

GSERM 2017

Regression III

Linear Regression Review

June 19, 2017 (morning session)

“Regression III”

- “Regression” course
- Texts: Weisberg (2013) and others
- Course materials at the github repo:
<https://github.com/PrisonRodeo/GSERM-2017-git>
- Software: R > Stata
- “Introduction to R and RStudio” today after the break...
- Grading: Two homework assignments plus a final examination

“Regression,” conceptually:

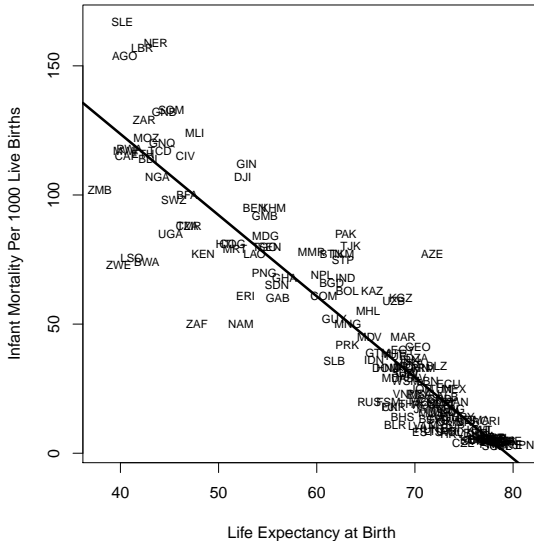
$$\Pr(Y|\mathbf{X}) = f(\mathbf{X})$$

Two important things:

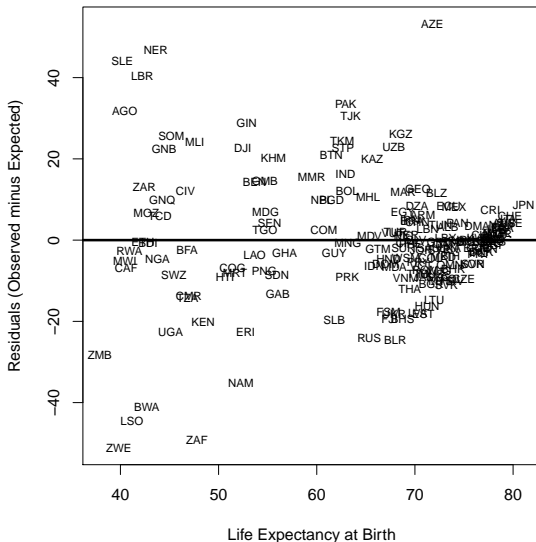
- The distribution of Y is *conditional on all variables in \mathbf{X}* , and
- The conditional distribution of Y is conditional on the *joint distribution* of the elements of \mathbf{X} .

→ Regression is hard...

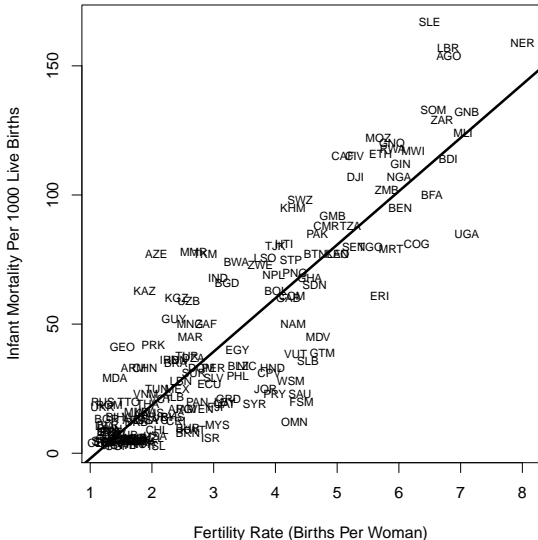
Infant Mortality and Life Expectancy



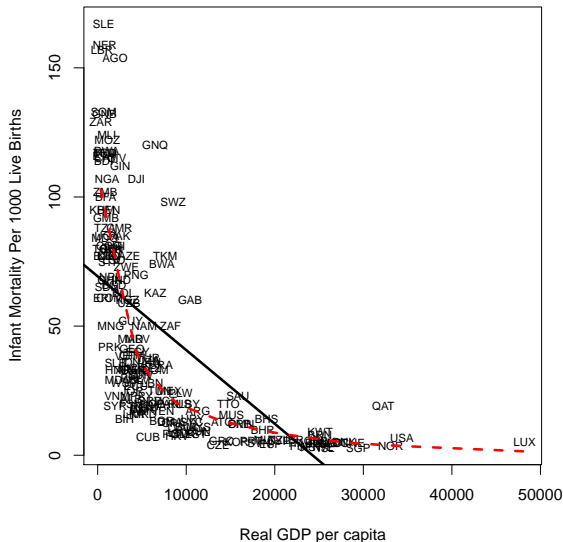
Infant Mortality and Life Expectancy: “Residuals”



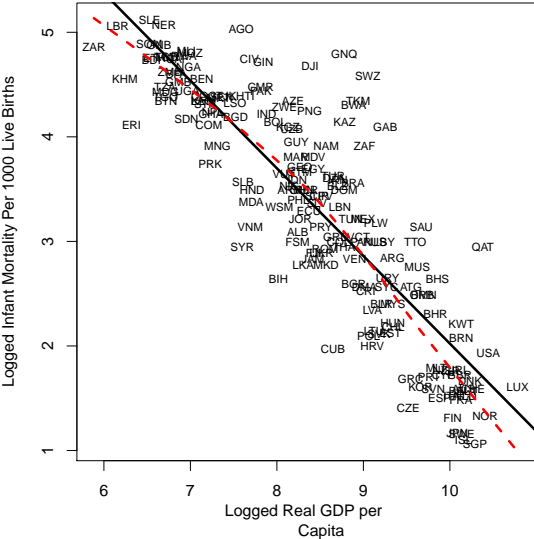
Infant Mortality and Fertility



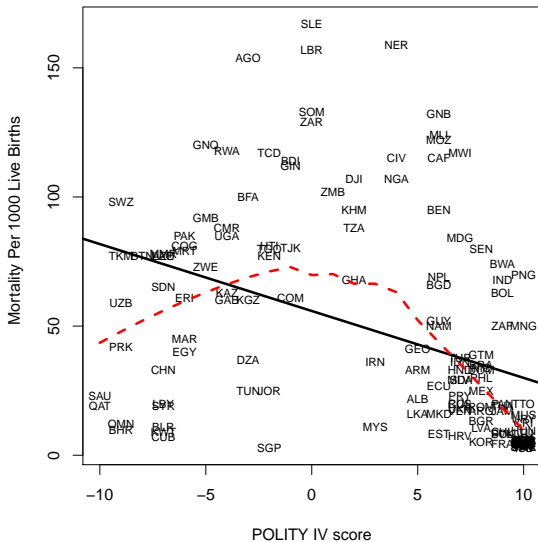
Infant Mortality and Wealth



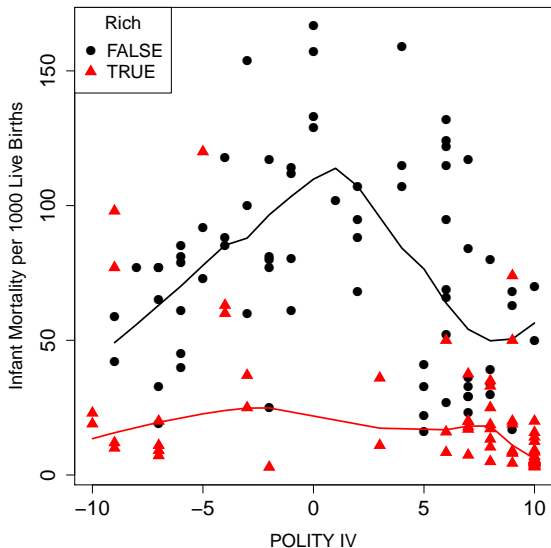
(Logged) Infant Mortality and (Logged) Wealth



Infant Mortality and Democracy



Infant Mortality, (Dichotomized) Wealth, and Democracy



Wherefore Regression?

	Description	Explanation	Prediction
Task	Summarize data	Correlation/causation	Forecast OOS / future data
Emphasis	Data	Theory / Hypotheses	Outcomes
Focus	Univariate	Multivariate	Multivariate
Typical Application	Summarize / “reduce” data	Discuss marginal associations between predictors and an outcome of interest	Optimize out-of- sample predictive power / minimize prediction error

$$Y_i = \mu + u_i \quad (1)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

so:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (2)$$

Goals:

- Estimate $\hat{\beta}_0$ and $\hat{\beta}_1$
- Estimate the *variability* $\hat{\beta}_0$ and $\hat{\beta}_1$

Linear Regression (continued)

If we have $\hat{\beta}_0$ and $\hat{\beta}_1$, then:

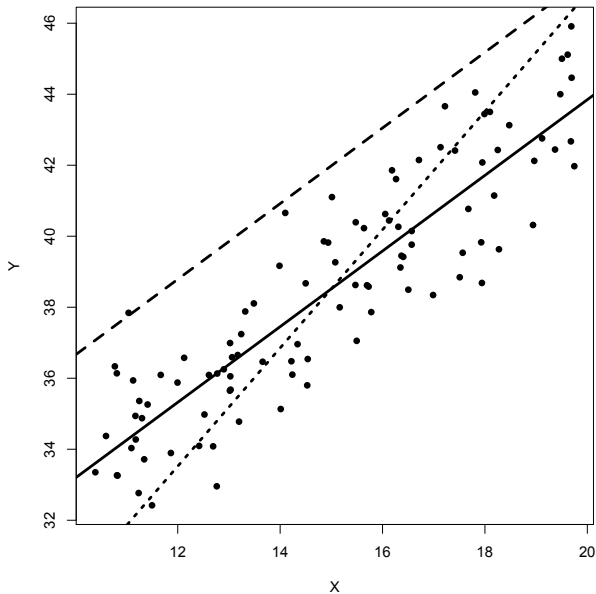
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (3)$$

and

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \end{aligned} \quad (4)$$

Q: How to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$?

Scatterplot: X and Y (with regression lines)



Ordinary Least Squares

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize $\hat{S} = \sum_{i=1}^N \hat{u}_i^2$.

$$\begin{aligned}\hat{S} &= \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^N (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ &= \sum_{i=1}^N (Y_i^2 - 2Y_i\hat{\beta}_0 - 2Y_i\hat{\beta}_1 X_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 X_i + \hat{\beta}_1^2 X_i^2)\end{aligned}$$

Differentiate:

$$\begin{aligned}\frac{\partial \hat{S}}{\partial \hat{\beta}_0} &= \sum_{i=1}^N (-2Y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 X_i) \\ &= -2 \sum_{i=1}^N (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= -2 \sum_{i=1}^N \hat{u}_i\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \hat{S}}{\partial \hat{\beta}_1} &= \sum_{i=1}^N (-2Y_i X_i + 2\hat{\beta}_0 X_i + 2\hat{\beta}_1 X_i^2) \\ &= -2 \sum_{i=1}^N (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i \\ &= -2 \sum_{i=1}^N \hat{u}_i X_i\end{aligned}$$

Yields:

$$\sum_{i=1}^N Y_i = N\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^N X_i$$

and

$$\sum_{i=1}^N Y_i X_i = \hat{\beta}_0 \sum_{i=1}^N X_i + \hat{\beta}_1 \sum_{i=1}^N X_i^2$$

Solving yields:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ &= \frac{\text{Covariance of } X \text{ and } Y}{\text{Variance of } X}\end{aligned}\tag{5}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}\tag{6}$$

$$\text{Var}(\hat{\beta}_1)$$

$$u_i \sim \text{i.i.d. } N(0, \sigma^2)$$

meaning:

$$\text{Var}(Y|X, \beta) = \sigma^2$$

so:

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var} \left[\frac{\sum_{i=1}^N (X_i - \bar{X}) Y_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \right] \\ &= \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]^2 \sum (X_i - \bar{X})^2 \text{Var}(Y) \\ &= \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]^2 \sum (X_i - \bar{X})^2 \sigma^2 \\ &= \frac{\sigma^2}{\sum (X_i - \bar{X})^2}. \end{aligned}$$

$$\text{Var}(\hat{\beta}_0) \text{ and } \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

Similarly:

$$\text{Var}(\hat{\beta}_0) = \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \sigma^2$$

and :

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \sigma^2$$

Important Things

- $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1) \propto \sigma^2$
- $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1) \propto -\sum(X_i - \bar{X})$
- $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1) \propto -N$
- $\text{sign}[\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)] = \text{sign}(\bar{X})$

If $u_i \sim N(0, \sigma^2)$, then:

$$\hat{\beta}_0 \sim N[\beta_0, \text{Var}(\hat{\beta}_0)]$$

and

$$\hat{\beta}_1 \sim N[\beta_1, \text{Var}(\hat{\beta}_1)]$$

Means:

$$\begin{aligned} z_{\hat{\beta}_1} &= \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\text{Var}(\hat{\beta}_1)}} \\ &= \frac{(\hat{\beta}_1 - \beta_1)}{\text{s.e.}(\hat{\beta}_1)} \\ &= \sim N(0, 1) \end{aligned}$$

A Small Problem...

$$\sigma^2 = ???$$

Solution: use

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N - k}$$

Gives:

$$\widehat{\text{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2},$$

and

$$\widehat{\text{Var}(\hat{\beta}_0)} = \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \hat{\sigma}^2$$

$$\begin{aligned}
 \widehat{\text{s.e.}}(\hat{\beta}_1) &= \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} \\
 &= \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}} \\
 &= \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}
 \end{aligned}$$

implies:

$$\begin{aligned}
 t_{\hat{\beta}_1} &\equiv \frac{(\hat{\beta}_1 - \beta_1)}{\widehat{\text{s.e.}}(\hat{\beta}_1)} = \frac{(\hat{\beta}_1 - \beta_1)}{\frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}} \\
 &= \frac{(\hat{\beta}_1 - \beta_1) \sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}} \\
 &\sim t_{N-k}
 \end{aligned}$$

Predictions and Variance

Point prediction:

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k$$

Y_k is unbiased:

$$\begin{aligned} E(\hat{Y}_k) &= E(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\ &= E(\hat{\beta}_0) + X_k E(\hat{\beta}_1) \\ &= \beta_0 + \beta_1 X_k \\ &= E(Y_k) \end{aligned}$$

Variability:

$$\begin{aligned} \text{Var}(\hat{Y}_k) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\ &= \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \sigma^2 + \left[\frac{\sigma^2}{\sum (X_i - \bar{X})^2} \right] X_k^2 + 2 \left[\frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \sigma^2 \right] X_k \\ &= \sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \end{aligned}$$

Variability of Predictions

$$\text{Var}(\hat{Y}_k) = \sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

means that $\text{Var}(\hat{Y}_k)$:

- Decreases in N
- Decreases in $\text{Var}(X)$
- Increases in $|X - \bar{X}|$

Standard error of the prediction:

$$\widehat{\text{s.e.}(\hat{Y}_k)} = \sqrt{\sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]}$$

→ (e.g.) confidence intervals:

$$95\% \text{ c.i.}(\hat{Y}_k) = \hat{Y}_k \pm [1.96 \times \widehat{\text{s.e.}(\hat{Y}_k)}]$$

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(\hat{Y} + \hat{u}) \\ &= \text{Var}(\hat{Y}) + \text{Var}(\hat{u}) + 2 \text{Cov}(\hat{Y}, \hat{u}) \\ &= \text{Var}(\hat{Y}) + \text{Var}(\hat{u})\end{aligned}$$

$$\begin{array}{ccccc}\mathbf{TSS} & = & \mathbf{MSS} & + & \mathbf{RSS} \\ \text{("Total")} & & \text{("Estimated," or "Model")} & & \text{("Residual")}\end{array}$$

$$\begin{aligned} R^2 &= \frac{\text{MSS}}{\text{TSS}} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{RSS}}{\text{TSS}} \\ &= 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \end{aligned}$$

R-squared:

- is “the proportion of variance explained”
- $\in [0, 1]$
 - $R^2 = 1.0 \equiv$ a “perfect (linear) fit”
 - $R^2 = 0 \equiv$ no (linear) $X - Y$ association

For a single X ,

$$\begin{aligned} R^2 &= \hat{\beta}_1^2 \frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} \\ &= r_{XY}^2 \end{aligned}$$

$$R_{adj.}^2 = 1 - \frac{(1 - R^2)(N - c)}{(N - k)}$$

where $c = 1$ if there is a constant in the model and $c = 0$ otherwise.

$R_{adj.}^2$:

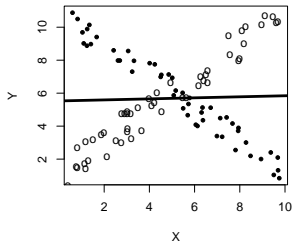
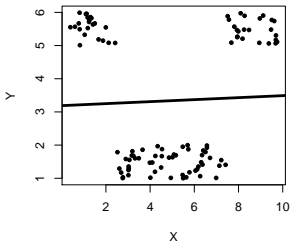
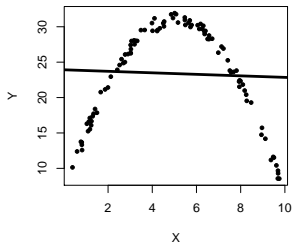
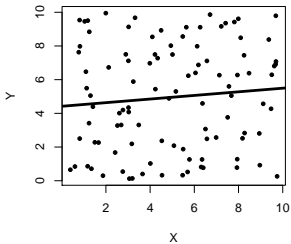
- $R_{adj.}^2 \rightarrow R^2$ as $N \rightarrow \infty$
- $R_{adj.}^2$ can be > 1 , or < 0 ...
- $R_{adj.}^2$ increases with model "fit," but
- The extent of that increase is discounted by a factor proportional to the number of covariates.

- Standard Error of the Estimate:

$$\text{SEE} = \sqrt{\frac{\text{RSS}}{N - k}}$$

- F -tests
- ROC / AUC
- Graphical methods

Caution: Different Ways to get $R^2 = 0$



Linear Regression: k Predictors

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{KN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}.$$

Residuals:

$$\mathbf{u} = \mathbf{Y} - \mathbf{X}\beta$$

The inner product of \mathbf{u} :

$$\begin{aligned} \mathbf{u}'\mathbf{u} &= \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \\ &= u_1^2 + u_2^2 + \dots + u_N^2 \\ &= \sum_{i=1}^N u_i^2 \end{aligned}$$

$$\begin{aligned}\mathbf{u}'\mathbf{u} &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y}' + \beta'\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

Now get:

$$\frac{\partial \mathbf{u}'\mathbf{u}}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta$$

Solve:

$$\begin{aligned}-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta &= 0 \\ -\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\beta &= 0 \\ \mathbf{X}'\mathbf{X}\beta &= \mathbf{X}'\mathbf{Y} \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \beta &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\end{aligned}$$

The Importance of $\mathbf{V}(\hat{\beta})$

$$\begin{aligned}\mathbf{V}(\hat{\beta}) &= E[\hat{\beta} - E(\hat{\beta})]^2 \\ &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\}\end{aligned}$$

Rewrite:

$$\begin{aligned}\mathbf{V}(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

The Importance of $\mathbf{V}(\hat{\beta})$

Taking expectations:

$$\begin{aligned}\mathbf{V}(\hat{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Estimating $\mathbf{V}(\hat{\beta})$

Empirical estimate:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{N - K}$$

Yields:

$$\widehat{\mathbf{V}(\hat{\beta})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

1. Zero Expectation Disturbances

$$E(\mathbf{u}) = \mathbf{0}$$

2. Homoscedasticity / No Error Correlation

$$\begin{aligned} \mathbf{uu}' &= \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix} \\ &= \begin{bmatrix} u_1^2 & u_1 u_2 & \cdots & u_1 u_N \\ u_2 u_1 & u_2^2 & \cdots & u_2 u_N \\ \vdots & \vdots & \ddots & \vdots \\ u_N u_1 & u_N u_2 & \cdots & u_N^2 \end{bmatrix} \end{aligned}$$

Expectation must be:

$$E(\mathbf{uu}') = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

3. "Fixed" \mathbf{X} ...

- No *measurement error* in the \mathbf{X} s, and
- $\text{Cov}(\mathbf{X}, \mathbf{u}) = \mathbf{0}$.

4. \mathbf{X} is full column rank.

Means:

- no exact linear relationship among \mathbf{X} , and
- $K < N$.

5. Normal Disturbances

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$


Under these assumptions, the OLS estimate of $\hat{\beta}$ is:

- **Unbiased**
- **Fully Efficient**

(i.e., “BLUE”)

BREAK

- Preferred: R + RStudio
- Also viable: Stata
- Highly discouraged: SAS, SPSS/PSPP, etc.



The image shows a screenshot of the R Console window on a macOS system. The title bar at the top reads "R Console". Below the title bar is a menu bar with standard macOS icons (stop, back, forward, search, etc.). The main content area displays the R version 3.3.1 (2016-06-21) -- "Bug in Your Hair" copyright notice, followed by the R Foundation's disclaimer and instructions on how to use R. The console also shows the R.app GUI version 1.68 (7238) and the workspace restored from the user's home directory. The prompt ">|" is visible at the bottom.

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

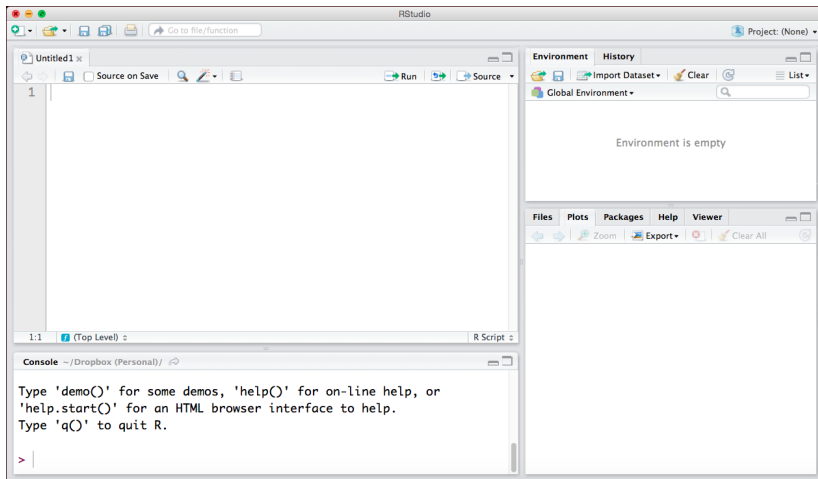
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.68 (7238) x86_64-apple-darwin13.4.0]

[Workspace restored from /Users/cuz101cl/.RData]
[History restored from /Users/cuz101cl/.Rapp.history]

> |
```



RStudio (annotated)

Source window:

- Click here to save your source code. Save often!
- Source on Save
- Run

This is the "Source" window.

- It's the place where you'll type the code that will then be sent to R.
- It's basically a text editor. You can open text files of any kind here if you want.
- Files that appear here end in (and should be saved with) the extension ".R" (as in "MyCode.R").

You'll spend most of your time working here.

Environment window:

This is the "Environment" window. It is where you can find all the various "objects" that you create, grouped by object type (data frames, lists, graphs, etc.). Environment is empty.

There's also a "History" tab above; switching to that will show what has transpired in the Console window recently.

Console window:

This is the "working directory." Anything you save will be saved here, unless you tell the program to save it somewhere else.

Console - /Dropbox (Personal) /

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

This is the "Console." When you run the code in the Source window, the results that aren't graphics appear here.

Other tabs:

- Files
- Plots
- Packages
- Help
- Viewer

Zoom Export Clear All

This:

```
> table(df$X)
```

... means “Type the phrase ‘table(df\$X)’ on the command line,” or – equivalently – “Type the phrase ‘table(df\$X)’ into your Source code, and then run it.”

More often, you'll see:

```
with(df, plot(Y~X,pch=19,col="red")) # draw a scatterplot
abline(h=0,lty=2) # add a horizontal line at zero
abline(v=0,lty=2) # add a vertical line at zero
text(df$X,df$Y,labels=df$names,pos=1) # add labels
```

... which means “Put this block of text into your Source code, and then run it.”

Note:

- R / RStudio ignores line breaks
- Anything to the right of a “#” is a comment

Very basic R examples...
(see GSERM-2017-R-Intro.R)

Help For Learning R(Studio)

In rough order of preference:

- Quick-R (<http://www.statmethods.net/>)
- The “Level-Zero” R Tutorial (doesn’t integrate RStudio, but is otherwise very good)
- Statistics with R
- The Do It Yourself Introduction to R
- Also be sure to consult the Regression III “Useful R Resources” guide (on GitHub).

Example Data: Infant Mortality

```
> url <- getURL("https://raw.githubusercontent.com/PrisonRodeo/
  GSERM-2017-git/master/Data/CountryData2000.csv")
> Data <- read.csv(text = url) # read the "votes" data
> rm(url)
>
> # Summary statistics
>
> # install.packages("psych") <- Install psych package, if necessary
> library(psych)

> with(Data, describe(infantmortalityperK))
  vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
1     1 179 43.83 40.39     29   38.38 34.26 2.9 167 164.1     1     0.06 3.02

> with(Data, describe(DPTpct))
  vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
1     1 181 81.71 19.77     90   85.23 11.86 24  99   75 -1.31     0.57 1.47
```

OLS Regression

```
> IMDPT<-lm(infantmortalityperK~DTPpct,data=Data,na.action=na.exclude)
> summary.lm(IMDPT)
```

Call:

```
lm(formula = infantmortalityperK ~ DTPpct, data = Data)
```

Residuals:

Min	1Q	Median	3Q	Max
-56.801	-16.328	-5.105	11.777	86.590

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	173.2771	8.4893	20.41	<2e-16 ***
DTPpct	-1.5763	0.1009	-15.62	<2e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 26.19 on 175 degrees of freedom

(14 observations deleted due to missingness)

Multiple R-squared: 0.5824, Adjusted R-squared: 0.58

F-statistic: 244.1 on 1 and 175 DF, p-value: < 2.2e-16

Analysis of Variance

```
> anova(IMDPT)
```

```
Analysis of Variance Table
```

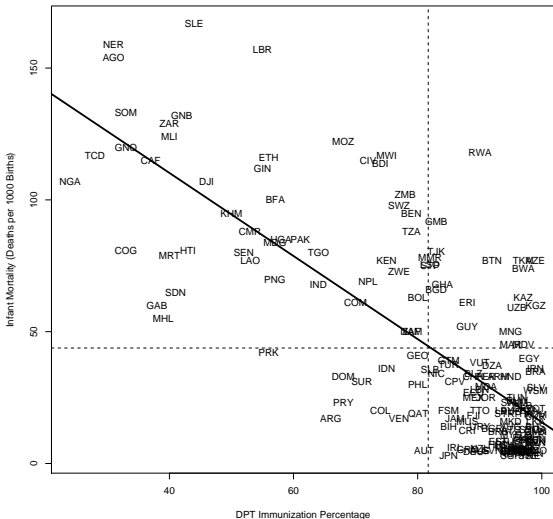
```
Response: infantmortalityperK
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DPTpct	1	167423	167423	244.09	< 2.2e-16 ***
Residuals	175	120033	686		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

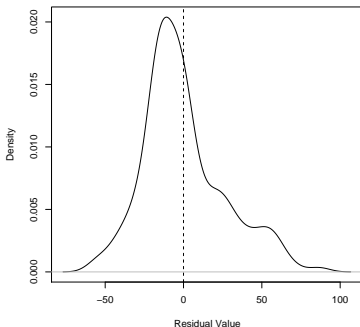
Regression of Infant Mortality on DPT Immunization Rates



Fitted Values, Residuals, etc.

```
> # Residuals (u):  
> Data$IMDPTres <- with(Data, residuals(IMDPT))  
> describe(Data$IMDPTres)
```

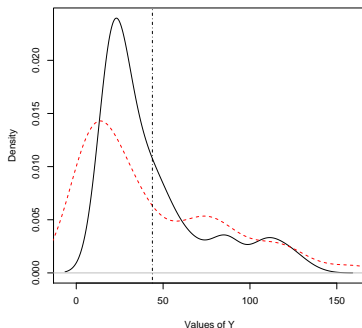
	var	n	mean	sd	median	mad	min	max	range	skew	kurtosis	se
1	1	177	0	26.12	-5.1	19.42	-56.8	86.59	143.4	0.75	0.44	1.96



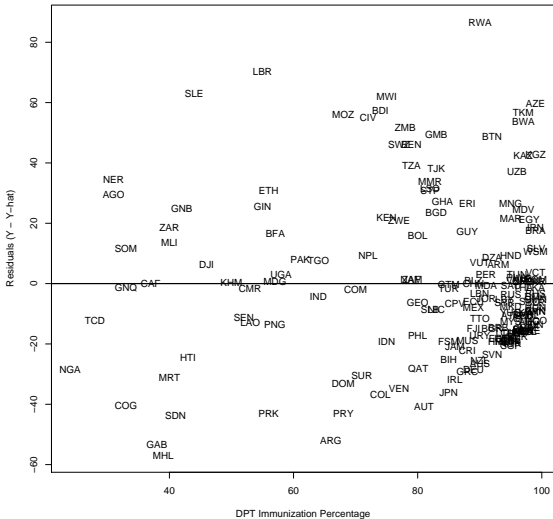
```
> # Fitted Values:  
> Data$IMDPThat<-fitted.values(IMDPT)  
> describe(Data$IMDPThat)
```

	var	n	mean	sd	median	mad	min	max	range	skew	kurtosis	se
1	1	177	44.26	30.84	31.41	18.7	17.22	135.4	118.2	1.3	0.59	2.32

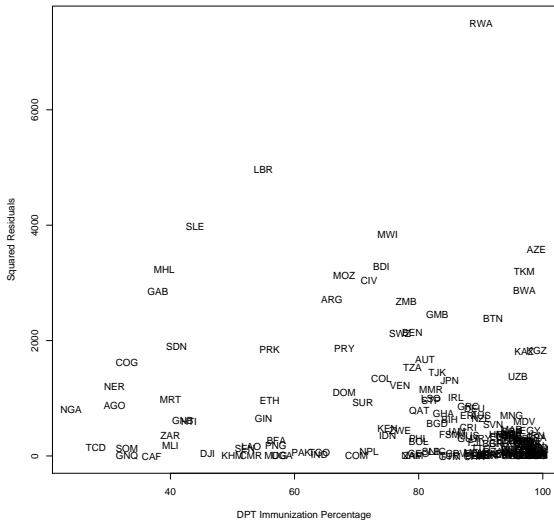
Density Plot: Actual (Y) and Fitted Values (\hat{Y})



Regression Residuals (\hat{u}) vs. DPT Percentage



Squared Residuals vs. DPT Percentage



$\text{Var}(\hat{\beta})$:

```
> vcov(IMDPT)
```

	(Intercept)	DPTpct
(Intercept)	72.0677	-0.83317
DPTpct	-0.8332	0.01018

95 percent c.i.s:

```
> confint(IMDPT)
```

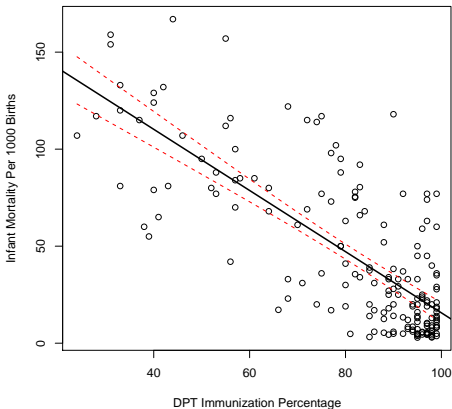
	2.5 %	97.5 %
(Intercept)	156.523	190.032
DPTpct	-1.775	-1.377

```
> SEs<-predict(IMDPT,interval="confidence")
> SEs
```

	fit	lwr	upr
1	25.10	20.53	29.68
3	17.22	12.05	22.40
4	23.53	18.84	28.21
.			
.			
<rows omitted>			
.			
.			
189	21.95	17.15	26.75
190	39.29	35.36	43.23
191	17.22	12.05	22.40

A Plot, With CIs

Scatterplot of Infant Mortality and DPT Immunizations, along with Least-Squares Line and 95% Prediction Confidence Intervals



Multivariate Example: Africa Data

```
> library(RCurl)
> temp<-getURL("https://raw.githubusercontent.com/PrisonRodeo/GSERM-2017-git/master/Data/africa2001.csv")
> Data<-read.csv(text=temp, header=TRUE)
> Data<-with(Data, data.frame(adrate, polity,
+                             subsaharan=as.numeric(subsaharan), muslperc, literacy))
```

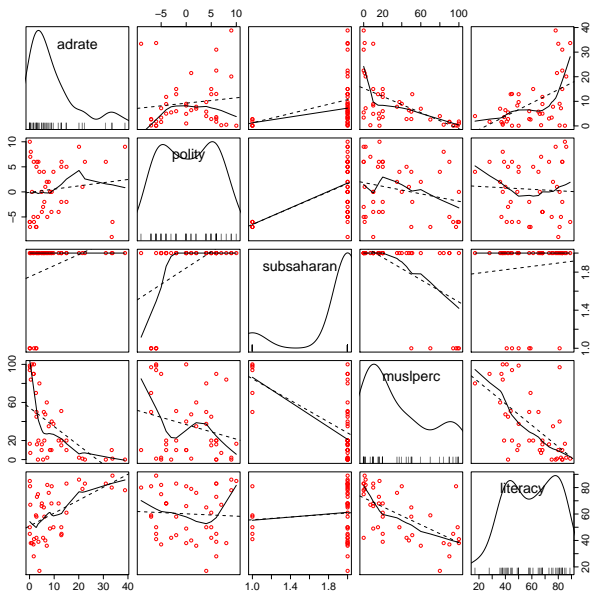
```
> summary(Data)
```

adrate	polity	subsaharan	muslperc	literacy
Min. : 0.100	Min. : -9.0000	Min. : 1.00	Min. : 0.00	Min. : 17.00
1st Qu.: 2.700	1st Qu.: -4.5000	1st Qu.: 2.00	1st Qu.: 10.00	1st Qu.: 43.00
Median : 6.000	Median : 0.0000	Median : 2.00	Median : 20.00	Median : 61.00
Mean : 9.365	Mean : 0.5116	Mean : 1.86	Mean : 35.96	Mean : 60.07
3rd Qu.: 12.900	3rd Qu.: 5.5000	3rd Qu.: 2.00	3rd Qu.: 55.50	3rd Qu.: 78.50
Max. : 38.800	Max. : 10.0000	Max. : 2.00	Max. : 100.00	Max. : 89.00

```
> cor(Data)
```

	adrate	polity	subsaharan	muslperc	literacy
adrate	1.0000000	0.11794182	0.33129420	-0.5709233	0.51489444
polity	0.1179418	1.00000000	0.52819844	-0.2391715	-0.05079354
subsaharan	0.3312942	0.52819844	1.00000000	-0.5772513	0.09472968
muslperc	-0.5709233	-0.23917151	-0.57725134	1.0000000	-0.61960385
literacy	0.5148944	-0.05079354	0.09472968	-0.6196039	1.00000000

Africa Data



A Regression

```
> model<-lm(adrate~polity+subsaharan+muslperc+literacy,data=Data)
> summary(model)
```

Call:

```
lm(formula = adrate ~ polity + subsaharan + muslperc + literacy,
    data = Data)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-15.4681	-4.3947	-0.5251	3.4246	22.9358

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.39843	14.94744	-0.294	0.7702
polity	-0.01390	0.27969	-0.050	0.9606
subsaharan	3.72969	5.43093	0.687	0.4964
muslperc	-0.08689	0.06282	-1.383	0.1747
literacy	0.16575	0.09433	1.757	0.0869 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.264 on 38 degrees of freedom

Multiple R-squared: 0.3771, Adjusted R-squared: 0.3115

F-statistic: 5.751 on 4 and 38 DF, p-value: 0.001013

Variance-Covariance Matrix of $\hat{\beta}$

```
> options(digits=4)
> vcov(model)
```

	(Intercept)	polity	subsaharan	muslperc	literacy
(Intercept)	223.4259	1.088030	-72.2628	-0.771309	-1.002421
polity	1.0880	0.078229	-0.6642	-0.000293	0.001968
subsaharan	-72.2628	-0.664212	29.4950	0.206067	0.171765
muslperc	-0.7713	-0.000293	0.2061	0.003946	0.004098
literacy	-1.0024	0.001968	0.1718	0.004098	0.008898

Test $H_0 : \beta_{\text{polity}} = \beta_{\text{subsaharan}} = 0$:

```
> library(lmtest)
> modelsmall<-lm(adrate~muslperc+literacy,data=Data)
> waldtest(model,modelsmall)
```

Wald test

Model 1: adrate ~ polity + subsaharan + muslperc + literacy

Model 2: adrate ~ muslperc + literacy

	Res.Df	Df	F	Pr(>F)
1	38			
2	40	-2	0.27	0.76

Test $H_0 : \beta_{\text{muslperc}} = 0.1$:

```
> library(car)
> linearHypothesis(model,"muslperc=0.1")
```

Linear hypothesis test

Hypothesis:
muslperc = 0.1

Model 1: restricted model

Model 2: adrate ~ polity + subsaharan + muslperc + literacy

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	39	3200				
2	38	2595	1	605	8.85	0.0051 **

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Test $H_0 : \beta_{\text{literacy}} = \beta_{\text{muslperc}}$:

```
> linearHypothesis(model,"literacy=muslperc")
```

Linear hypothesis test

Hypothesis:

- muslperc + literacy = 0

Model 1: restricted model

Model 2: adrate ~ polity + subsaharan + muslperc + literacy

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	39	3534				
2	38	2595	1	938	13.7	0.00067 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1