manipulations of the sum of squares total:

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{n}(\mathbf{Y}_i^2 - 2\mathbf{Y}_i\bar{\mathbf{Y}} + \bar{\mathbf{Y}}^2) \\
&= \sum_{i=1}^{n}\mathbf{Y}_i^2 - 2\sum_{i=1}^{n}\mathbf{Y}_i\bar{\mathbf{Y}} + n\bar{\mathbf{Y}}^2 \\
&= \sum_{i=1}^{n}\mathbf{Y}_i^2 - 2n\bar{\mathbf{Y}}^2 + n\bar{\mathbf{Y}}^2 \\
&= \sum_{i=1}^{n}\mathbf{Y}_i^2 - n\bar{\mathbf{Y}}^2 \qquad \text{(scalar description)} \\
&= \boxed{\mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}} \qquad \text{(matrix algebra description)}
\end{aligned}$$

  where $\mathbf{J}$ is a $n \times n$ matrix of all 1's.

- Note that pre-multiplying by $\mathbf{J}$ produces a same-sized matrix where the values are the sum by column, and post-multiplying by $\mathbf{J}$ 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

$$\sum_{i=1}^{n} \mathbf{Y}_i^2 - n\bar{\mathbf{Y}}^2 = \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}.$$

Fit & Decomposition, Regression Variability

- Sum of Squares Regression:

$$\begin{aligned}
\text{SSR} &= \sum_{i=1}^{n}(\hat{\mathbf{Y}}_i^2 - 2\hat{\mathbf{Y}}_i\bar{\mathbf{Y}} + \bar{\mathbf{Y}}^2) \\
&= \hat{\mathbf{Y}}'\hat{Y} - 2\bar{\mathbf{Y}}\sum_{i=1}^{n}\hat{Y}_i + n\bar{\mathbf{Y}}^2 \\
&= (\hat{\boldsymbol{\beta}}'\mathbf{X}')(\hat{Y}) - 2n\bar{\mathbf{Y}}^2 + n\bar{\mathbf{Y}}^2 \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{Y} - n\bar{\mathbf{Y}}^2 \\
&= \boxed{\hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{Y}} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}}
\end{aligned}$$

Fit & Decomposition, Remaining Variability

- Sum of Squares Error (using the Normal Equation, $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$):

$$\text{SSE} = \sum_{i=1}^{n}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 = \mathbf{e}'\mathbf{e}$$

$$= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

now do the Normal Equation substitution...

$$= \mathbf{Y}'\mathbf{Y} - (\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

$$= \mathbf{Y}'\mathbf{Y} - (\hat{\boldsymbol{\beta}}'\mathbf{X}')(\mathbf{X}\hat{\boldsymbol{\beta}})$$

$$= \boxed{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{Y}}}$$

Total %*&^#%ing Magic!

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

$$\begin{aligned} SSR + SSE &= (\hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{Y}} - n\mathbf{Y}'\mathbf{J}\mathbf{Y}) + (\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{Y}}) \\ &= \mathbf{Y}'\mathbf{Y} - n\mathbf{Y}'\mathbf{J}\mathbf{Y} \\ &= \mathrm{SST} \end{aligned}$$

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

```
7^2 + 3^2
[1] 58
(7+3)^2
[1] 100
```

(except in unusual or pathological circumstances).

The F-Statistic

- The F-statistic is an *omnibus* test statistic where tail values indicate that at least one of the regression coefficients is statistically signficant.

- The null hypothesis is that none of them are (implying a null model with no explanatory variables just an intercept, $\beta_0 \neq 0$, so rejecting the null only means that one or more are implied to be statistically reliable in the given model.

- Meaning that:

  $$H_0{:}\beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0,$$

  for an explantory variable matrix of dimension $(n \times k)$.

- The equation against the null model is given by:

  $$F = \frac{(SSR - SSE)/df_{diff}}{SSE/df_{SSE}} = \frac{(SSR - SSE)/(k-1)}{SSE/(n-k-1)} = \frac{MSR}{MSE}.$$

  (although submodels could be substituted for the null model).

- For the Menapause linear model:

```
F-statistic: 48.6 on 2 and 17 DF,  p-value: 9.26e-08
```

A Measure of Fit

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

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^o\mathbf{Y}} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^o\mathbf{X}\hat{\boldsymbol{\beta}}}{\mathbf{Y}'\mathbf{M}^o\mathbf{Y}}$$

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

- Note: $\mathbf{M}^o$ is idempotent and transforms means to deviances for the explanatory variables:

```
M.0 <- diag(3) - (1/3)*c(1,1,1)%*%t(c(1,1,1))
M.0
         [,1]     [,2]     [,3]
[1,]  0.66667 -0.33333 -0.33333
[2,] -0.33333  0.66667 -0.33333
[3,] -0.33333 -0.33333  0.66667
```

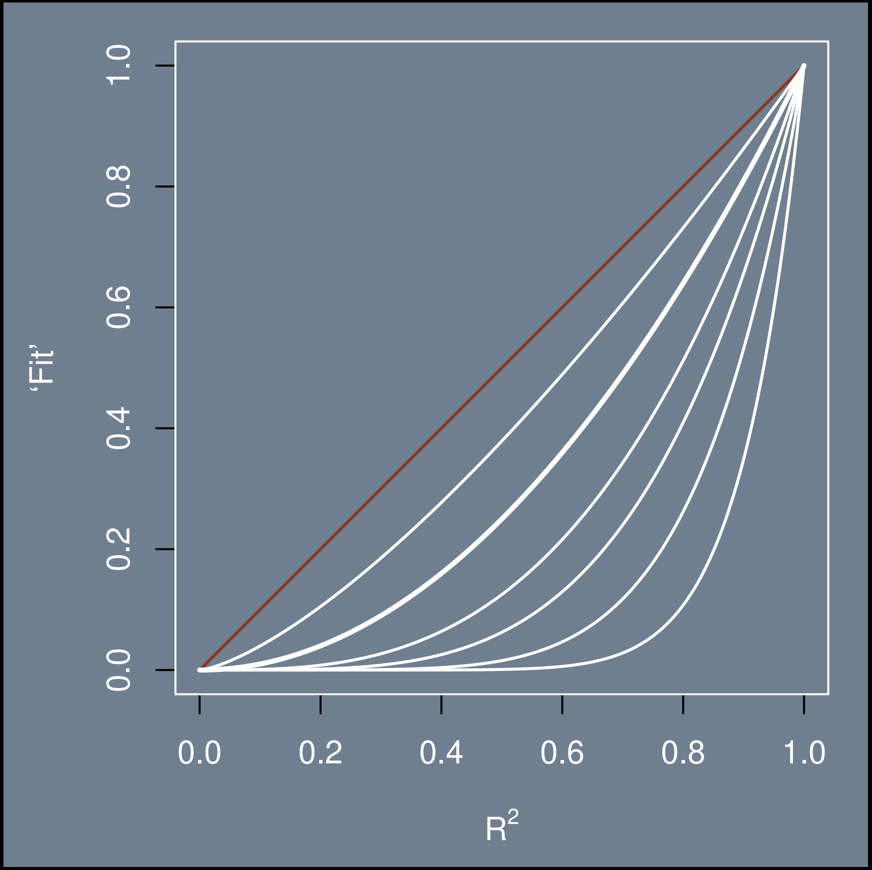
A Measure of Fit

- Also, there is another version that accounts for sample size and the number of explanatory variables (k):
$$R^2_{adj} = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-k)}{\mathbf{Y}'M^o\mathbf{Y}/(n-1)}$$
which is useful with small datasets.

- Bivariate relationships for R^2 from `lm(Y ~ X)`:
$$\frac{s_Y}{s_X}\text{cor}X,Y = \beta$$
$$R^2 = \text{cor}(X,Y)^2$$
$$\frac{s_Y}{s_X}\sqrt{R^2} = \beta$$
$$\sqrt{R^2} = \frac{s_X}{s_Y}\beta$$
$$R^2 = \left(\frac{s_X}{s_Y}\beta\right)^2$$

Warnings about R^2

- There is not a *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 due to the nonlinearity.



Does R^2 Have a Distribution?

- Surprisingly, yes.
- Note that the F-statistic can be expressed as:

$$F = \frac{(n-k)}{(k-1)} \times \frac{R^2}{1-R^2}.$$

- Algebraically rearranging:

$$R^2 = \frac{(k-1)F}{(n-k)+(k-1)F}.$$

- This means that:

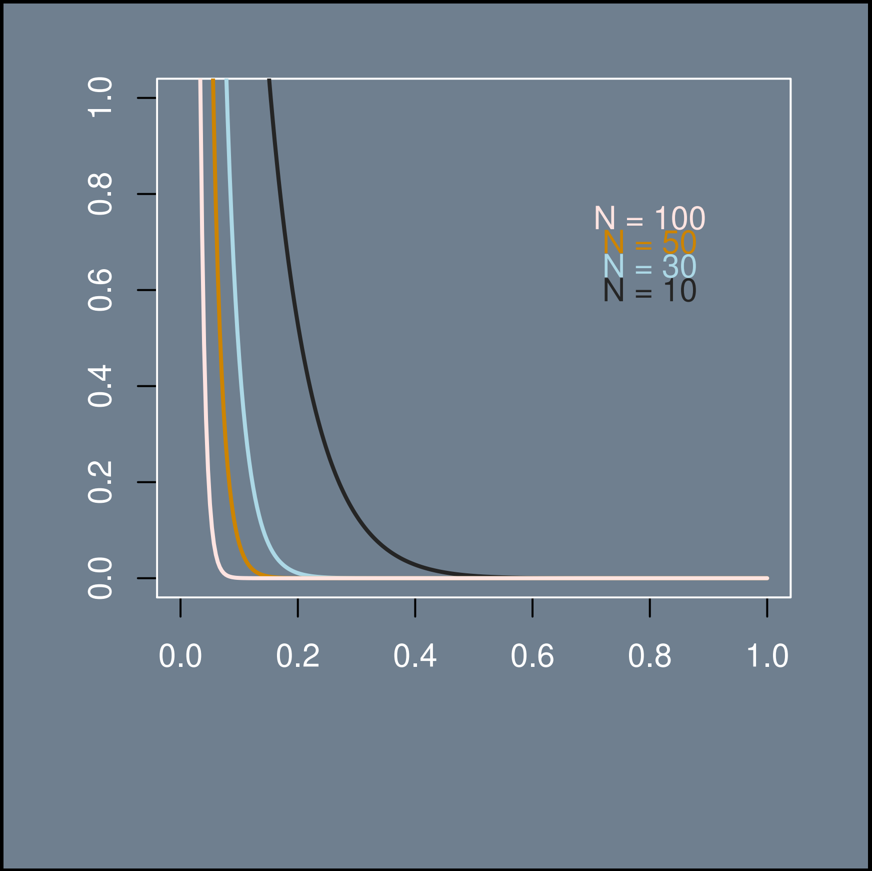
$$R^2 \sim \text{Beta}\left(\frac{k-1}{2}, \frac{n-k}{2}\right)$$

under the assumption that the effect of all the regressors is zero ($\boldsymbol{\beta} = 0$), inherited from the condition of the F-test.

Does R^2 Have a Distribution?

- For simplicity restrict ourselves to $k = 1$, and again H_0 : all $\boldsymbol{\beta} = 0$.

- What does R^2 look like for different sample sizes?



Properties of the Estimator, Unbiasedness

- $\hat{\boldsymbol{\beta}}$ is an estimator for $\boldsymbol{\beta}$, which can be rewritten according to:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\end{aligned}$$

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

- Taking expectations:

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}]\\ &= \mathbb{E}[\boldsymbol{\beta}]+\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}]\\ &= \boldsymbol{\beta}+\mathbb{E}[K\boldsymbol{\epsilon}]\\ &= \boldsymbol{\beta}+K\mathbb{E}[\boldsymbol{\epsilon}]=\boldsymbol{\beta}\end{aligned}$$

  shows that it is unbiased.

Properties of the Estimator, Variance

- By definition (using an outer product):

$$\begin{aligned}\text{Var}[\hat{\boldsymbol\beta}|\mathbf{X}] &= \mathbb{E}[(\hat{\boldsymbol\beta}-\boldsymbol\beta)(\hat{\boldsymbol\beta}-\boldsymbol\beta)'|\mathbf{X}] - \mathbb{E}[(\hat{\boldsymbol\beta}-\boldsymbol\beta|\mathbf{X})]^2 \\ &= \mathbb{E}[(\hat{\boldsymbol\beta}-\boldsymbol\beta)(\hat{\boldsymbol\beta}-\boldsymbol\beta)'|\mathbf{X}] - \mathbb{E}[0]^2.\end{aligned}$$

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

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

$$\begin{aligned}\text{Var}[\hat{\boldsymbol\beta}|\mathbf{X}] &= \mathbb{E}\left[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol\epsilon)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol\epsilon)'|\mathbf{X}\right] \\ &= \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol\epsilon\boldsymbol\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol\epsilon\boldsymbol\epsilon'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\text{Var}[\boldsymbol\epsilon|\mathbf{X}] + \mathbb{E}[\boldsymbol\epsilon|\mathbf{X}]^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Properties of the Estimator, General

- The OLS estimator is consistent (converges in probability):

$$\operatorname*{plim}_{n\to\infty}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

with Gauss-Markov assumption #5: $\text{Cov}[\epsilon_i, \mathbf{X}] = 0, \forall i$, and there is no perfect multicollinearity.

- The OLS estimate is optimal:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) \leq \text{Var}(\hat{\boldsymbol{\beta}}_{\text{All Other}})$$

with Gauss-Markov assumptions #3 $\text{Var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$, and #4 $\text{Cov}[\epsilon_i, \epsilon_j] = 0, \quad \forall i \neq j$.

- Given all of the Gauss-Markov assumptions and $\sigma^2 < \infty$, we say that $\hat{\boldsymbol{\beta}}$ is BLUE (Best Linear Unbiased Estimator) for $\boldsymbol{\beta}$ if calculated from OLS or MLE.

- Given sufficient sample size $\hat{\boldsymbol{\beta}}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

Dealing With the Variance

- We want to use some $\text{Est.Var}[\hat{\boldsymbol{\beta}}]$, since σ^2 is generally not known.
- What about calculating the standard error of the regression?
- We will use the sample quantities for estimation:

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

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

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

$$f(\mathbf{e}) = \prod_{i=1}^{n} f(\mathbf{e}_i) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\mathbf{e}'\mathbf{e}/2\sigma^2]$$

based on sample quantities.

Estimating From Sample Quantities

- Population derived variance/covariance matrix: $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- We also know: $\mathbb{E}[\mathbf{e}_i] = \boldsymbol{\epsilon}_i$, although in practice we always have finite samples.
- And by assumption: $\mathbb{E}[\mathbf{e}_i^2] = \text{Var}[\boldsymbol{\epsilon}_i] + (\mathbb{E}[\boldsymbol{\epsilon}_i])^2 = \sigma^2$
  (again using the elementary property $\text{Var}[A] = \mathbb{E}[A^2] - (\mathbb{E}[A])^2$).
- Therefore the sum is $\sum \mathbb{E}[\mathbf{e}_i^2] = \text{tr}(\sigma^2\mathbf{I}) = n\sigma^2$.
- So why not use: $\hat{\sigma}^2 \approx \frac{1}{n}\sum \mathbf{e}_i^2$.
- But:

$$\begin{aligned}\mathbf{e}_i &= \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} \qquad \text{(now insert population values for } \mathbf{Y}_i\text{)}\\ &= (\mathbf{X}_i'\boldsymbol{\beta} + \boldsymbol{\epsilon}_i) - \mathbf{X}_i\hat{\boldsymbol{\beta}}\\ &= \boldsymbol{\epsilon}_i - \mathbf{X}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\end{aligned}$$

  meaning that $\text{plim}[\mathbf{e}_i] = \boldsymbol{\epsilon}_i$ since $\hat{\boldsymbol{\beta}} \underset{n\rightarrow\infty}{\longrightarrow} \boldsymbol{\beta}$.
- So asymptotically this substitution is fine, but is it okay in finite samples?

Some Needed Relations

- Recall that:

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

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

$$\begin{aligned}\mathbf{MY} &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{e}\end{aligned}$$

- From similar calculations we get:

$$\begin{aligned}\mathbf{Me} &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{e} \\ &= \mathbf{e} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \\ &= \mathbf{e} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{e} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{e} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{e}\end{aligned}$$

Some Needed Relations

- Equating $\mathbf{MY} = \mathbf{e}$ and $\mathbf{Me} = \mathbf{e}$ from the last slide, we get $\mathbf{MY} = \mathbf{Me}$

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

$$\begin{aligned}
\mathbf{MY} &= \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\
&= \mathbf{MX}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon} \\
&= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon} \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon} \\
&= \mathbf{M}\boldsymbol{\epsilon}
\end{aligned}$$

- So $\mathbf{e}'\mathbf{e} = (\mathbf{M}\boldsymbol{\epsilon})'\mathbf{M}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}$.

Estimating From Sample Quantities

- This means that we can use $\mathbf{e}'\mathbf{e} = \mathbf{e}'\mathbf{M}\mathbf{M}\mathbf{e} = \mathbf{e}'\mathbf{M}\mathbf{e}$ accordingly:

$$
\begin{aligned}
\mathbb{E}[\mathbf{e}'\mathbf{e}|\mathbf{X}] &= \mathbb{E}[\mathbf{e}'\mathbf{M}\mathbf{e}|\mathbf{X}] \\
&= \mathbb{E}[\mathrm{tr}(\mathbf{e}'\mathbf{M}\mathbf{e})|\mathbf{X}] && (\text{Gauss-Markov assumption \#4: } \mathrm{Cov}[\epsilon_i,\epsilon_j]=0, \forall i \neq j) \\
&= \mathbb{E}[\mathrm{tr}(\mathbf{M}\mathbf{e}'\mathbf{e})|\mathbf{X}] && (\text{property of traces: } \mathrm{tr}(ABC)=\mathrm{tr}(BAC)) \\
&= \mathrm{tr}(\mathbf{M}\mathbb{E}[\mathbf{e}'\mathbf{e}|\mathbf{X}]) && (\mathbf{M} \text{ is fixed for observed } \mathbf{X}) \\
&= \mathrm{tr}(\mathbf{M})\mathbf{I}\sigma^2 && (\text{Gauss-Markov assumption \#3: } \mathrm{Var}[\boldsymbol{\epsilon}]=\sigma^2\mathbf{I}) \\
&= \mathrm{tr}(\mathbf{I}-\mathbf{H})\mathbf{I}\sigma^2 \\
&= [\mathrm{tr}(\mathbf{I}_{n\times n}) - \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')]\mathbf{I}\sigma^2 && (\text{property of traces: } \mathrm{tr}(A-B)=\mathrm{tr}(A)-\mathrm{tr}(B)) \\
&= [\mathrm{tr}(\mathbf{I}_{n\times n}) - k)]\sigma^2 && (\text{for linear models the trace of the hat matrix} \\
& && \qquad\qquad\qquad \text{is the of rank } X) \\
&= [n-k]\sigma^2
\end{aligned}
$$

Estimating From Sample Quantities

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

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

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

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

- The Wald-style traditional linear inference, for the kth coefficient is:

$$z_k = \frac{\hat{\boldsymbol{\beta}}_k - \beta_k^{\text{null}}}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_k^{-1}}} \sim N(0,1),$$

with the assumption that we know σ^2 (which we usually do not).

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

$$\text{we know that } X^2 = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k} \qquad \text{then } \frac{z_k}{\sqrt{X^2/df}} \sim t_{(n-k)}(0)$$

provided the random variables z_k and X^2 are independent.

Estimating From Sample Quantities

- Making the obvious substitution gives:

$$t_{(n-k)} = \frac{\hat{\boldsymbol{\beta}}_k - \beta_k^{\text{null}}}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}}} \times \frac{1}{\sqrt{\frac{(n-k)s^2}{\sigma^2}/(n-k)}} = \frac{\hat{\boldsymbol{\beta}}_k - \beta_k^{\text{null}}}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}}}$$

- Typical (Wald) regression test:

$$H_0\colon\ \beta_k = 0 \qquad\qquad H_1\colon\ \beta_k \neq 0$$

making:

$$t_{(n-k)} = \frac{\hat{\boldsymbol{\beta}}_k - \beta_k^{\text{null}}}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}}} = \frac{\hat{\boldsymbol{\beta}}_k}{SE(\beta_k)}$$

- Alternatives usually look like:

$$H_0\colon\ \beta_k < 7 \qquad\qquad H_1\colon\ \beta_k \geq 7$$

making:

$$t_{(n-k)} = \frac{\hat{\boldsymbol{\beta}}_k - 7}{SE(\beta_k)}$$

Summary Statistics

- $(1-\alpha)$ Confidence Interval for $\hat{\boldsymbol{\beta}}_k$:

$$\left[\hat{\boldsymbol{\beta}}_k - SE(\hat{\boldsymbol{\beta}})t_{\alpha/2,df} \mathbin{:} \hat{\boldsymbol{\beta}}_k + SE(\hat{\boldsymbol{\beta}})t_{\alpha/2,df}\right]$$

- $(1-\alpha)$ Confidence Interval for σ^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 $\hat{\beta}_0$ 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 $\mathbf{x}_i$ *in the sample*:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \qquad\qquad \hat{y}_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$$

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

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

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

$$\begin{aligned}
\text{Var}[\mathbf{e}_i|\mathbf{X},\mathbf{x}_i] &= \text{Var}[\mathbf{x}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})] \\
&= \text{Var}[\mathbf{x}_i\boldsymbol{\beta}] + \text{Var}[\mathbf{x}_i\hat{\beta}] \\
&= s^2 + \mathbf{x}_i\text{Var}[\hat{\boldsymbol{\beta}}]\mathbf{x}_i' \\
&= s^2 + \mathbf{x}_i(s^2(\mathbf{X}\mathbf{X})^{-1})\mathbf{x}_i' \\
&= s^2[1 + \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i']
\end{aligned}$$

Linear Model Predictions/Forecasts

- We want the predicted value for $\mathbf{x}^0$ *not in the sample*:

  $$y^0 = \mathbf{x}^0\boldsymbol{\beta} + \boldsymbol{\epsilon}^0 \qquad\qquad \hat{y}^0 = \mathbf{x}^0\hat{\boldsymbol{\beta}}$$

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

- The *prediction error* is:

  $$e^0 = y^0 - \hat{y}^0 = \mathbf{x}^0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{\epsilon}^0.$$

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

- The Prediction variance is:

  $$\text{Var}[e^0|\mathbf{X},\mathbf{x}^0] = s^2 + \text{Var}[\mathbf{x}^0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})|\mathbf{X},\mathbf{x}^0] = s^2 + s^2(\mathbf{x}^0)(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}^0)'$$

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

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

  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

- Using:

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

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

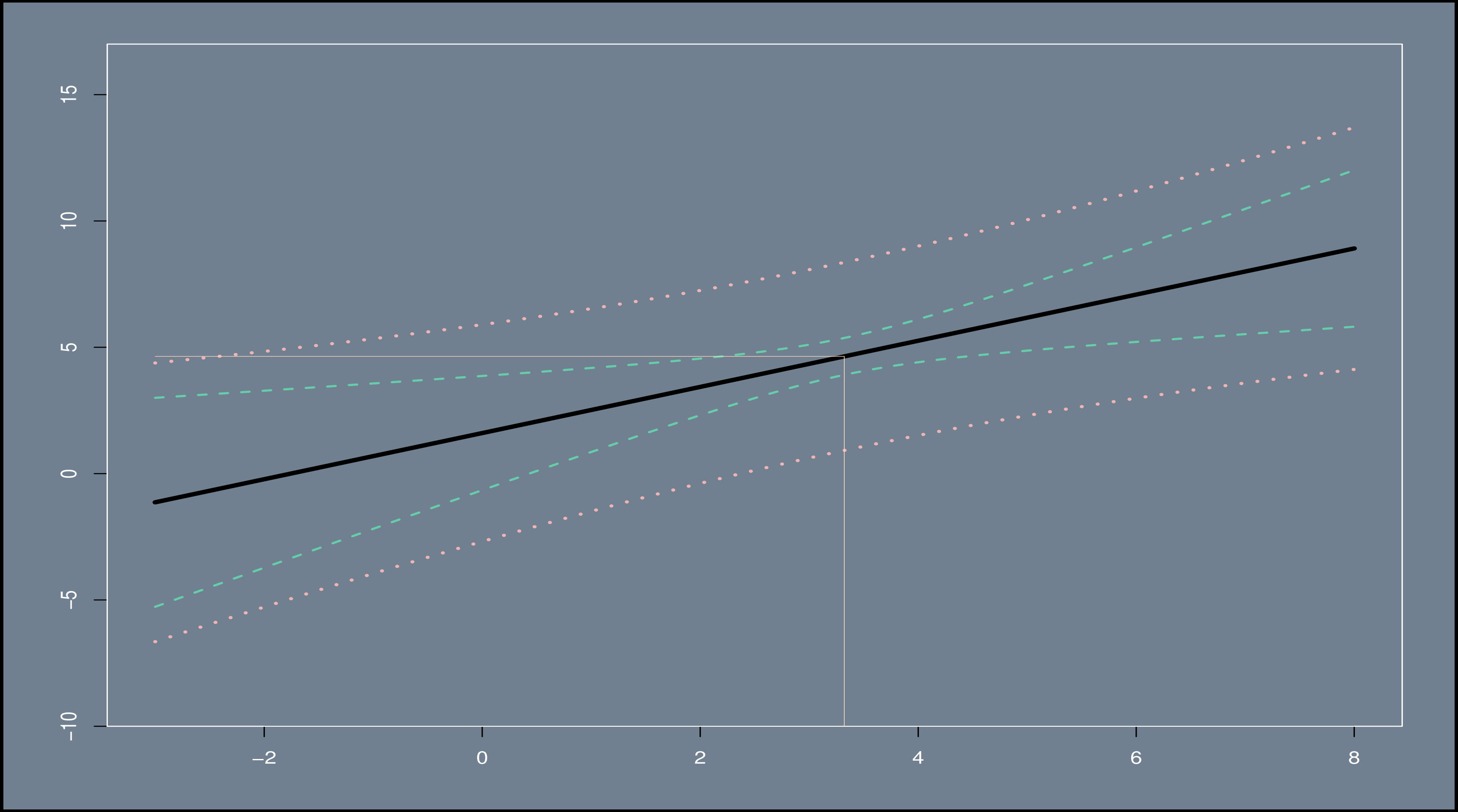
$$CI[\hat{y}^0] = \hat{y}^0 \pm t_{\alpha/2}\sqrt{\text{Var}[e^0|\mathbf{X},\mathbf{x}^0]}.$$

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

- It is important to also distinguish between two interval estimates around the regression line: the CI for $\hat{y} = \mathbf{X}\boldsymbol{\beta}$ 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:

```
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 INTER ALS
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")
dev.off()
```

Multicollinearity Issues

- If one explanatory variable is a linear combination of another then $\text{rank}(\mathbf{X}) = k - 1$.
- Therefore $\text{rank}(\mathbf{X}'\mathbf{X}) = k - 1$ (matrix size $k \times k$), and it is singular and non-invertible.
- Now no parameter estimates are possible, and the 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): add/drop variables, add data cases that break the pattern, restrict the range of some variables, combine variables possibly with PCA.

- Center explanatory variables, or standardize (slope coefficient is interpreted in units of standard deviations of the covariate, the intercept is the mean of the outcome y when all covariate values are zero).

- 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):

  $$\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X} + \mathbf{RI}]^{-1}\mathbf{X}'\mathbf{Y}$$

  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.”

More on Ridge Regression

- Suppose that we are concerned with a single problematic explanatory variable, so that **R** is just a vector of ones with λ in the place.
- Then:

$$\lambda \longrightarrow 0, b_j^{\text{ridge}} \longrightarrow b_j^{\text{OLS}} \qquad \lambda \longrightarrow \infty, b_j^{\text{ridge}} \longrightarrow 0.$$

- Under the assumption of an orthonormal **X** matrix (each column vector has length 1 and is orthogonal to all the other column vectors and the inverse is the transpose):

$$b_j^{\text{ridge}} = \frac{b_j^{\text{OLS}}}{1+\lambda},$$

which shows that increasing λ *shrinks* the estimator towards zero, increasing bias but reducing the variance.

Even More on Ridge Regression

- Define $\mathbf{W} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$, such that:

$$\text{Var}(b) = \sigma^2\mathbf{W}\mathbf{X}'\mathbf{X}\mathbf{W}.$$

- This leads to an important result:

$$\text{bias}(b) = -\lambda\mathbf{W}\boldsymbol{\beta}.$$

- Also there always exists a λ such that the MSE ridge estimator is always less than the the MSE of the regular OLS estimator, where:

$$\text{MSE}(b_j) = \mathbb{E}_{b_j}[(b_j - \beta_j)^2] = \text{Var}(b_j) + \text{bias}(b_j)^2.$$

Simple Ridge Regression Example

```
anaemia <- read.table("https://jeffgill.org/files/jeffgill/files/anaemia.txt",
                      header=TRUE,row.names=1)
library(MASS)
a.lm3.out <- lm.ridge(Hb ~ Age + Menapause + I(Age+rnorm(nrow(anaemia))),
                      data=anaemia)
a.lm3.out$GCV
0.07936452

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:

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

G3: $\mathbf{R}$ is the sample correlation matrix of the observed columns of $\mathbf{X}$, excluding the leading column of 1s. Then $\lim\limits_{n\longrightarrow\infty} \mathbf{R} = \mathbf{C}$, 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:

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

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:

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

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

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

- Taking the limit now as $n \;\rightarrow\; \infty$:

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

- Summarizing:

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

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

$$\therefore\; Est.Asy.Var[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Testing Linear Restrictions

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

  $$H_0\colon\ \beta_k = 0 \qquad \text{versus} \qquad H_1\colon\ \beta_k \neq 0,$$

  for example.

- Most restrictions involve *nested* parameter spaces:

  $$\text{unrestricted:}[\beta_0, \beta_1, \beta_2, \beta_3]$$

  $$\text{restricted:}[\beta_0, 0, \beta_2, \beta_3]$$

  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:

$$\text{unrestricted: } [\beta_0, \beta_1, \beta_2, \beta_3] \qquad \text{restricted: } [\beta_0, 0, \beta_2, \beta_3]$$

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

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

  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:

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

  or in more succinct matrix algebra form: $\underset{(J\times k)}{\mathbf{R}}\underset{(k\times 1)}{\boldsymbol{\beta}} = \underset{(J\times 1)}{\mathbf{q}}$.

- Notes:
  - Each row of $\mathbf{R}$ is one restriction.
  - $J < k$ for $\mathbf{R}$ 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 $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2)$.

- General test where tail values reject the restriction set:

$$H_0\colon \mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{0} \qquad\qquad H_1\colon \mathbf{R}\boldsymbol{\beta} - \mathbf{q} \neq \mathbf{0}$$

Testing Linear Restrictions, Examples

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

$$\mathbf{R} = [0, 0, 0, \underset{j}{1}, \dots, 0, 0, 0], \qquad \mathbf{q} = 0$$

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

$$\mathbf{R} = [0, 0, 0, \underset{j}{1}, \dots, \underset{k}{-1}, \dots, 0, 0, 0], \qquad \mathbf{q} = 0$$

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

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots 0 \\ 0 & 1 & 0 & 0 & \dots 0 \\ 0 & 0 & 1 & 0 & \dots 0 \end{bmatrix} = [\mathbf{I}_3 \mathpunct{:} 0], \qquad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Testing Linear Restrictions, Examples

- A set of coefficients sum to one, $\beta_2 + \beta_3 + \beta_4 = 1$:

$$\mathbf{R} = [0, 1, 1, 1, 0, \ldots, 0], \qquad \mathbf{q} = 1$$

- Several restrictions, $\beta_2 + \beta_3 = 4$, $\beta_4 + \beta_6 = 0$, $\beta_5 = 9$:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \qquad \mathbf{q} = \begin{bmatrix} 4 \\ 0 \\ 9 \end{bmatrix}$$

- All coefficients except the constant are zero:

$$\mathbf{R} = [0{:}\mathbf{I}], \qquad \mathbf{q} = [\mathbf{0}]$$

Testing Linear Restrictions, Examples

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

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q} = \mathbf{m} \approx 0,$$

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

- This makes it straightforward to think about:

$$\mathbb{E}[\mathbf{m}|\mathbf{X}] = R\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathbf{q} = \mathbf{R}\hat{B} - \mathbf{q} = 0$$

$$\text{Var}[\mathbf{m}|\mathbf{X}] = \text{Var}[\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}|\mathbf{X}] = \mathbf{R}\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\mathbf{R}' = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$$

- Wald Test:

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

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 σ^2, so we use a test with s^2, by modifying W:

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

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

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

$$= \frac{1}{J}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'(s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$

- Where:

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

$$X_{\text{denominator}} = X_d/(n-k) \sim \chi^2_{n-k}.$$

- Which is useful because:

$$F = \frac{X_n/J}{X_d/(n-k)} \sim F_{J,n-k}$$

Testing Linear Restrictions, Examples

- With only 1 linear restriction, this simplifies down to:

$$H_0\colon\, r_1\beta_1 + r_2\beta_2 + \ldots + r_k\beta_k = \mathbf{r}\boldsymbol{\beta} = q$$

$$F_{1,n-k} = \frac{\sum_j (r_j\hat{\boldsymbol{\beta}}_j - \mathbf{q})^2}{\sum_j \sum_k r_j r_k Est.Cov.[b_j, b_k]}$$

  and suppose the restriction is on the ℓth coefficient:

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

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

- We use $\hat{\boldsymbol{\beta}}$ to test for $\boldsymbol{\beta}$.

- Giving:

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q} = \hat{\boldsymbol{\beta}}_\ell - \mathbf{q}, \qquad\qquad F_{1,n-k} = \frac{(\hat{\boldsymbol{\beta}}_\ell - \mathbf{q})^2}{Est.Var.[\hat{\boldsymbol{\beta}}_\ell]^{-\frac{1}{2}}}.$$

Testing a Restriction with a Slightly Different Anaemia Model

```
anaemia <-
    read.table("https://jeffgill.org/files/jeffgill/files/anaemia.txt",
    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:

$$\beta_2 = 0, \qquad J = 1, \qquad \mathbf{R} = [0, 0, 1], \qquad \mathbf{q} = 0$$

- Pieces:

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

- Test Statistic:

$$F[J, n-K] = \frac{1}{J}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'(s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$

$$F[1, 20-3] = \frac{1}{1}(0.6)'\left((1.198)[0,0,1]\begin{bmatrix} 0.930 & -0.027 & 0.631 \\ -0.027 & 0.001 & -0.024 \\ 0.631 & -0.024 & 0.845 \end{bmatrix}[0,0,1]'\right)^{-1}(0.6)$$

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

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
                Age Menapause
           20   876        10
Age       876 43074       572
Menapause  10   572        10
```

```
round(solve(t(X) %*% X),3)
                    Age Menapause
           0.930 -0.027     0.631
Age       -0.027  0.001    -0.024
Menapause  0.631 -0.024     0.845
```

```
( 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: $c(\boldsymbol{\beta}) = q$ where $c()$ is some nonlinear function.
- Simple 1-restriction case:

$$z = \frac{c(\hat{\boldsymbol{\beta}})}{est.SE(c(\hat{\boldsymbol{\beta}}))} \sim t_{n-k}$$

  (or equivalently $z^2 \sim F_{1,n-k}$).

- But getting $est.SE(c(\hat{\boldsymbol{\beta}}))$ is hard, so start with a Taylor series expansion:

$$f(b) = f(a) + f'(a)\frac{(b-a)^1}{1!} + f''(a)\frac{(b-a)^2}{2!} + f'''(a)\frac{(b-a)^3}{3!} + \ldots$$

  and then drop all but the first two terms to get an approximation:

$$f(b) \approx f(a) + f'(a)\frac{(b-a)^1}{1!} = f(a) + f'(a)(b-a).$$

- Substituting $\boldsymbol{\beta} = a$, $\hat{\boldsymbol{\beta}} = b$, and $c() = f()$, gives:

$$c(\hat{\boldsymbol{\beta}}) \approx c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

  and also

Testing *Nonlinear* Restrictions

- Now we can calculate the needed variance term:

$$
\begin{aligned}
\text{Var}(c(\hat{\boldsymbol{\beta}})) &= \mathbb{E}\left[c(\hat{\boldsymbol{\beta}})^2\right] - (\mathbb{E}[c(\hat{\boldsymbol{\beta}})])^2 \\
&\approx \mathbb{E}\left[\left(c(\boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^2\right] - (\mathbb{E}[c(\hat{\boldsymbol{\beta}})])^2 \\
&= \mathbb{E}\left[c(\boldsymbol{\beta})^2 - 2c(\boldsymbol{\beta})\left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\right] - (\mathbb{E}[c(\hat{\boldsymbol{\beta}})])^2 \\
&= c(\boldsymbol{\beta})^2 - 2c(\boldsymbol{\beta})^2\left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T (0) + \mathbb{E}\left[\left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\right] - c(\boldsymbol{\beta})^2 \\
&= \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \text{Var}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)
\end{aligned}
$$

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

- This means that we can use sample estimates for $\partial c(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and plug in $s^2(\mathbf{X}'\mathbf{X})^{-1}$ for $\text{Var}(\hat{\boldsymbol{\beta}})$ and then test with a normal distribution, provided reasonable sample size.

Testing a Nonlinear Restriction with the Anaemia Data

- Test: $H_0\colon c(\boldsymbol{\beta}) = q$ where $c(\beta) = \beta^{\frac{1}{2}}$ and $q = 0$, versus $H_1\colon c(\boldsymbol{\beta}) \neq q$.
- Calculate: $\frac{\partial c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{2}(\boldsymbol{\beta})^{-\frac{1}{2}}$.
- 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 a scalar test, use (with $c(\boldsymbol{\beta}) = 0$ as the restriction):

$$z = \frac{c(\hat{\boldsymbol{\beta}}) - 0}{est.SE(c(\hat{\boldsymbol{\beta}}))} \approx \frac{c(\hat{\boldsymbol{\beta}})}{\sqrt{\left(\frac{\partial c(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)' \text{Var}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial c(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)}}$$

$$= \frac{(0.11716)^{\frac{1}{2}}}{\sqrt{\frac{1}{2}(0.11716)^{-\frac{1}{2}}(0.03588)^2\frac{1}{2}(0.11716)^{-\frac{1}{2}}}} = 0.38257 \sim t_{n-k}$$

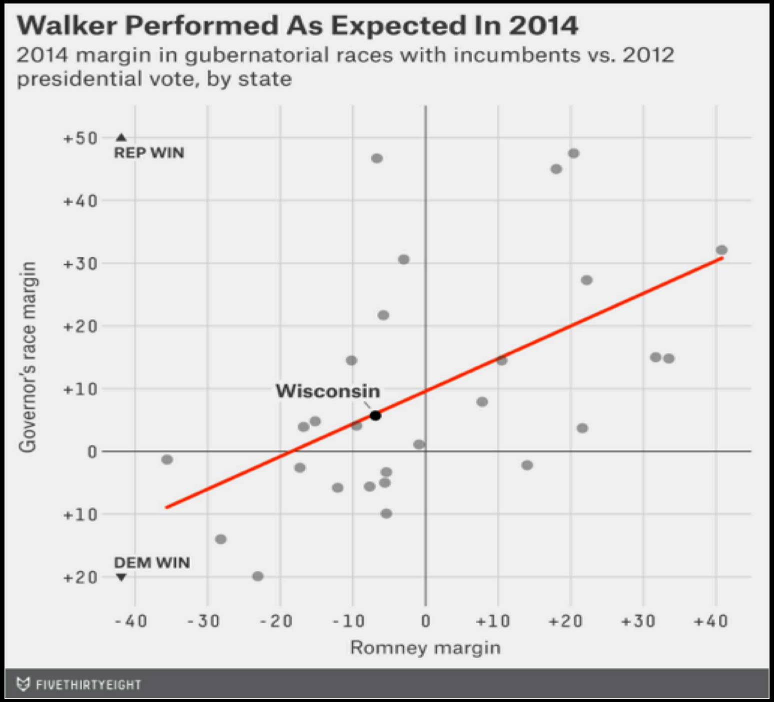
- We *fail* to reject the restriction since:

```
pt(0.38257,17,lower.tail=FALSE)
[1] 0.35339
```

Heteroscedasticity

- If the variance of the residuals in the linear model is not constant over the range of one more **X** variables, then the regression model is heteroscedastic.

- Consider this old example from FiveThirtyEight.com:



Walker Performed As Expected In 2014

2014 margin in gubernatorial races with incumbents vs. 2012 presidential vote, by state

Weighted Least Squares for Known Heteroscedasticity

- A standard technique for compensating for non-constant error variance in LMs is to insert a diagonal matrix of weights, $\boldsymbol{\Omega}$, into the calculation of $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ such that the heteroscedasticity is mitigated.

- The $\boldsymbol{\Omega}$ matrix is created by taking the error variance of the i^{th} case (estimated or known), v_i, and assigning the inverse to the i^{th} diagonal: $\boldsymbol{\Omega}_{ii} = \frac{1}{v_i}$. The idea is that large error variances are reduced by multiplication of the reciprocal.

- Starting with $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$, observe that there is heteroscedasticity in the error term so: $\epsilon_i = \epsilon v_i$, where the shared (minimum) variance is ϵ (i.e. non-indexed), and differences are reflected in the v_i term.

- Really simple example: a heteroscedastic error vector: $\mathbf{E} = [1, 2, 3, 4]$. Then $\epsilon = 1$, and the $\mathbf{v}$ vector is $[1, 2, 3, 4]$. So by the logic above, the $\boldsymbol{\Omega}$ matrix for this example is:

$$\boldsymbol{\Omega} = \begin{bmatrix} \frac{1}{v_1} & 0 & 0 & 0 \\ 0 & \frac{1}{v_2} & 0 & 0 \\ 0 & 0 & \frac{1}{v_3} & 0 \\ 0 & 0 & 0 & \frac{1}{v_4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

Review of Weighted Least Squares

- Premultiply each term by the square root of the $\mathbf{\Omega}$ matrix (a Cholesky factorization given that $\mathbf{A}$ is a positive definite, but greatly simplified here since $\mathbf{\Omega}$ is diagonal).

$$\mathbf{\Omega}^{\frac{1}{2}}\mathbf{Y} = \mathbf{\Omega}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \mathbf{\Omega}^{\frac{1}{2}}\boldsymbol{\epsilon}.$$

- So if the heteroscedasticity in the error term is expressed as the diagonals of a matrix: $\boldsymbol{\epsilon} \sim (0, \sigma^2\mathbf{V})$, then this gives: $\boldsymbol{\epsilon} \sim (0, \mathbf{\Omega}\sigma^2\mathbf{V}) = (0, \sigma^2)$, and the heteroscedasticity is "removed."

- Now instead of minimizing

$$(\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

  we minimize

$$(\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta})'\mathbf{\Omega}(\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

  and the weighted least squares estimator is found by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{Y}.$$

GLS in R

```
diastolic.pressure.df <-
    read.table("https://jeffgill.org/files/jeffgill/files/bloodpressure.txt",
    header=FALSE)
dimnames(diastolic.pressure.df)[[2]] <- c("age","pressure")
summary(diastolic.pressure.df)
```

```
        age           pressure
Min.   :20.00   Min.   : 63.00
1st Qu.:30.25   1st Qu.: 71.00
Median :40.00   Median : 77.00
Mean   :39.57   Mean   : 79.11
3rd Qu.:49.00   3rd Qu.: 85.75
Max.   :59.00   Max.   :109.00
```

GLS in R

```
attach(diastolic.pressure.df)
unweighted.lm <- lm(pressure~age)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.15693    3.99367  14.061  < 2e-16
age          0.58003    0.09695   5.983 2.05e-07
---
Residual standard error: 8.146 on 52 degrees of freedom
Multiple R-Squared: 0.4077,     Adjusted R-squared: 0.3963
F-statistic: 35.79 on 1 and 52 DF,  p-value: 2.05e-07

plot(age,pressure,pch=3)
abline(unweighted.lm)
```

GLS in R

```
# REGRESS ABSOLUTE VALUE RESIDUALS ON PREDICTOR -> SD FUNCTION
resid.fit <- lm(abs(unweighted.lm$residuals)~age)

# OBTAIN FITTED VALUES FOR THE WEIGHTS
weights.fit <- 1/(resid.fit$fitted.values)^2

# USE THESE WEIGHTS FOR A GLS REGRESSION
weighted.lm <- lm(pressure~age,weights=weights.fit)
```

GLS in R

```
summary(weighted.lm)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0230 -0.9939 -0.0327  0.9250  2.2008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.56577    2.52092  22.042  < 2e-16
age          0.59634    0.07924   7.526 7.19e-10
---
Residual standard error: 1.213 on 52 degrees of freedom
Multiple R-Squared: 0.5214,     Adjusted R-squared: 0.5122
F-statistic: 56.64 on 1 and 52 DF,  p-value: 7.187e-10
```

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 Heteroscedascity 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

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

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

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

where:

- ▷ $\mathbf{U}$ is an $M \times k$ matrix,
- ▷ such that each row is produced by $\mathbf{X}_m * \mathbf{e}_m$ 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 $\mathbf{e}_m$ corresponding residuals vector.

Dealing With Heteroscedascity From Group Effects

- A non-clustered, non-robusted model:

```
tweed <- read.table("https://jeffgill.org/files/jeffgill/files/tweed2.txt",
    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 Heteroscedascity From Group Effects

- A model with robust standard errors:

```
lapply(c("sandwich","lmtest","plm"),library, character.only=TRUE)
# GET FUNCTION FROM http://jeffgill.org/files/jeffgill/files/clx.r.txt
source("Class.Multilevel/Code/clx.R")
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 Heteroscedascity 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:

$$VC^{\texttt{Stata}} = \frac{M}{(M-1)}\frac{(N-1)}{(N-K)} \times f_{\text{robust}}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{U}'\mathbf{U})(\mathbf{X}'\mathbf{X})^{-1}$$

- This is easily obtained in R with the following:

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

```
(Intercept)                                  252.06960 219.06530
year                                          -0.12333   0.11027
arrests                                        5.35855   6.04840
factor(attitude)Ethnic/regionalist Separatist -6.37254   3.05351
factor(attitude)Left wing extremist Other     -6.78818   3.14835
factor(attitude)Right wing extremist Other    -4.70957   3.41273
```

Notes on Clustered Standard Errors

- There are many variations with many different names.
- They usually rely on the same "sandwich" structure.
- This does not come without a cost! There can be bias, inflated standard errors, and more.
- Importantly, see:

  Gary King and Margaret E. Roberts. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis* (2015).