GSERM 2024

Regression for Publishing

June 17, 2024

- Instructor: Prof. Christopher Zorn
  - Email: zorn@psu.edu
  - Phone: +1-803-553-4077
  - Twitter / Instagram / etc.: @prisonrodeo

- Class: June 17-21, 2024, 09:15 - 15:15 CET, at the University of St. Gallen SQUARE building, room 11-0101, "Gallus."

- The course outline / syllabus is here.

- More important: The syllabus, slides, readings, code, data, etc. are all available on the course github repo (viewable at https://github.com/PrisonRodeo/GSERM-RFP-2024).

Assessment

Evaluation at GSERM isn't easy... the plan:

- One "homework exercise"
  - Practical exercise – "real" data analysis and discussion
  - Assigned Tuesday (June 18); due Friday (June 21)
  - Worth 300 possible points

- Final Examination
  - Multiple essay-style questions + "real" data analysis
  - Some choice of questions to answer
  - Assigned Friday (June 21) around 12:00 CET
  - Due either Friday, June 21, 2024 by 20:00 CET ("in-class" alternative) or Friday, June 28, 2024 ("take-home" alternative)
  - Worth 700 possible points

- Total course = 1000 possible points

- Grades assessed on Swiss (1 - 6) scale

Software

R

- All examples, plots, etc. are generated using R
- Current version is 4.4.0
- Desktop: Be sure to get the RStudio / Posit IDE...
- Alternatively: Can be run in a browser, using Posit Cloud
- The course Github repo contains a bit of introductory code for people who may never have used R, and a list of R resources.

Stata

- Current version is 18
- Main linear regression command is `-regress-` (or `reg`)

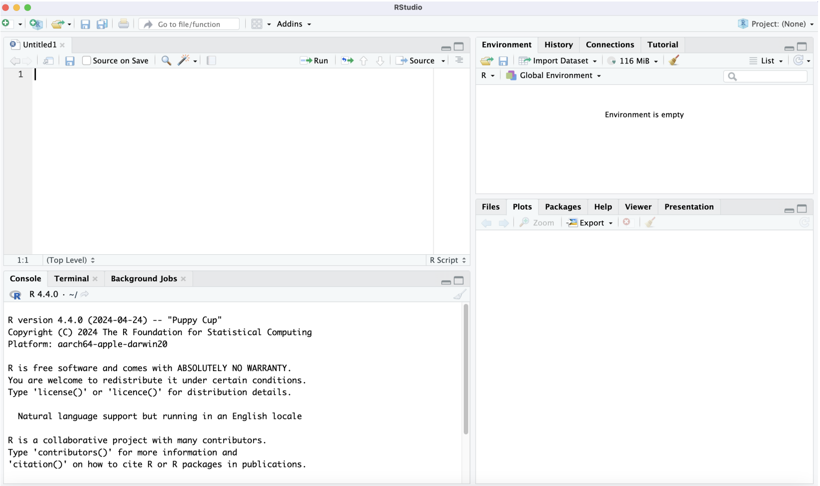
R

R Console

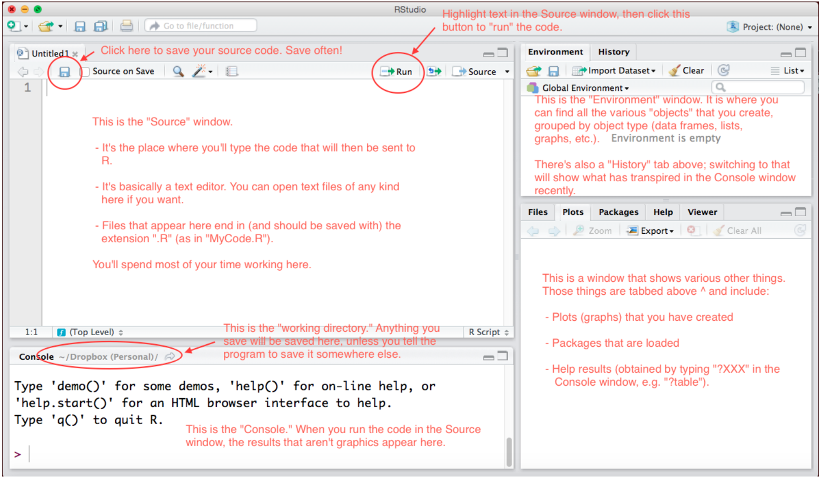
Help Search

```
R version 4.4.0 (2024-04-24) -- "Puppy Cup"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20

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 1.80 (8376) aarch64-apple-darwin20]

[History restored from /Users/cuz10/.Rapp.history]

> 
```

RStudio


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
116 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Console
Terminal
Background Jobs
R 4.4.0 · ~/
R version 4.4.0 (2024-04-24) -- "Puppy Cup"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
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.

RStudio (annotated)



Things We Will And Won't Do

Will: "Regression":

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

Won't: Multivariate regression:

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

Won't: Measurement (e.g. PCA, factor analysis, IRT, etc.):

$$\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$$

Won't: Classification:

- Cluster Analysis / Network Models / etc.
- Classification and Regression Trees $\rightarrow$ Random Forests.
- Pattern Recognition
- Machine Learning (beyond regression), Support Vector Machines, etc.

Friday = "Participants' Choice"

PennState

Please rank order these possible choices for topics for the Friday (June 21) class session (1 = most preferred to 8 = least preferred). When you have finished ranking them, please click on the arrow below.

- Extended GLMs
- Panel / TSCS Models
- Causal Inference
- Measurement Models
- Bayesian Statistics
- Sample Selection Models
- Survival Analysis
- Analysis of Text / LLMs

Why 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

"Regression," conceptually:

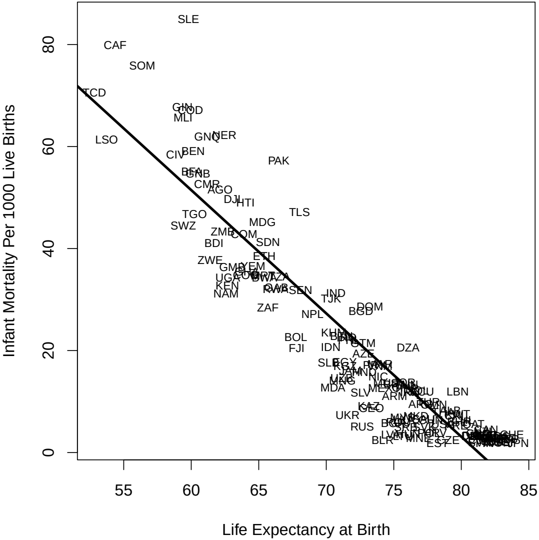
$$\Pr(Y|\mathbf{X}) = f(\mathbf{X})$$

Two important things:

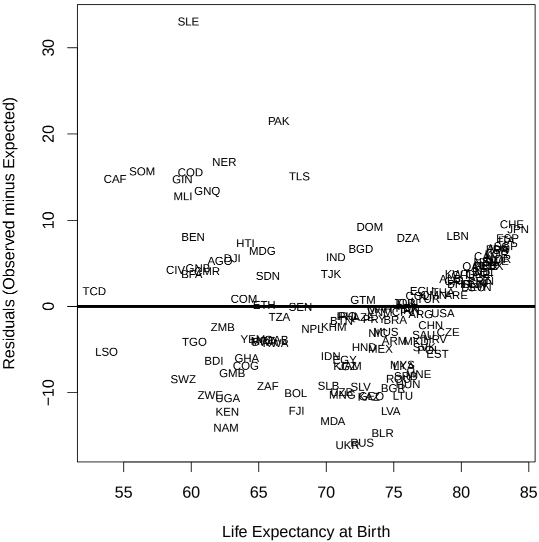
- The distribution of Y is *conditional on all variables in* $\mathbf{X}$, and
- The conditional distribution of Y is conditional on the *joint distribution* of the elements of $\mathbf{X}$.

$\rightarrow$ Regression is hard...

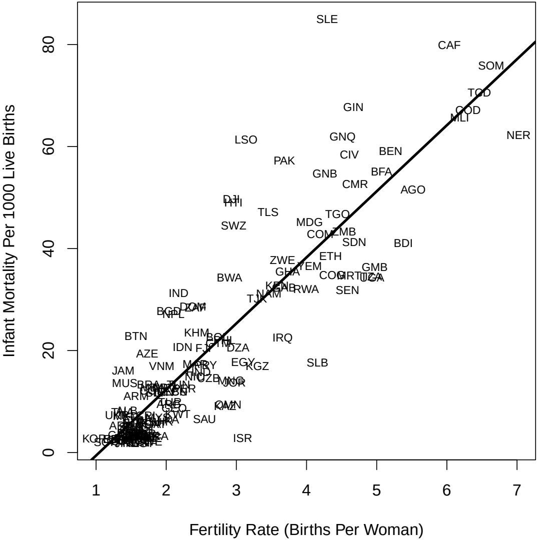
Example: Infant Mortality and Life Expectancy (2018)



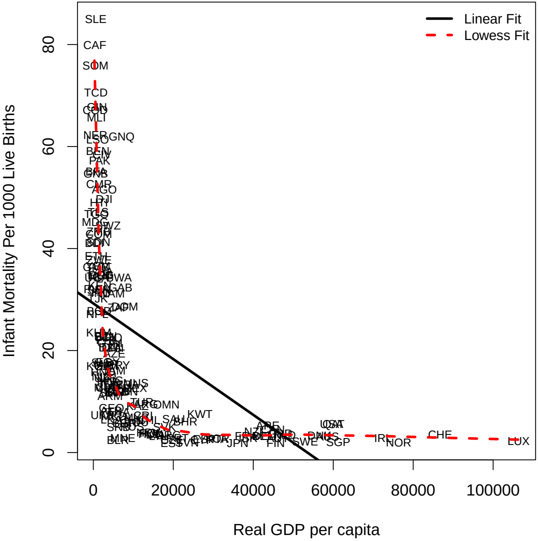
Infant Mortality and Life Expectancy: “Residuals”



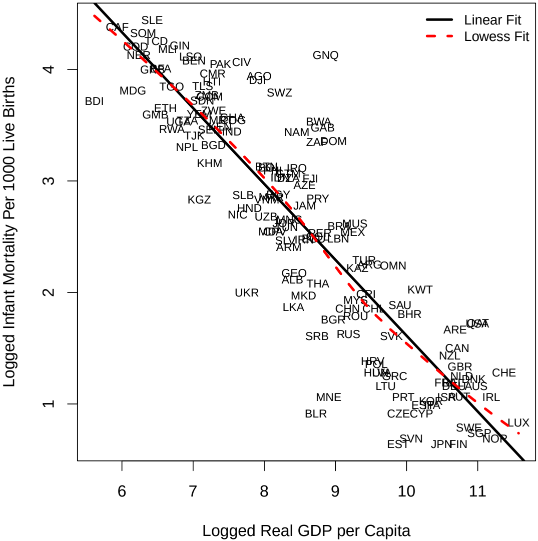
Infant Mortality and Fertility



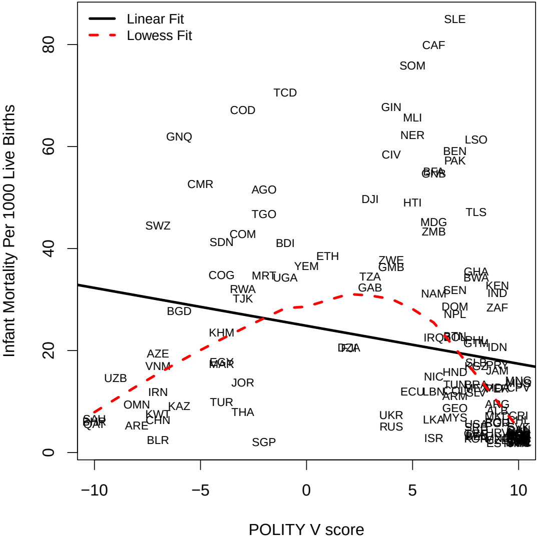
Infant Mortality and Wealth



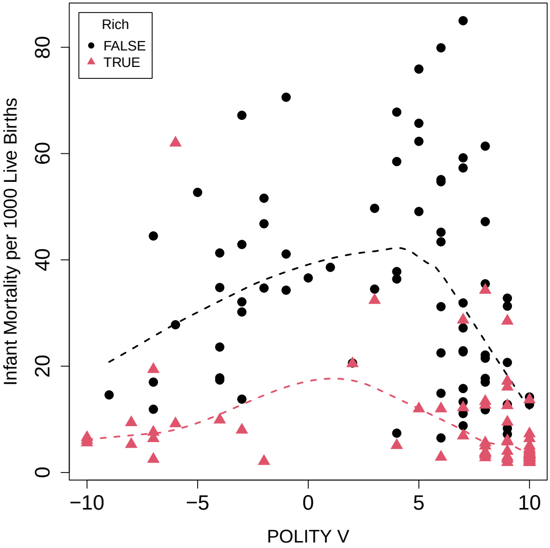
(Logged) Infant Mortality and (Logged) Wealth



Infant Mortality and Democracy



Infant Mortality, (Dichotomized) Wealth, and Democracy



Consider random variable Y:

$$Y_i = \mu + u_i \tag{1}$$

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

so:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2}$$

Goals:

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

For a bivariate regression:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\text{Covariance of } X \text{ and } Y}{\text{Variance of } X} \tag{3}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{4}$$

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

Assume (for now):

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

meaning:

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

so:

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

Var($\hat{\beta}_0$) and 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 :

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

Note that:

- $\text{Var}(\hat{\beta}_0)$ 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:

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

and

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

Means:

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

$\sigma^2 =$???

Solution: use

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

Gives:

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

and

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

So:

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

which 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:

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

$\hat{Y}_k$ is unbiased:

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

Variability:

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

Variability of Predictions

Prediction variation:

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

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:

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

$\rightarrow$ (e.g.) confidence intervals:

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

Variation in Y

We can decompose variation in Y:

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

TSS	$=$	**MSS**	$+$	**RSS**
(“Total”)		(“Estimated,” or “Model”)		(“Residual”)

Model Fit: R^2

"R-squared":

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

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,

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

"Adjusted" R^2:

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

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

$R^2_{adj.}$:

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

R^2 Alternatives

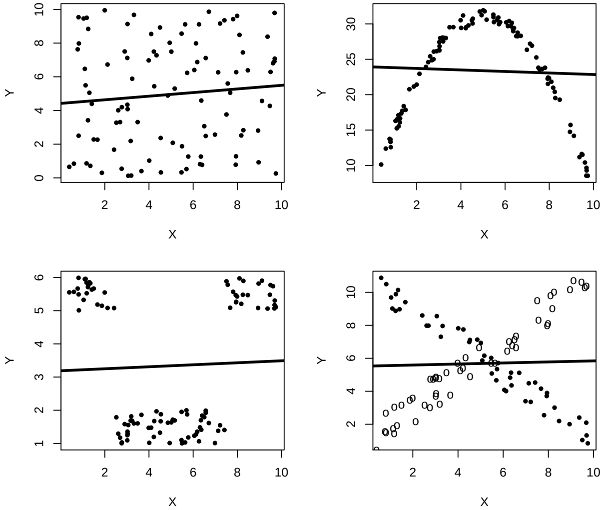
Alternative single measures of model fit include:

- The Standard Error of the Estimate:

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

- F-tests
- ROC / AUC
- Graphical methods

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



Now consider:

$$\underset{N\times 1}{\mathbf{Y}} = \underset{N\times K}{\mathbf{X}}\underset{K\times 1}{\boldsymbol{\beta}} + \underset{N\times 1}{\mathbf{u}}$$

equivalently:

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

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

Residuals:

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

The inner product of **u**:

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

Sum of squared residuals:

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

Now get:

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

Solve:

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

Variance-covariance matrix:

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

Rewrite:

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

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

Taking expectations:

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

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

Empirical estimate:

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

Yields:

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

OLS Assumptions

1. Zero Expectation Disturbances

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

2. Homoscedasticity / No Error Correlation

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

OLS Assumptions (continued)

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 of 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"**)

Multivariate Regression, Conceptually

Begin with:

- An *outcome* Y
- A *predictor* X
- Another variable Z

We are mainly interested in $\text{Cov}(Y, X|Z)$...

- Possibly *but not necessarily* the causal relationship (cf. Berk 2010)
- Key question is *model specification*...

The Easiest Case

Things are easiest if:

$$\begin{tikzcd} & Z & \\ X \arrow[rr] & & Y \end{tikzcd} \qquad \text{or} \qquad \begin{tikzcd} & Z & \\ X \arrow[ur] \arrow[rr] & & Y \end{tikzcd}$$

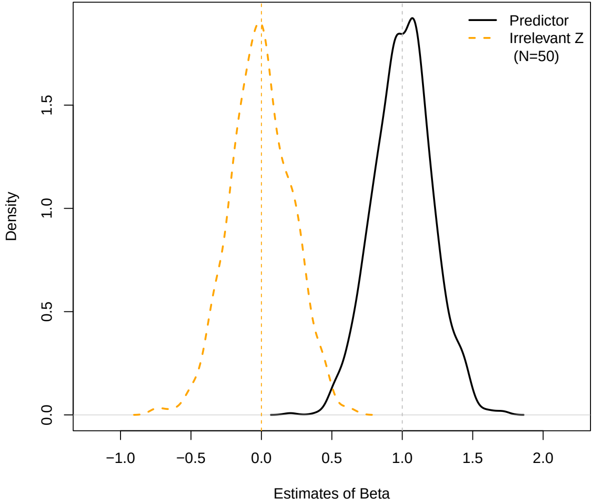
Implies that:

- Z is unimportant to understanding $\text{Cov}(X, Y)$
- $\rightarrow Z$ is *ignorable*

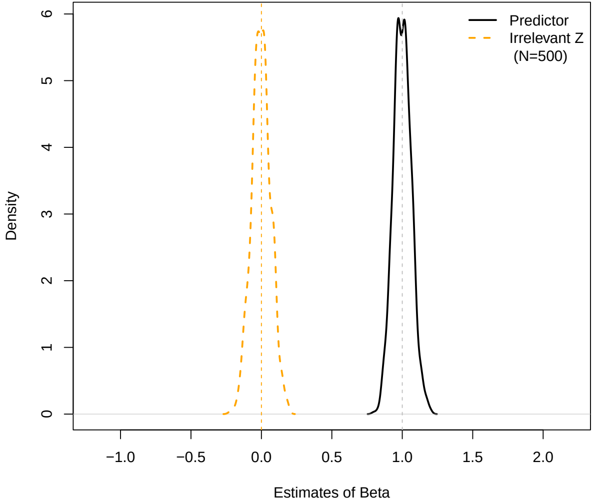
Simulations For Everyone!

```
> N<-50
> set.seed(7222009)
> X<-rnorm(N)         # Predictor
> Z<-(X+rnorm(N))/1.5 # Z "caused by" X
> Y<-X+rnorm(N)       # Outcome Y (unrelated to Z)

> print(summary(lm(Y~X)))

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    0.028      0.144    0.19       0.85
X              0.978      0.162    6.05 0.00000021 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 48 degrees of freedom
Multiple R-squared:  0.432, Adjusted R-squared:  0.421
F-statistic: 36.6 on 1 and 48 DF,  p-value: 0.000000212


> print(summary(lm(Y~X+Z)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0335     0.1460    0.23     0.82
X             0.9322     0.2161    4.31 0.000082 ***
Z             0.0659     0.2019    0.33     0.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 47 degrees of freedom
Multiple R-squared:  0.434, Adjusted R-squared:  0.41
F-statistic:   18 on 2 and 47 DF,  p-value: 0.00000157
```

Do That 999 More Times

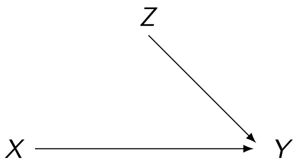


Same, But With $N = 500$



Slightly More Challenging

Suppose instead that:



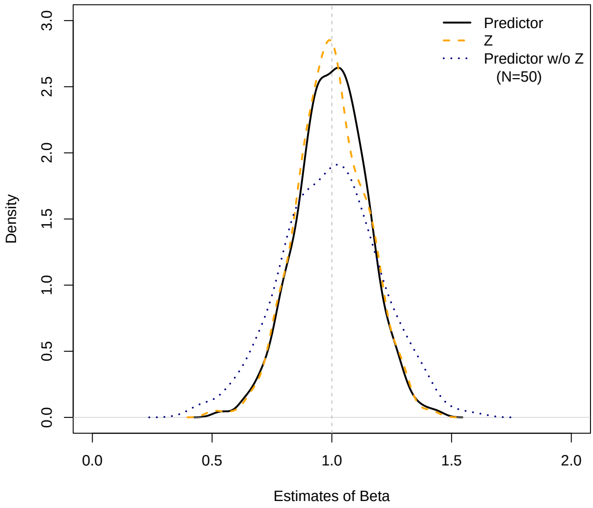
<u>This means that</u>:

- Z is important to / influential on understanding Y, *but*
- Z is unrelated to X...

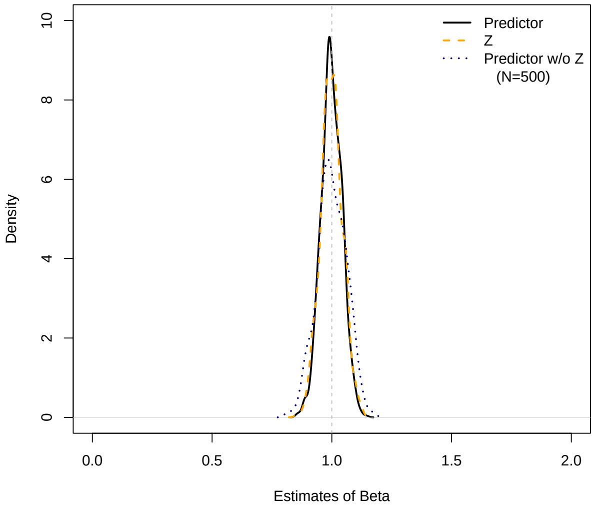
```
> N<-50
> set.seed(7222009)
> X<-rnorm(N)          # Predictor
> Z<-rnorm(N)          # Z orthogonal to X
> Y<-X+Z+rnorm(N)      # Outcome Y
```

```
> print(summary(lm(Y~X)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0956     0.2167   -0.44  0.66119
X             1.0301     0.2439    4.22  0.00011 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.5 on 48 degrees of freedom
Multiple R-squared:  0.271, Adjusted R-squared:  0.256
F-statistic: 17.8 on 1 and 48 DF,  p-value: 0.000107
```

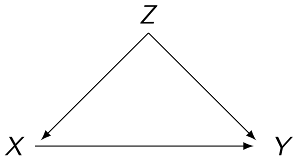
```
> print(summary(lm(Y~X+Z)))

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)   0.0335     0.1460    0.23          0.82
X             0.9761     0.1634    5.97 0.00000029670 ***
Z             1.0439     0.1346    7.75 0.00000000059 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1 on 47 degrees of freedom
Multiple R-squared:  0.68, Adjusted R-squared:  0.666
F-statistic:   50 on 2 and 47 DF,  p-value: 0.00000000000233
```


Predictor
Z
Predictor w/o Z
(N=50)
Density
Estimates of Beta
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Same, But With $N = 500$



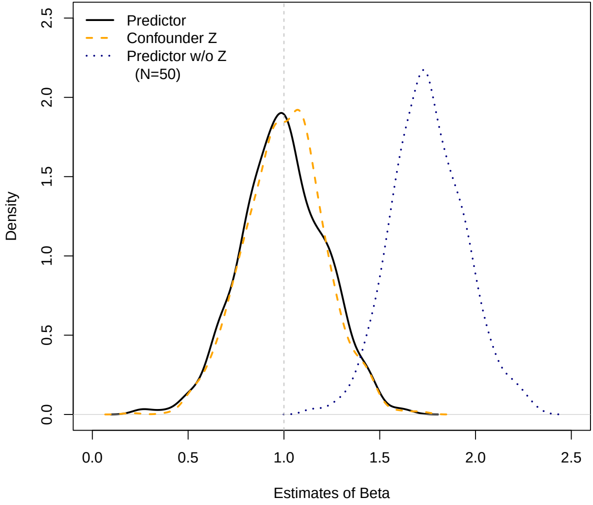
The classic example is:



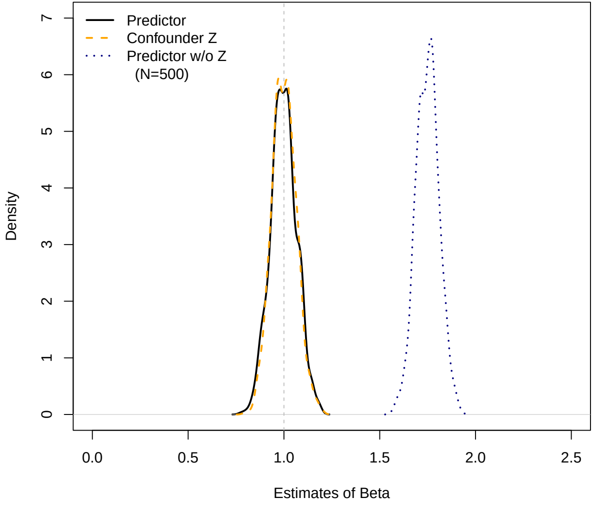
This means that:

- Z is important to / influential on both X and Y
- The marginal association $\text{Cov}(X, Y|Z)$ (obviously) depends on Z...
- More specifically, $\text{Cov}(X, Y|Z) \neq \text{Cov}(X, Y)$

So Much Confounding

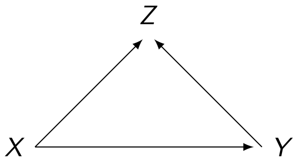


Same, But With $N = 500$



“Collider Bias”

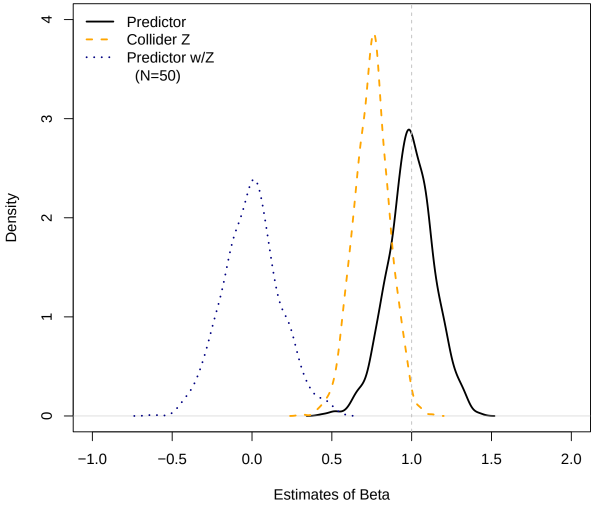
Z is a “collider”:



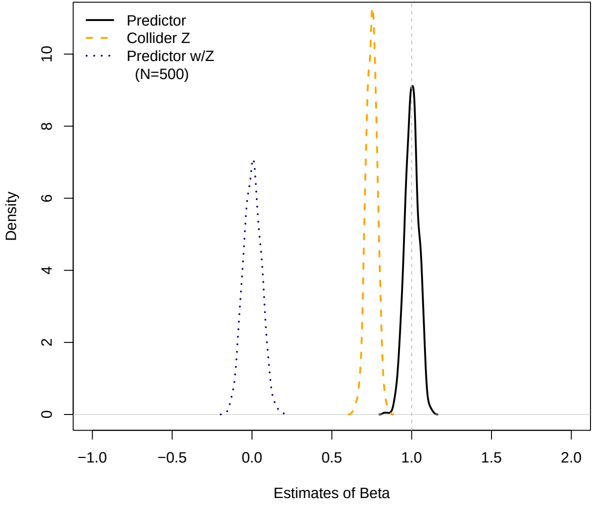
<u>This means that</u>:

- Z is influenced *by* both X and Y
- Once again, $\text{Cov}(X, Y|Z) \neq \text{Cov}(X, Y)$
- Sometimes referred to as Berkson’s paradox

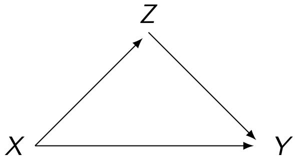
So Much Colliding



Same, But With $N = 500$



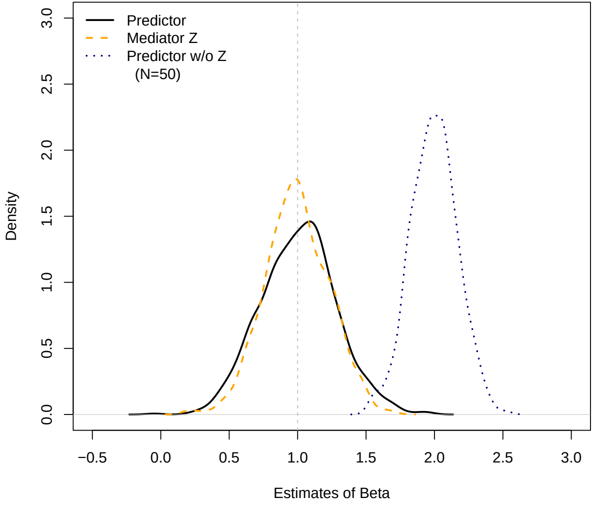
Z is a "mediator":



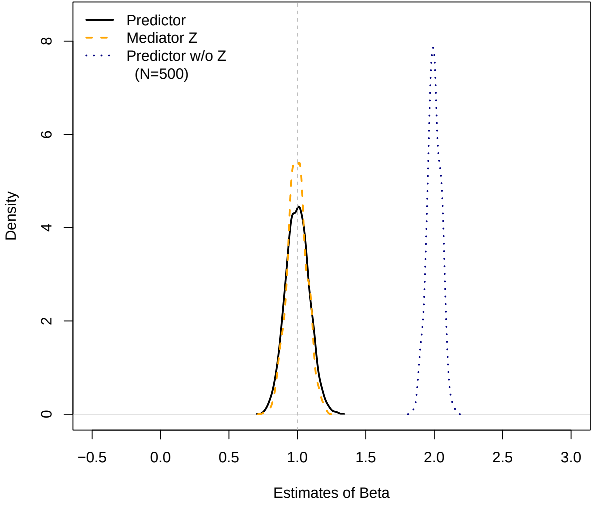
This means that:

- Z is influenced by X; Y by X and Z
- Once again, $\text{Cov}(X, Y|Z) \neq \text{Cov}(X, Y)$
- Think of $\text{Cov}(X, Y) =$ "total effect" and $\text{Cov}(X, Y|Z) =$ "direct effect"

Mediation Illustrated



Same, But With $N = 500$



Lessons...

Some takeaways... *In a linear model*:

- Variables that are irrelevant (to Y), are irrelevant...
- Variables that are relevant to Y but unrelated to X need not be modeled
- Confounders require that we condition on them, or else there's bias
- Colliders require that we *do not* condition on them, or else there's bias
- Mediators may or may not be good to condition on...

Mostly: **Model specification is hard.**

Example Data: The WDI

Data are drawn from the World Development Indicators:

- Cross-national country-level time series data
- $N = 143$ countries (due to missing data); here we'll focus on one recent year (2018)
- Available as file `WDI-2018-Day-One-24.csv` in the `Data` folder at the course Github repo.

Variables (among others):

- `ISO3` - The country's International Standards Organization (ISO) three-letter identification code (e.g., `CHE` for Switzerland).
- `Country` - The name of the country
- `Fertility` - Fertility rate (mean births per woman).
- `Child Mortality` - Average deaths of children under 5 per 1000 live births.
- `Infant Mortality` - Average deaths of infants per 1000 live births.
- `LifeExpectancy` - Life Expectancy at birth (years).
- `DPTPercent` - Percent of children aged 12-24 months w/DPT immunization.
- `GDPPerCapita` - GDP per capita (constant 2010 $US).
- `FDIIn` - Inward Foreign Direct Investment (FDI) (percent of GDP).
- `NaturalResourceRents` - Total natural resource rents (percent of GDP).
- `UrbanPopulation` - Urban Population (percent of total).
- `GovtExpenditures` - Government Expenditures (percent of GDP).
- `PaidParentalLeave` - Paid Parental Leave ($0 =$ No, $1 =$ Yes).
- `DemocScore` - POLITY V democracy score (0-10).
- `AutocScore` - POLITY V autocracy score (0-10).
- `POLITY` - POLITY V total score (`DemocScore` - `AutocScore`).

Summary Statistics

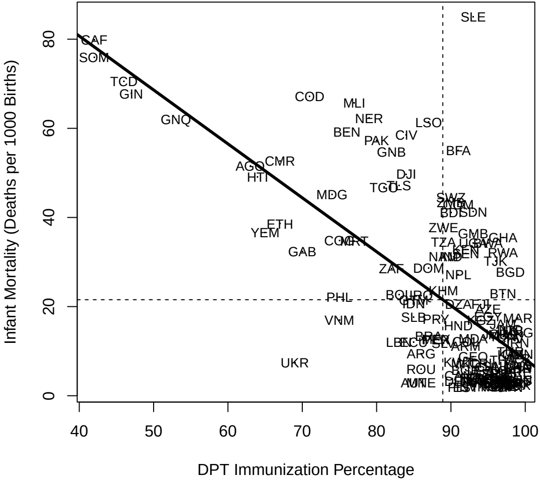
```
> describe(IR2018)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ISO3*	1	143	72.00	41.42	72.00	72.00	53.37	1.00	143.00	142.00	0.00	-1.23	3.46
Country*	2	143	72.00	41.42	72.00	72.00	53.37	1.00	143.00	142.00	0.00	-1.23	3.46
Fertility	3	143	2.71	1.39	2.15	2.52	1.02	0.98	7.02	6.05	1.09	0.31	0.12
ChildMortality	4	143	28.96	30.52	15.40	23.57	17.20	2.40	123.20	120.80	1.35	0.99	2.55
InfantMortality	5	143	21.55	20.47	13.40	18.45	14.97	1.90	85.00	83.10	1.10	0.27	1.71
LifeExpectancy	6	143	72.37	7.82	73.85	72.82	8.91	52.83	84.21	31.39	-0.45	-0.82	0.65
DPTPercent	7	143	88.90	11.73	93.00	91.19	5.93	42.00	99.00	57.00	-2.06	4.53	0.98
GDPPerCapita	8	143	14093.79	19826.50	5233.28	9823.82	6185.18	274.13	106376.78	106102.65	2.08	4.29	1657.97
FDIIn	9	143	1.37	12.07	2.38	2.45	2.11	-117.37	29.21	146.58	-7.05	64.58	1.01
NaturalResourceRents	10	143	6.73	10.55	2.26	4.34	2.98	0.00	62.77	62.77	2.55	7.40	0.88
UrbanPopulation	11	143	60.40	21.62	61.58	61.13	26.32	13.03	100.00	86.97	-0.24	-0.88	1.81
GovtExpenditures	12	143	15.72	6.45	15.48	15.29	5.72	3.60	56.31	52.71	2.00	10.03	0.54
PaidParentalLeave	13	143	0.29	0.46	0.00	0.24	0.00	0.00	1.00	1.00	0.90	-1.20	0.04
DemocScore	14	143	6.01	3.75	7.00	6.26	4.45	0.00	10.00	10.00	-0.58	-1.23	0.31
AutocScore	15	143	1.55	2.55	0.00	0.97	0.00	0.00	10.00	10.00	1.74	2.05	0.21
POLITY	16	143	4.45	6.01	7.00	5.27	4.45	-10.00	10.00	20.00	-0.97	-0.46	0.50
Rich*	17	143	1.50	0.50	1.00	1.50	0.00	1.00	2.00	1.00	0.01	-2.01	0.04
IMDPTres	18	143	0.00	14.75	-3.75	-1.39	10.15	-38.23	68.41	106.64	1.16	2.43	1.23
IMDPThat	19	143	21.55	14.19	16.59	18.78	7.17	9.33	78.29	68.96	2.06	4.53	1.19

```
> IMDPT<-lm(InfantMortality~DPTPercent,data=IR2018)
> summary.lm(IMDPT)

Call:
lm(formula = InfantMortality ~ DPTPercent, data = IR2018)

Residuals:
   Min     1Q Median     3Q    Max
-38.23  -9.64  -3.75   5.18  68.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  129.104      9.491    13.6   <2e-16 ***
DPTPercent    -1.210      0.106   -11.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.8 on 141 degrees of freedom
Multiple R-squared:  0.481, Adjusted R-squared:  0.477
F-statistic:  131 on 1 and 141 DF,  p-value: <2e-16
```

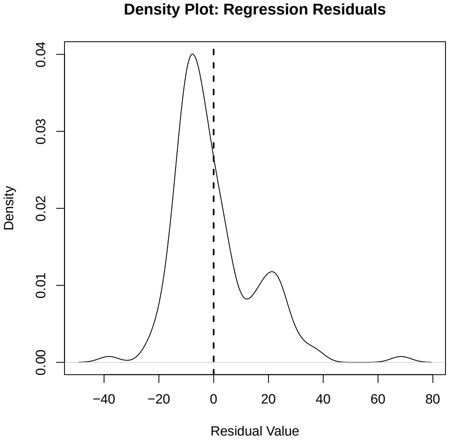
```
> anova(IMDPT)

Analysis of Variance Table

Response: InfantMortality
            Df Sum Sq Mean Sq F value Pr(>F)
DPTPercent   1  28607   28607     131 <2e-16 ***
Residuals  141  30874     219
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regression of Infant Mortality on DPT Immunization Rates



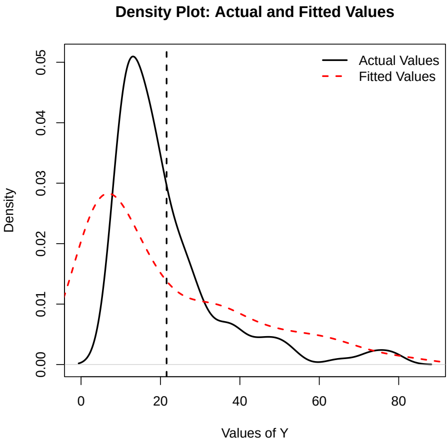
Fitted Values, Residuals, etc.

```
> # Residuals (u):
> IR2018$IMDPTres <- with(IR2018, residuals(IMDPT))
> describe(IR2018$IMDPTres)
   vars   n mean   sd median trimmed  mad   min  max range skew kurtosis   se
X1    1 143    0 14.8  -3.75   -1.39 10.2 -38.2 68.4   107 1.16     2.43 1.23
```

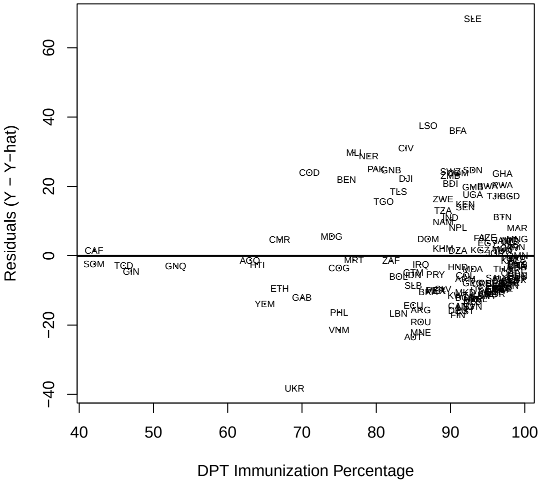


Fitted Values

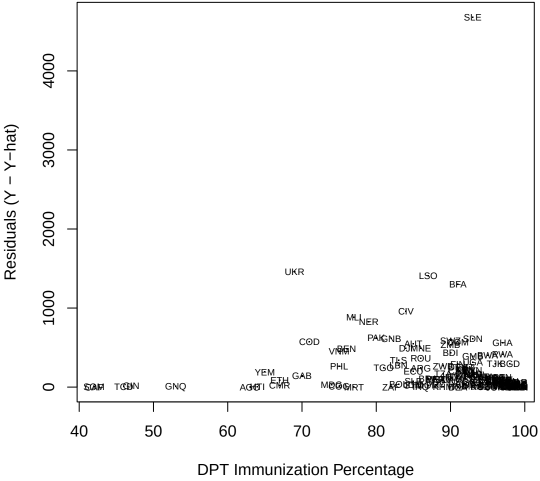
```
> # Fitted Values:
> IR2018$IMDPThat<-fitted.values(IMDPT)
> describe(IR2018$IMDPThat)
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 143 21.6 14.2   16.6    18.8 7.17 9.33 78.3    69 2.06     4.53 1.19
```



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



Squared Residuals vs. DPT Percentage



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

```
> vcov(IMDPT)
```

```
            (Intercept) DPTPercent
(Intercept)      90.078    -0.9960
DPTPercent       -0.996     0.0112
```

95 percent c.i.s:

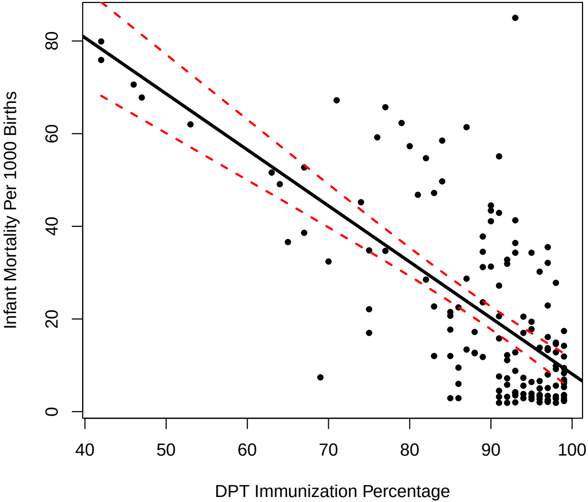
```
> confint(IMDPT)
```

```
             2.5 % 97.5 %
(Intercept) 110.34    148
DPTPercent   -1.42     -1
```

```
> SEs<-predict(IMDPT,interval="confidence")
> SEs
      fit   lwr  upr
3   52.88 46.94 58.8
4    9.33  6.10 12.6
6    9.33  6.10 12.6
.
.
<rows omitted>
.
.
212 50.46 44.90 56.0
213 29.90 27.06 32.7
214 20.22 17.76 22.7
215 21.43 18.98 23.9
```

A Plot, With CIs

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



Multivariate Model: Child Mortality

```
> model<-lm(InfantMortality~DPTPercent+GDPPerCapita+FDIIn+
+           NaturalResourceRents+UrbanPopulation+GovtExpenditures+
+           POLITY+PaidParentalLeave,data=IR2018)
> summary(model)

Call:
lm(formula = InfantMortality ~ DPTPercent + GDPPerCapita + FDIIn +
    NaturalResourceRents + UrbanPopulation + GovtExpenditures +
    POLITY + PaidParentalLeave, data = IR2018)

Residuals:
   Min     1Q Median     3Q    Max
-30.77  -5.97  -1.01   5.25  57.79

Coefficients:
                        Estimate  Std. Error t value Pr(>|t|)
(Intercept)          120.1847454   8.2458035   14.58  < 2e-16 ***
DPTPercent            -0.8508569   0.0936331   -9.09  1.2e-15 ***
GDPPerCapita          -0.0000915   0.0000717   -1.28   0.2044
FDIIn                 -0.1000156   0.0897276   -1.11   0.2670
NaturalResourceRents   0.1521474   0.1142324    1.33   0.1852
UrbanPopulation       -0.3295500   0.0571374   -5.77  5.3e-08 ***
GovtExpenditures       0.0036648   0.1727382    0.02   0.9831
POLITY                -0.0928692   0.1947269   -0.48   0.6342
PaidParentalLeave     -7.9426051   2.4968797   -3.18   0.0018 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.5 on 134 degrees of freedom
Multiple R-squared:  0.703, Adjusted R-squared:  0.685
F-statistic: 39.6 on 8 and 134 DF,  p-value: <2e-16
```

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

```
> vcov(model)
                     (Intercept)   DPTPercent   GDPPerCapita       FDIIn NaturalResourceRents
(Intercept)            67.993276 -0.693611256  0.00012395662  0.00773960         -0.234304691
DPTPercent             -0.693611  0.008767158 -0.00000058798 -0.00019383          0.002875263
GDPPerCapita            0.000124 -0.000000588  0.00000000515  0.00000264          0.000000642
FDIIn                   0.007740 -0.000193834  0.00000264192  0.00805104         -0.000201076
NaturalResourceRents   -0.234305  0.002875263  0.00000064215 -0.00020108          0.013049034
UrbanPopulation        -0.084284 -0.000733489 -0.00000204452 -0.00069983         -0.000774746
GovtExpenditures       -0.014527 -0.003418874 -0.00000080316  0.00036455         -0.007817267
POLITY                 -0.276336  0.002795967 -0.00000030251  0.00079257          0.010641537
PaidParentalLeave       3.270567 -0.037486336 -0.00004999336 -0.01489057          0.014976307

                     UrbanPopulation GovtExpenditures       POLITY PaidParentalLeave
(Intercept)              -0.08428441     -0.014527322 -0.276336134           3.27057
DPTPercent               -0.00073349     -0.003418874  0.002795967          -0.03749
GDPPerCapita             -0.00000204     -0.000000803 -0.000000303          -0.00005
FDIIn                    -0.00069983      0.000364547  0.000792569          -0.01489
NaturalResourceRents     -0.00077475     -0.007817267  0.010641537           0.01498
UrbanPopulation           0.00326469     -0.000559546 -0.000337250          -0.00822
GovtExpenditures         -0.00055955      0.029838489 -0.010635268          -0.02024
POLITY                   -0.00033725     -0.010635268  0.037918563          -0.07426
PaidParentalLeave        -0.00822062     -0.020241192 -0.074256519           6.23441
```

Test $H_0 : \beta_{\texttt{FDIIn}} = \beta_{\texttt{POLITY}} = 0$:

```
> library(lmtest)
> modelsmall<-lm(InfantMortality~DPTPercent+GDPPerCapita+NaturalResourceRents+
+                        UrbanPopulation+GovtExpenditures+PaidParentalLeave,
+                        data=IR2018)

> waldtest(model,modelsmall)  # from -lmtest- package

Wald test

Model 1: InfantMortality ~ DPTPercent + GDPPerCapita + FDIIn + NaturalResourceRents +
    UrbanPopulation + GovtExpenditures + POLITY + PaidParentalLeave
Model 2: InfantMortality ~ DPTPercent + GDPPerCapita + NaturalResourceRents +
    UrbanPopulation + GovtExpenditures + PaidParentalLeave
  Res.Df Df    F Pr(>F)
1    134
2    136 -2 0.71   0.49
```

Test $H_0 : \beta_{\texttt{NaturalResourceRents}} = 1.0$:

```
> library(car)
> linearHypothesis(model,"NaturalResourceRents=1") # from -car-
Linear hypothesis test

Hypothesis:
NaturalResourceRents = 1

Model 1: restricted model
Model 2: InfantMortality ~ DPTPercent + GDPPerCapita + FDIIn + NaturalResourceRents +
    UrbanPopulation + GovtExpenditures + POLITY + PaidParentalLeave

  Res.Df   RSS Df Sum of Sq    F  Pr(>F)
1    135 24946
2    134 17678  1      7268 55.1 1.2e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test $H_0 : \beta_{\texttt{NaturalResourceRents}} = -\beta_{\texttt{UrbanPopulation}}$:

```
> linearHypothesis(model,"NaturalResourceRents = -UrbanPopulation")

Linear hypothesis test

Hypothesis:
NaturalResourceRents  + UrbanPopulation = 0

Model 1: restricted model
Model 2: InfantMortality ~ DPTPercent + GDPPerCapita + FDIIn + NaturalResourceRents +
    UrbanPopulation + GovtExpenditures + POLITY + PaidParentalLeave

  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1    135 17960
2    134 17678  1       281 2.13   0.15
```

Reporting: Making Tables

R

- LaTeX: `texreg`, `xtable`, `stargazer`, and `modelsummary` packages
- MS Word: generally cut-and-paste (see, e.g., here: https://sejdemyr.github.io/r-tutorials/basics/tables-in-r/); also KableExtra
- A pretty good summary of many others is here: https://rfortherestofus.com/2019/11/how-to-make-beautiful-tables-in-r/.

Stata

- `estout` and `esttab` commands are standard
- Others: `outreg2`, `tabout`, `orth_out`, etc. (a summary is here: https://lukestein.github.io/stata-latex-workflows/)
- MS Word: putdocx

Reporting (continued)

The output:

```
> summary(model)

Call:
lm(formula = InfantMortality ~ DPTPercent + GDPPerCapita + FDIIn +
    NaturalResourceRents + UrbanPopulation + GovtExpenditures +
    POLITY + PaidParentalLeave, data = IR2018)

Residuals:
   Min     1Q Median     3Q    Max
-30.77  -5.97  -1.01   5.25  57.79

Coefficients:
                        Estimate  Std. Error t value Pr(>|t|)
(Intercept)          120.1847454   8.2458035   14.58  < 2e-16 ***
DPTPercent            -0.8508569   0.0936331   -9.09  1.2e-15 ***
GDPPerCapita          -0.0000915   0.0000717   -1.28   0.2044
FDIIn                 -0.1000156   0.0897276   -1.11   0.2670
NaturalResourceRents   0.1521474   0.1142324    1.33   0.1852
UrbanPopulation       -0.3295500   0.0571374   -5.77  5.3e-08 ***
GovtExpenditures       0.0036648   0.1727382    0.02   0.9831
POLITY                -0.0928692   0.1947269   -0.48   0.6342
PaidParentalLeave     -7.9426051   2.4968797   -3.18   0.0018 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.5 on 134 degrees of freedom
Multiple R-squared:  0.703, Adjusted R-squared:  0.685
F-statistic: 39.6 on 8 and 134 DF,  p-value: <2e-16
```

The table:

Table 1: OLS Regression Model of Child Mortality Rates, 2018

	Model I
(Constant)	120.00^{***}
	(8.25)
DPT Immunization Rate	-0.85^{***}
	(0.09)
GDP Per Capita	-0.0001
	(0.0001)
Inward FDI	-0.10
	(0.09)
Natural Resource Rents (Pct. GDP)	0.15
	(0.11)
Urban Population (Pct.)	-0.33^{***}
	(0.06)
Government Expenditures (Pct. GDP)	0.004
	(0.17)
POLITY V (Democracy) Score	-0.09
	(0.19)
Paid Parental Leave	-7.94^{***}
	(2.50)
Observations	143
R^2	0.70
Adjusted R^2	0.69
Residual Std. Error	11.50 (df = 134)
F Statistic	39.60^{***} (df = 8; 134)

Note: Cell entries are coefficient estimates; numbers in parentheses are estimated standard errors.
$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. See text for details.

Multiple Models (`stargazer` defaults)

OLS Regression Models of Infant Mortality Rates, 2018

	Univariate	Full	Reduced
(Constant)	129.00***	120.00***	120.00***
	(9.49)	(8.25)	(8.10)
DPT Immunization Rate	−1.21***	−0.85***	−0.85***
	(0.11)	(0.09)	(0.09)
GDP Per Capita		−0.0001	−0.0001
		(0.0001)	(0.0001)
Inward FDI		−0.10	
		(0.09)	
Natural Resource Rents (Pct. GDP)		0.15	0.17*
		(0.11)	(0.10)
Urban Population (Pct.)		−0.33***	−0.34***
		(0.06)	(0.06)
Government Expenditures (Pct. GDP)		0.004	−0.02
		(0.17)	(0.16)
POLITY V (Democracy) Score		−0.09	
		(0.19)	
Paid Parental Leave		−7.94***	−8.29***
		(2.50)	(2.46)
Observations	143	143	143
R^2	0.48	0.70	0.70
Adjusted R^2	0.48	0.69	0.69
Residual Std. Error	14.80 (df = 141)	11.50 (df = 134)	11.50 (df = 136)
F Statistic	131.00*** (df = 1; 141)	39.60*** (df = 8; 134)	52.80*** (df = 6; 136)
Note:			*p<0.1; **p<0.05; ***p<0.01

Some Guidelines ("Rules"?)

Tables:

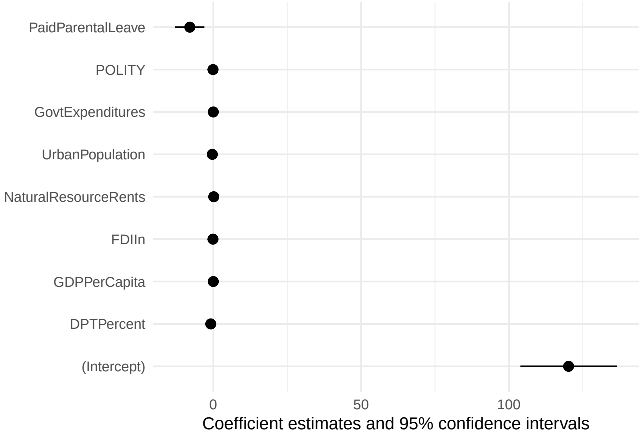
- *Use column headings descriptively.*
- *Use multiple rows / columns rather than multiple tables.*
- *Learn about significant digits, and don't report more than 4-5 of them (at the most).*
- *Use a figure to replace a table when you can.*
- *Be aware of norms about *s.*

Figures:

- *Report the scale of axes, and label them.*
- *Use as much "space" as you need, but no more.*
- *Use color sparingly.*

Plotting Regression Estimates

Ladderplot of OLS Results (using `modelplot` defaults)



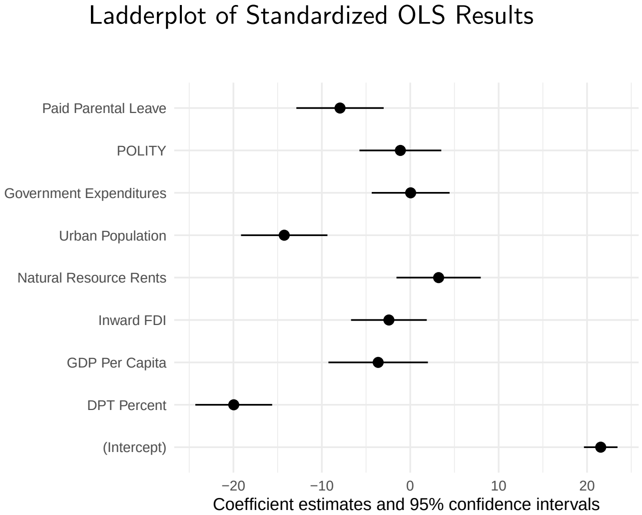
Rescaling Covariates

A la Gelman (2008):

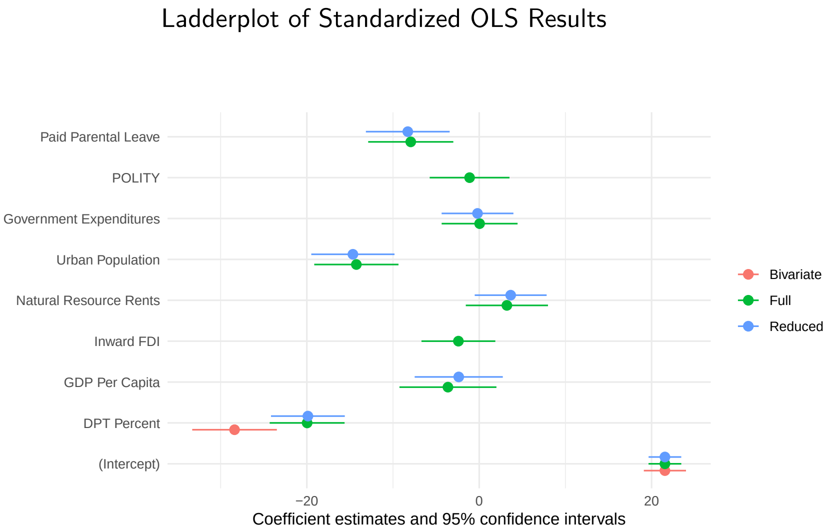
- Continuous = divide by two standard deviations
- Binary = mean 0, difference of 1 between the two categories

```
> modelS<-standardize(model)
> summary(modelS)

Call:
lm(formula = InfantMortality ~ z.DPTPercent + z.GDPPerCapita +
    z.FDIIn + z.NaturalResourceRents + z.UrbanPopulation + z.GovtExpenditures +
    z.POLITY + c.PaidParentalLeave, data = IR2018)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             21.5469     0.9605   22.43  < 2e-16 ***
z.DPTPercent           -19.9644     2.1970   -9.09  1.2e-15 ***
z.GDPPerCapita          -3.6274     2.8445   -1.28   0.2044
z.FDIIn                 -2.4140     2.1657   -1.11   0.2670
z.NaturalResourceRents   3.2110     2.4108    1.33   0.1852
z.UrbanPopulation      -14.2529     2.4712   -5.77  5.3e-08 ***
z.GovtExpenditures       0.0472     2.2269    0.02   0.9831
z.POLITY                -1.1163     2.3407   -0.48   0.6342
c.PaidParentalLeave     -7.9426     2.4969   -3.18   0.0018 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.5 on 134 degrees of freedom
Multiple R-squared:  0.703, Adjusted R-squared:  0.685
F-statistic: 39.6 on 8 and 134 DF,  p-value: <2e-16
```

A Better Regression Plot



An Even Better Regression Plot



Some Meta-Rules

- *Be aware of the norms in your discipline / field, and follow them.*
- *If it has uncertainty, you should show it.*
- *Ask for advice.*
- *When in doubt, more information is (probably) better.*

Supplement: R Things

Inside the Source Window

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-June-2024-R-Intro.R` in the github repo)

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 for Publishing "Useful R Resources" guide (on GitHub).