GSERM 2022
Regression for Publishing

June 13, 2022

"Regression for Publishing"

- "Regression" course
- Texts: Mostly posted readings; also Weisberg (2014) and/or Faraway (2002)
- Course materials at the github repo: https://github.com/PrisonRodeo/GSERM-RFP-2022 and on CANVAS
- Software:
  - Support: R $\geq$ Stata $>$ others...
  - GSERM virtual machines at https://vdi.unisg.ch/gserm; more details are available at the *GSERM 2022 Virtual Software Guide*
  - Also: There's an "Introduction to R and LaTeX" on the Github repo as well.
- Assessment: One homework assignment plus a final examination.

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.

"Regression," conceptually:

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

Two important things:

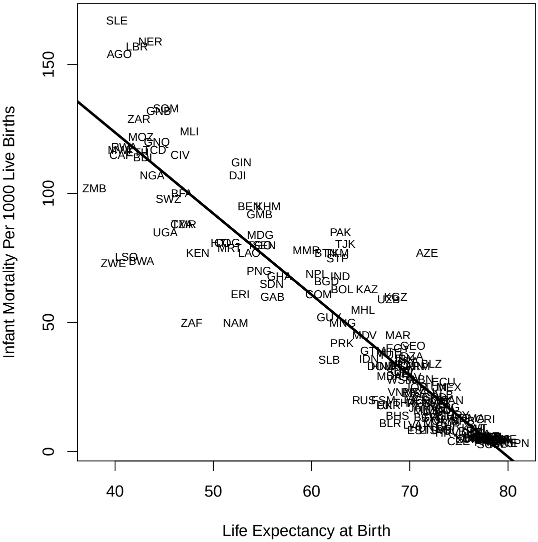
- The distribution of Y is *conditional on all variables in* $\mathbf{X}$, and
- The conditional distribution of Y is conditional on the *joint distribution* of the elements of $\mathbf{X}$.

$\rightarrow$ Regression is <u>hard</u>...

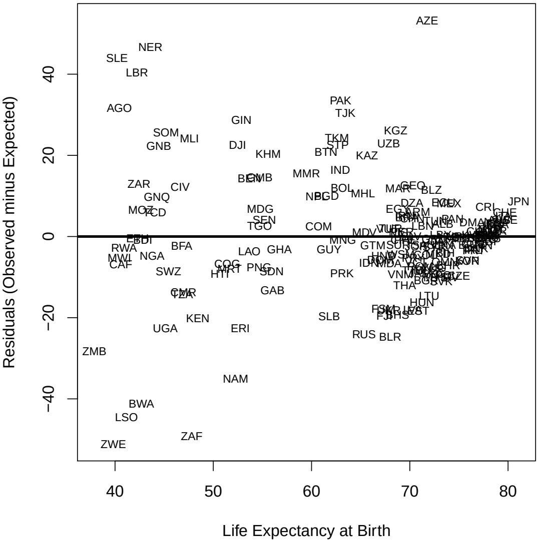
Why Regression?

	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

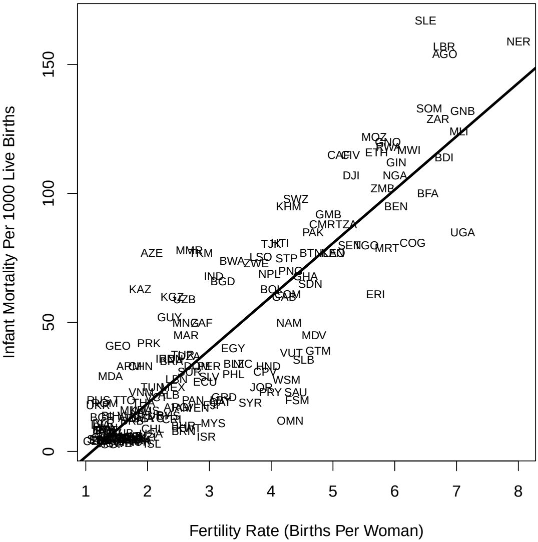
Example: Infant Mortality and Life Expectancy



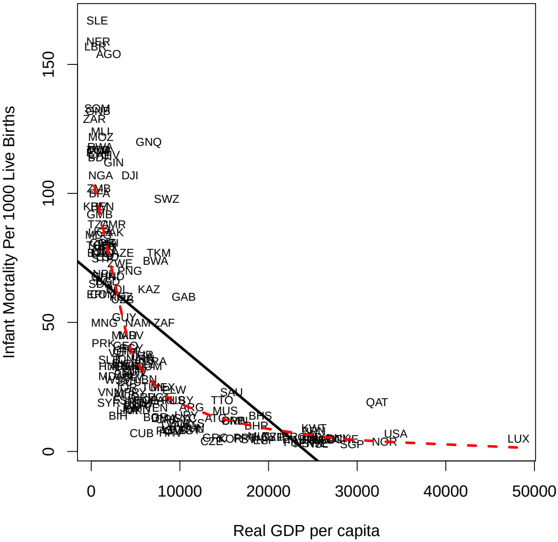
Infant Mortality and Life Expectancy: “Residuals”



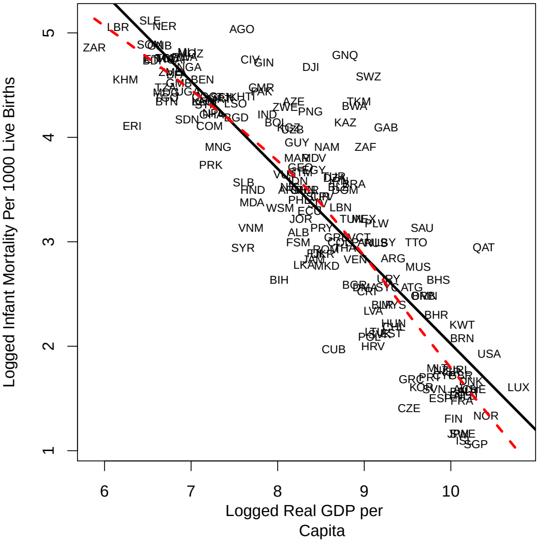
Infant Mortality and Fertility



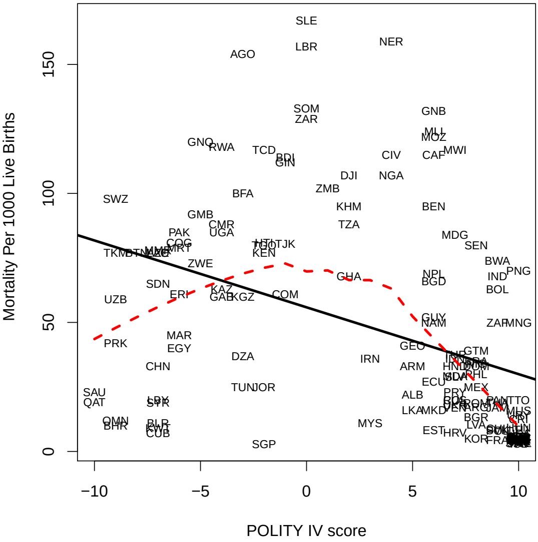
Infant Mortality and Wealth



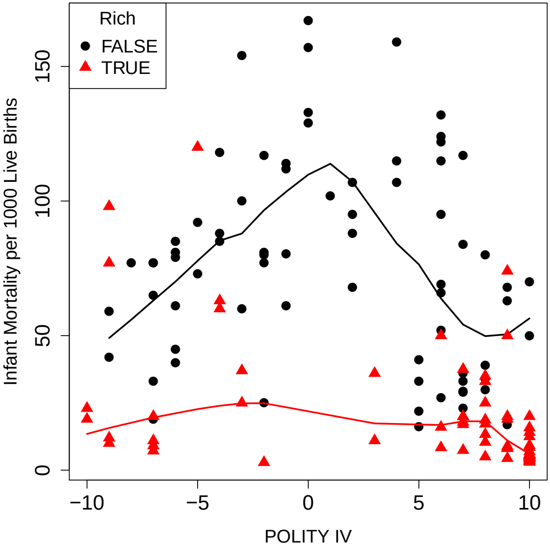
(Logged) Infant Mortality and (Logged) Wealth



Infant Mortality and Democracy



Infant Mortality, (Dichotomized) Wealth, and Democracy



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

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

so:

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

<u>Goals</u>:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \quad (3)$$

$$= \frac{\text{Covariance of } X \text{ and } Y}{\text{Variance of } X}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (4)$$

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

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

meaning:

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

so:

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

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 :

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

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

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

Inference (continued)

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

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:

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

Variability of Predictions

Prediction variation:

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

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

- Decreases in N
- Decreases in $\mathrm{Var}(X)$
- 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

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

$$\underset{(\text{“Total”})}{\textbf{TSS}} = \underset{(\text{“Estimated,” or “Model”})}{\textbf{MSS}} + \underset{(\text{“Residual”})}{\textbf{RSS}}$$

"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

"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.

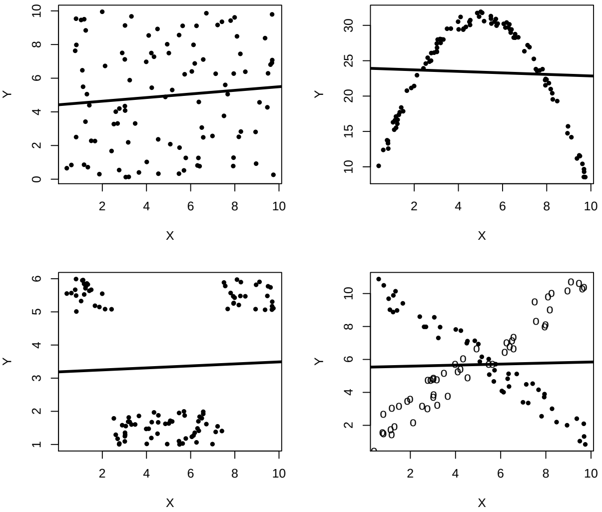
R^2 Alternatives

- 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

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

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

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

Residuals:

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

The inner product of **u**:

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

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

Now get:

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

Solve:

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

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

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

Rewrite:

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

Taking expectations:

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

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

Empirical estimate:

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

Yields:

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

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

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

OLS: Properties

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

- **Unbiased**

- **Fully Efficient**

(i.e., **"BLUE"**)

Example Data: Infant Mortality

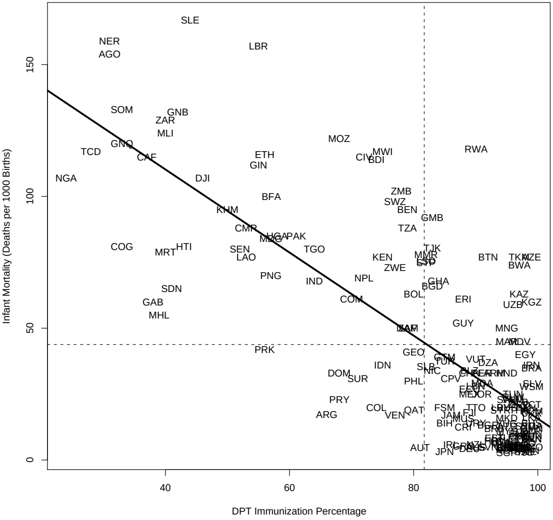
```
> # Summary statistics
>
> # install.packages("psych") <- Install psych package, if necessary
> library(psych)

> with(IR2000, 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(IR2000, 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~DPTpct,data=IR2000,na.action=na.exclude)
> summary.lm(IMDPT)

Call:
lm(formula = infantmortalityperK ~ DPTpct, 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 ***
DPTpct       -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
```

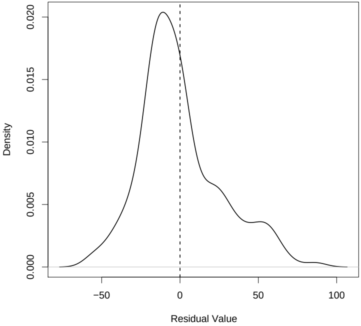
```
> anova(IMDPT)
Analysis of Variance Table

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.

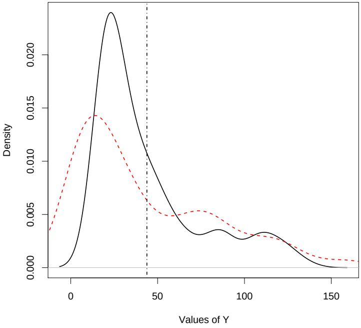
```
> # Residuals (u):
> IR2000$IMDPTres <- with(IR2000, residuals(IMDPT))
> describe(IR2000$IMDPTres)

  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
```



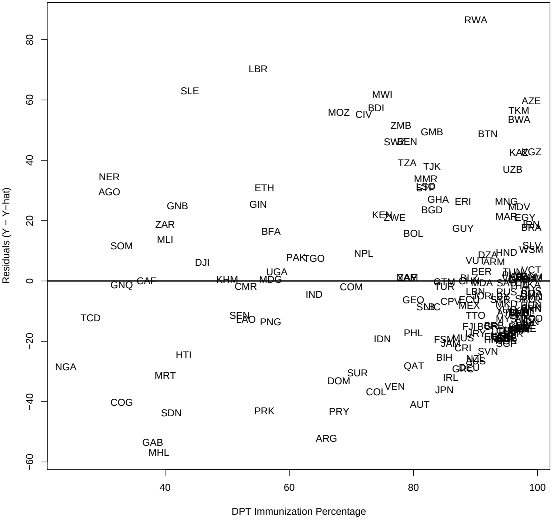
Fitted Values

```
> # Fitted Values:
> IR2000$IMDPThat<-fitted.values(IMDPT)
> describe(IR2000$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
```

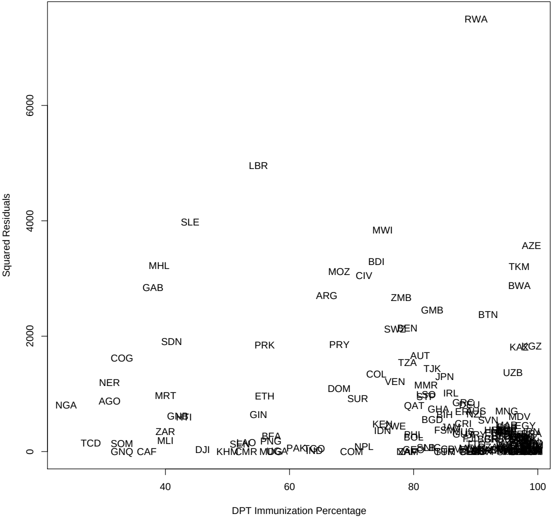
Density Plot: Actual (Y) and Fitted Values ($\hat{Y}$)



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



Squared Residuals vs. DPT Percentage



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

```
> vcov(IMDPT)
```

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

95 percent c.i.s:

```
> confint(IMDPT)
```

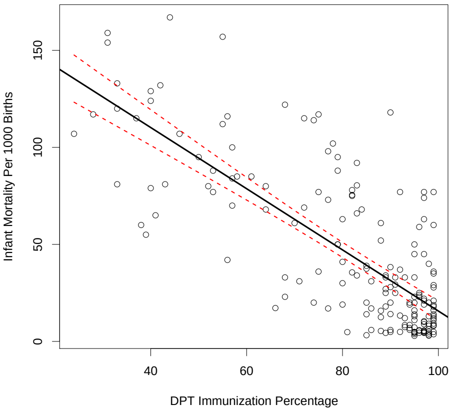
```
              2.5 %  97.5 %
(Intercept) 156.523 190.032
DPTpct       -1.775  -1.377
```

Predictions

```
> 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

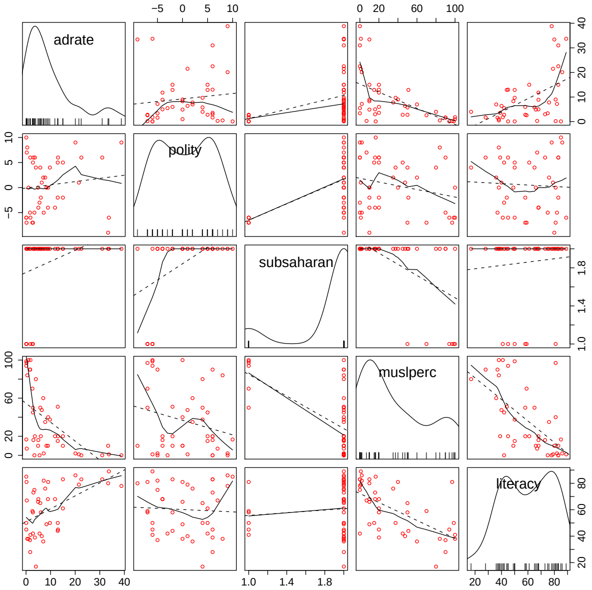
Scatterplot of Infant Mortality and DPT Immunizations, along with Least-Squares Line and 95% Prediction Confidence Intervals



Multivariate Example: Africa Data

```
> Data<-read_csv("https://github.com/PrisonRodeo/GSERM-RFP-2022/raw/main/Data/africa2001.csv")

> Data<-with(Data, data.frame(adrate,polity,
+                             subsaharan=as.numeric(as.factor(subsaharan))-1,
+                             muslperc,literacy))

> describe(Data)
           vars  n  mean    sd median trimmed   mad  min   max range  skew kurtosis   se
adrate        1 43  9.37  9.96      6    7.58  6.38  0.1  38.8  38.7  1.44     1.23 1.52
polity        2 43  0.51  5.41      0    0.46  7.41 -9.0  10.0  19.0  0.01    -1.38 0.82
subsaharan    3 43  0.86  0.35      1    0.94  0.00  0.0   1.0   1.0 -2.01     2.08 0.05
muslperc      4 43 35.96 34.58     20   32.87 29.65  0.0 100.0 100.0  0.68    -1.04 5.27
literacy      5 43 60.07 18.94     61   60.63 26.69 17.0  89.0  72.0 -0.20    -1.18 2.89

> cor(Data)
            adrate   polity subsaharan muslperc literacy
adrate      1.0000  0.11794    0.33129  -0.5709  0.51489
polity      0.1179  1.00000    0.52820  -0.2392 -0.05079
subsaharan  0.3313  0.52820    1.00000  -0.5773  0.09473
muslperc   -0.5709 -0.23917   -0.57725   1.0000 -0.61960
literacy    0.5149 -0.05079    0.09473  -0.6196  1.00000
```

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
-15.468  -4.395  -0.525   3.425  22.936

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6687    10.4113   -0.06    0.949
polity       -0.0139     0.2797   -0.05    0.961
subsaharan    3.7297     5.4309    0.69    0.496
muslperc     -0.0869     0.0628   -1.38    0.175
literacy      0.1657     0.0943    1.76    0.087 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.26 on 38 degrees of freedom
Multiple R-squared:  0.377,Adjusted R-squared:  0.312
F-statistic: 5.75 on 4 and 38 DF,  p-value: 0.00101
```

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 $H_0 : \beta_{\texttt{polity}} = \beta_{\texttt{subsaharan}} = 0$:

```
> 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 $H_0 : \beta_{\texttt{muslperc}} = 0.1$:

```
> 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 $H_0 : \beta_{\texttt{literacy}} = \beta_{\texttt{muslperc}}$:

```
> 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
```

The output:

```
> summary(model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-15.468  -4.395  -0.525   3.425  22.936

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6687    10.4113   -0.06    0.949
polity       -0.0139     0.2797   -0.05    0.961
subsaharan    3.7297     5.4309    0.69    0.496
muslperc     -0.0869     0.0628   -1.38    0.175
literacy      0.1657     0.0943    1.76    0.087 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.26 on 38 degrees of freedom
Multiple R-squared:  0.377,Adjusted R-squared:  0.312
F-statistic: 5.75 on 4 and 38 DF,  p-value: 0.00101
```

The table:

Table 1: OLS Regression Model of HIV/AIDS Rates in Africa, 2001

	Model I
(Constant)	-0.67
	(10.41)
POLITY Score	-0.01
	(0.28)
Subsaharan Africa	3.73
	(5.43)
Muslim Percentage of the Population	-0.09
	(0.06)
Literacy Rate	0.17^{*}
	(0.09)
Observations	43
R^2	0.38
Adjusted R^2	0.31
Residual Std. Error	8.26 (df = 38)
F Statistic	5.75^{*} (df = 4; 38)

Note: $N = 43$. Cell entries are coefficient estimates; numbers in parentheses are estimated standard errors. Asterisks indicate $p < .05$ (one-tailed). See text for details.

Multiple Models (`stargazer` defaults)

OLS Regression Models of HIV/AIDS Rates in Africa, 2001

	w/Literacy	w/o Literacy
(Constant)	-0.67	14.81^{**}
	(10.41)	(5.70)
POLITY Score	-0.01	-0.05
	(0.28)	(0.29)
Subsaharan Africa	3.73	0.53
	(5.43)	(5.25)
Muslim Percentage of the Population	-0.09	-0.16^{***}
	(0.06)	(0.05)
Literacy Rate	0.17^{*}	
	(0.09)	
Observations	43	43
R^2	0.38	0.33
Adjusted R^2	0.31	0.27
Residual Std. Error	8.26 (df = 38)	8.48 (df = 39)
F Statistic	5.75^{***} (df = 4; 38)	6.30^{***} (df = 3; 39)
Note:		*p<0.1; **p<0.05; ***p<0.01

R

- LaTeX: `texreg`, `xtable`, and `stargazer` 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

Some Guidelines ("Rules"?)

Tables:

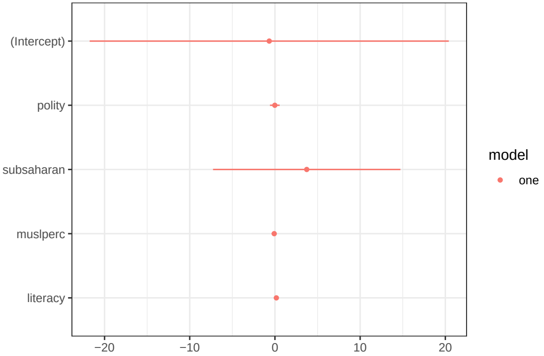
- *Use column headings descriptively.*
- *Use multiple rows / columns rather than multiple tables.*
- *Learn about significant digits, and don't report more than 4-5 of them (at the most).*
- *Use a figure to replace a table when you can.*
- *Be aware of norms about *s.*

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 `dotwhisker`)



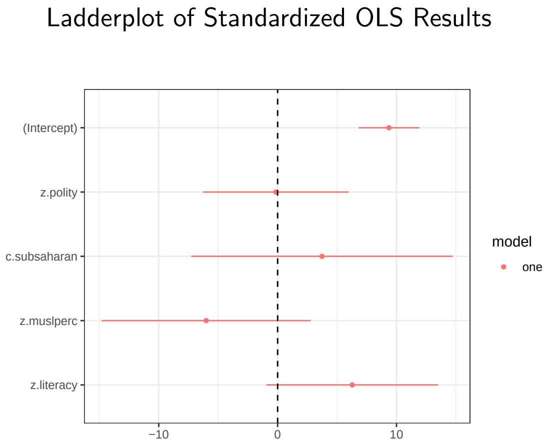
Rescale Covariates

A la Gelman (2008):

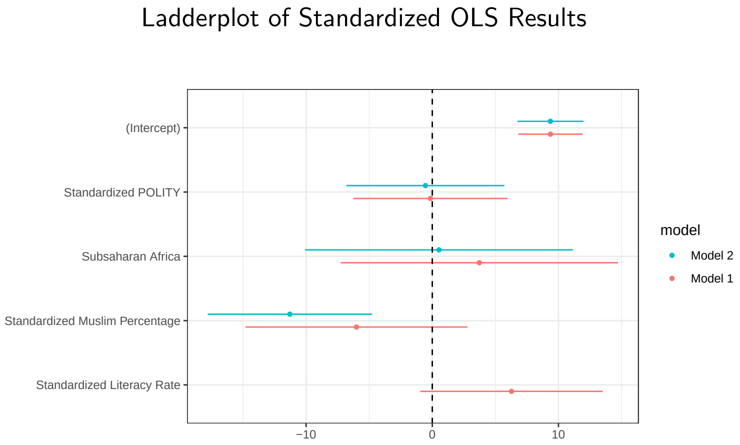
- Continuous = divide by two standard deviations
- Binary = mean 0, difference of 1 between the two categories

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

Call:
lm(formula = adrate ~ z.polity + c.subsaharan + z.muslperc +
    z.literacy, data = Data)

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)      9.37       1.26    7.43 0.0000000065 ***
z.polity        -0.15       3.03   -0.05        0.961
c.subsaharan     3.73       5.43    0.69        0.496
z.muslperc      -6.01       4.34   -1.38        0.175
z.literacy       6.28       3.57    1.76        0.087 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.26 on 38 degrees of freedom
Multiple R-squared:  0.377,Adjusted R-squared:  0.312
F-statistic: 5.75 on 4 and 38 DF,  p-value: 0.00101
```

A Better Regression Plot



An Even Better Regression Plot



Some Meta-Rules

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

Supplementary Materials

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

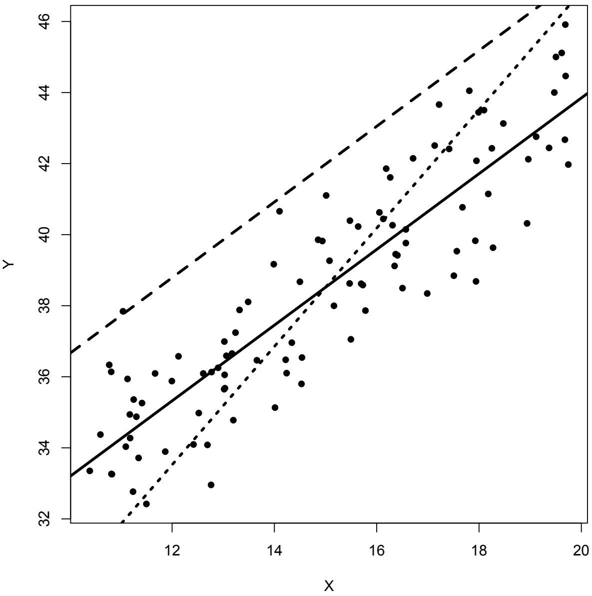
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and

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

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

Scatterplot: X and Y (with regression lines)



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

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

OLS (continued)

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

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

OLS (continued)

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:

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

and

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

“Stupid Regression Tricks”

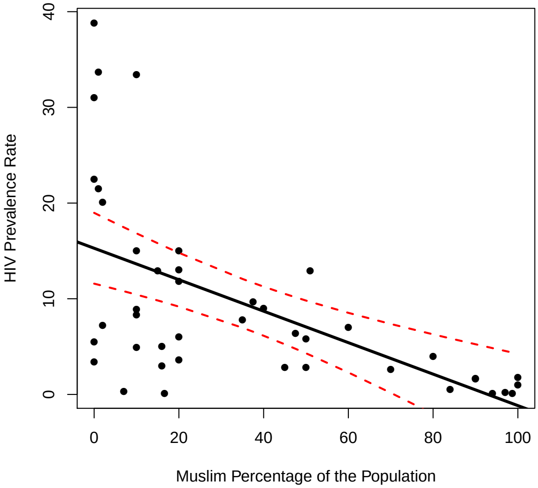
Africa (2001) Data

```
> africa<-read_csv("https://raw.githubusercontent.com/PrisonRodeo/GSERM-RFP-2022/master/Data/africa2001.csv")
> summary(africa)

    ccode          cabbr                         country      population           popthou      
 Min.   :404   AGO    : 1   Angola                  : 1   Min.   :   470000   Min.   :   470  
 1st Qu.:452   BDI    : 1   Benin                   : 1   1st Qu.:  3446000   1st Qu.:  3446  
 Median :510   BEN    : 1   Botswana                : 1   Median :  9662000   Median :  9662  
 Mean   :510   BWA    : 1   Burundi                 : 1   Mean   : 17388558   Mean   : 17390  
 3rd Qu.:556   CAF    : 1   Cameroon                : 1   3rd Qu.: 19150000   3rd Qu.: 19189  
 Max.   :651   CIV    : 1   Central African Republic: 1   Max.   :117000000   Max.   :116929  
               (Other):37   (Other)                 :37                                        
     popden            polity           gdppppd          tradegdp            war            adrate     
 Min.   :0.0022   Min.   :-9.000   Min.   : 0.500   Min.   :  4.03   Min.   :0.000   Min.   : 0.10  
 1st Qu.:0.0134   1st Qu.:-4.500   1st Qu.: 0.855   1st Qu.:  7.64   1st Qu.:0.000   1st Qu.: 2.70  
 Median :0.0357   Median : 0.000   Median : 1.200   Median : 13.56   Median :0.000   Median : 6.00  
 Mean   :0.0643   Mean   : 0.512   Mean   : 2.159   Mean   : 30.49   Mean   :0.116   Mean   : 9.37  
 3rd Qu.:0.0683   3rd Qu.: 5.500   3rd Qu.: 2.040   3rd Qu.: 30.01   3rd Qu.:0.000   3rd Qu.:12.90  
 Max.   :0.5740   Max.   :10.000   Max.   :10.800   Max.   :272.69   Max.   :1.000   Max.   :38.80  

   healthexp               subsaharan    muslperc         literacy      internalwar       intensity    
 Min.   :2.00   Not Sub-Saharan: 6   Min.   :  0.0   Min.   :17.0   Min.   :0.000   Min.   :0.000  
 1st Qu.:3.45   Sub-Saharan    :37   1st Qu.: 10.0   1st Qu.:43.0   1st Qu.:0.000   1st Qu.:0.000  
 Median :4.40                        Median : 20.0   Median :61.0   Median :0.000   Median :0.000  
 Mean   :4.60                        Mean   : 36.0   Mean   :60.1   Mean   :0.302   Mean   :0.581  
 3rd Qu.:5.80                        3rd Qu.: 55.5   3rd Qu.:78.5   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :8.60                        Max.   :100.0   Max.   :89.0   Max.   :1.000   Max.   :3.000  
```

A Simple Regression

```
> fit<-with(africa, lm(adrate~muslperc))
> summary(fit)

Call:
lm(formula = adrate ~ muslperc)

Residuals:
    Min      1Q  Median      3Q     Max
-13.828  -5.206   0.279   2.022  23.521

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  15.2787     1.8322    8.34 0.00000000023 ***
muslperc     -0.1644     0.0369   -4.45 0.00006390853 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.28 on 41 degrees of freedom
Multiple R-squared:  0.326,Adjusted R-squared:  0.31
F-statistic: 19.8 on 1 and 41 DF,  p-value: 0.0000639
```

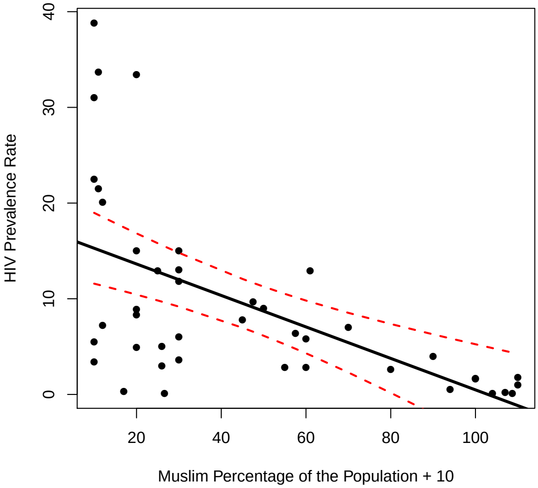
Scatterplot of HIV/AIDS Rates on Muslim Population Percentage, Africa 2001



Adding a Constant to X

```
> africa$muslplusten<-africa$muslperc+10
> fit2<-with(africa, lm(adrate~muslplusten,data=africa))
> summary(fit2)

Call:
lm(formula = adrate ~ muslplusten, data = africa)

Residuals:
    Min      1Q  Median      3Q     Max
-13.828  -5.206   0.279   2.022  23.521

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  16.9232     2.1152    8.00 0.00000000066 ***
muslplusten  -0.1644     0.0369   -4.45 0.00006390853 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.28 on 41 degrees of freedom
Multiple R-squared:  0.326,Adjusted R-squared:  0.31
F-statistic: 19.8 on 1 and 41 DF,  p-value: 0.0000639
```

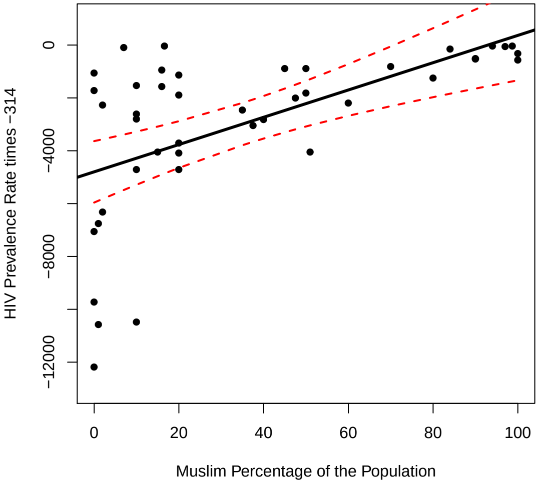
Scatterplot of HIV/AIDS Rates on Rescaled Muslim Population Percentage



Multiplying Y by a Constant

```
> africa$screwyrate<-africa$adrate*(-314)
> fit3<-with(africa, lm(screwyrate~muslperc))
> summary(fit3)

Call:
lm(formula = screwyrate ~ muslperc)

Residuals:
   Min     1Q Median     3Q    Max
 -7386   -635    -88   1635   4342

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -4797.5      575.3   -8.34 0.00000000023 ***
muslperc        51.6       11.6    4.45 0.00006390853 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2600 on 41 degrees of freedom
Multiple R-squared:  0.326,Adjusted R-squared:  0.31
F-statistic: 19.8 on 1 and 41 DF,  p-value: 0.0000639
```

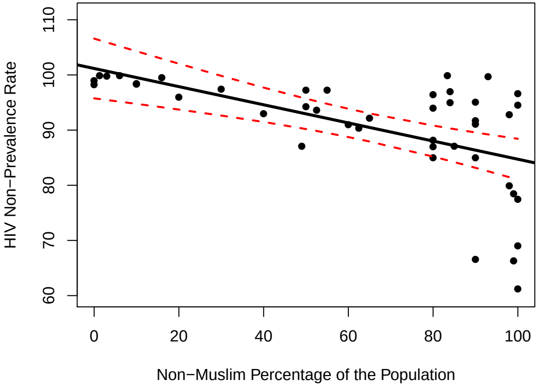
Scatterplot of Rescaled HIV/AIDS Rates on Muslim Population Percentage



Reversing the scales of X and Y

```
> africa$nonmuslimpct <- 100 - africa$muslperc
> africa$noninfected <- 100 - africa$adrate
> fit4<-lm(noninfected~nonmuslimpct,data=africa)
> summary(fit4)

Call:
lm(formula = noninfected ~ nonmuslimpct, data = africa)

Residuals:
    Min      1Q  Median      3Q     Max
-23.521  -2.022  -0.279   5.206  13.828

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  101.1660     2.6808   37.74  < 2e-16 ***
nonmuslimpct  -0.1644     0.0369   -4.45 0.000064 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.28 on 41 degrees of freedom
Multiple R-squared:  0.326,Adjusted R-squared:  0.31
F-statistic: 19.8 on 1 and 41 DF,  p-value: 0.0000639
```

Scatterplot of HIV/AIDS Non-Infection Rates on Non-Muslim Population Percentage



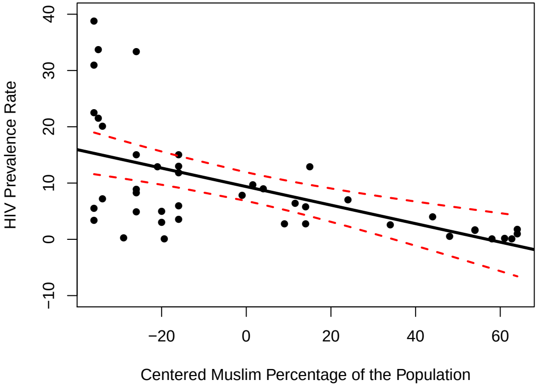
Linear Transformations

- Adding (subtracting) a positive constant to X shifts the X-axis to the left (right).
- Adding (subtracting) a positive constant to Y shifts the Y-axis downwards (upwards).
- Multiplying X (Y) times a positive constant greater than 1.0 stretches the X (Y) axis.
- Multiplying X (Y) times a positive constant less than 1.0 shrinks the X (Y) axis.
- Multiplying X (Y) times a negative constant inverts the X (Y) axis, and stretches / shrinks it as above.

Use: "Centering" a Variable

```
> africa$muslcenter<-africa$muslperc - mean(africa$muslperc, na.rm=TRUE)
> fit5<-lm(adrate~muslcenter,data=africa)
> summary(fit5)

Call:
lm(formula = adrate ~ muslcenter, data = africa)

Residuals:
    Min      1Q  Median      3Q     Max
-13.828  -5.206   0.279   2.022  23.521

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   9.3651     1.2622    7.42 0.0000000042 ***
muslcenter   -0.1644     0.0369   -4.45 0.0000639085 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.28 on 41 degrees of freedom
Multiple R-squared:  0.326,Adjusted R-squared:  0.31
F-statistic: 19.8 on 1 and 41 DF,  p-value: 0.0000639
```

Scatterplot of HIV/AIDS Infection Rates on (Centered) Muslim Population Percentage



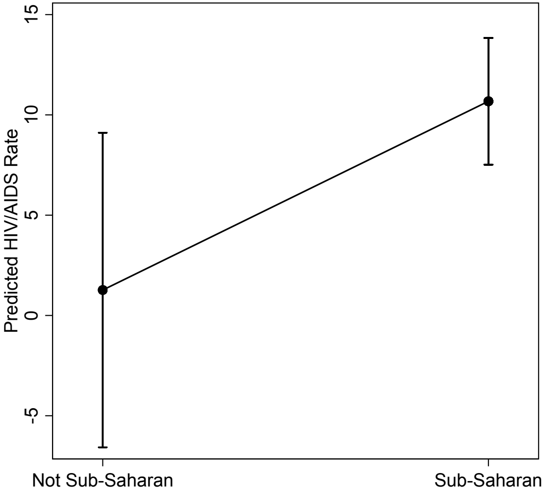
Use: Rescaling X for Interpretability

```
> fit6<-lm(adrate~population,data=africa)
> summary(fit6)

                Estimate    Std. Error t value Pr(>|t|)
(Intercept) 10.5883163475  1.9140361989    5.53 0.000002 ***
population  -0.0000000703  0.0000000671   -1.05      0.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.95 on 41 degrees of freedom
Multiple R-squared:  0.0261,Adjusted R-squared:  0.00234
F-statistic:  1.1 on 1 and 41 DF,  p-value: 0.301


> africa$popmil<-africa$population / 1000000
> fit7<-lm(adrate~popmil,data=africa)
> summary(fit7)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.5883     1.9140    5.53 0.000002 ***
popmil       -0.0703     0.0671   -1.05      0.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.95 on 41 degrees of freedom
Multiple R-squared:  0.0261,Adjusted R-squared:  0.00234
F-statistic:  1.1 on 1 and 41 DF,  p-value: 0.301
```

Dichotomous Xs: Bivariate Regression $\equiv$ t-test

```
> fit8<-lm(adrate~subsaharan,data=africa)
> summary(fit8)

Residuals:
   Min     1Q Median     3Q    Max
-10.58  -6.23  -1.78   2.22  28.12

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.27       3.88    0.33     0.75
subsaharanSub-Saharan     9.41       4.19    2.25     0.03 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.51 on 41 degrees of freedom
Multiple R-squared:  0.11,Adjusted R-squared:  0.088
F-statistic: 5.05 on 1 and 41 DF,  p-value: 0.03


> with(africa,
+      t.test(adrate~subsaharan, var.equal=TRUE))

Two Sample t-test

data:  adrate by subsaharan
t = -2.2, df = 41, p-value = 0.03
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.8659  -0.9576
sample estimates:
mean in group Not Sub-Saharan     mean in group Sub-Saharan
                        1.267                        10.678
```

Expected Values of HIV/AIDS Infection Rates in Saharan and Sub-Saharan Africa



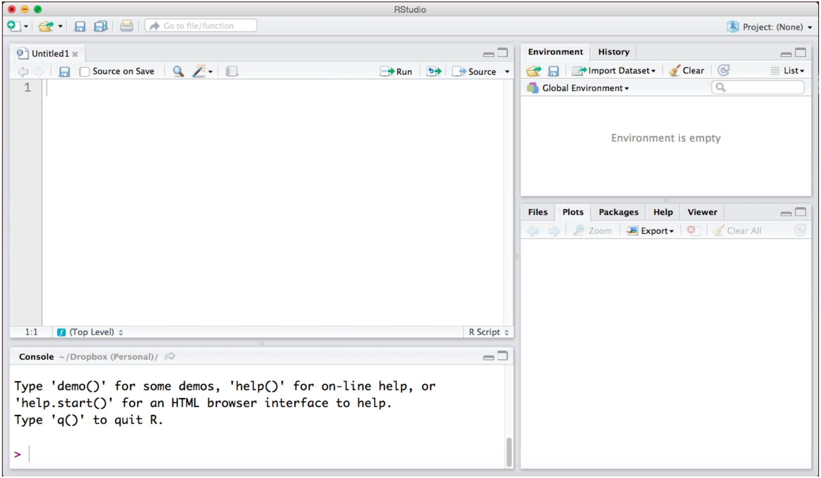
R Things

R

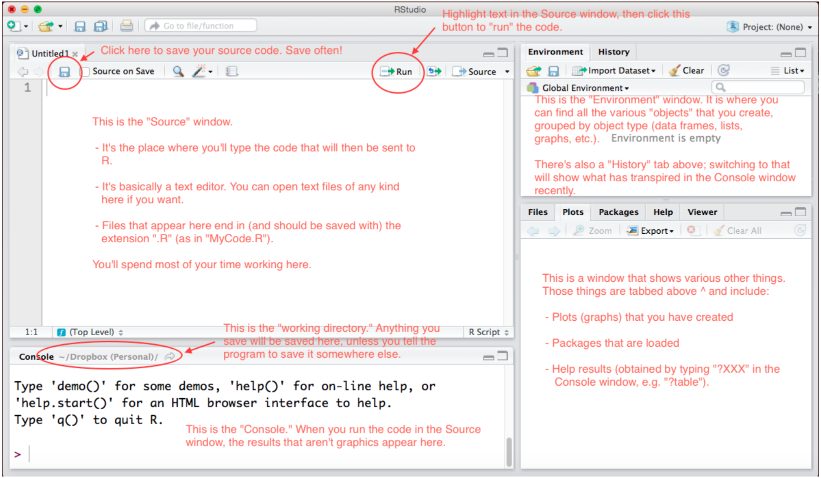
R Console

```
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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 1.70 (7735) x86_64-apple-darwin15.6.0]

[Workspace restored from /Users/cuz10lcl/.RData]
[History restored from /Users/cuz10lcl/.Rapp.history]

>
```

RStudio


RStudio
Go to file/function
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Import Dataset
Clear
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Zoom
Export
Clear All
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.
>

RStudio (annotated)



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-2022-R-Intro.R` in the github repo)

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).