PLSC 503 – Spring 2022

“Stupid Regression Tricks” + Multivariate Regression

February 2, 2022

Stupid Regression Tricks

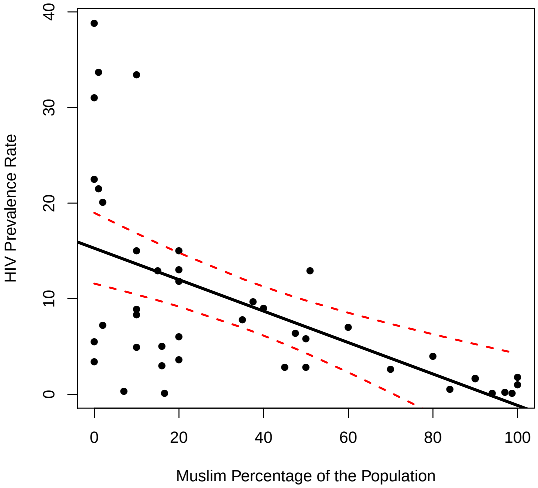
Africa (2001) Data

```
> temp<-getURL("https://raw.githubusercontent.com/PrisonRodeo/PLSC503-2022-git/master/Data/africa2001.csv")
> africa<-read.csv(text=temp, header=TRUE)
> 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          gdpppppd          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 Very 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
```

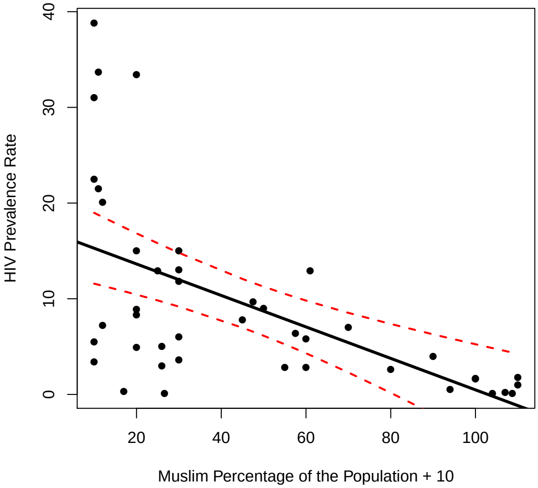
Scatterplot of HIV/AIDS Rates on Muslim Population Percentage, Africa 2001



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
```

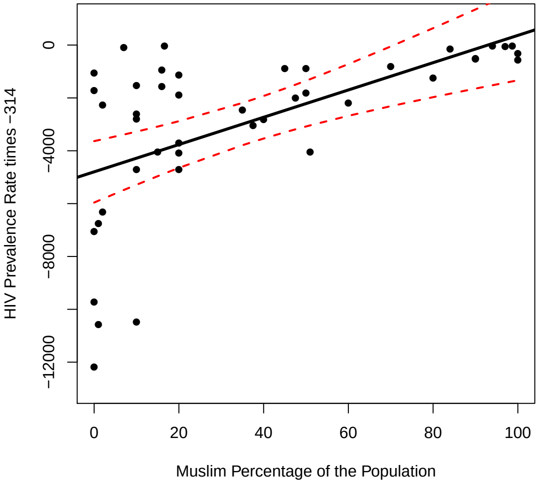
Scatterplot of HIV/AIDS Rates on Rescaled Muslim Population Percentage



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
```

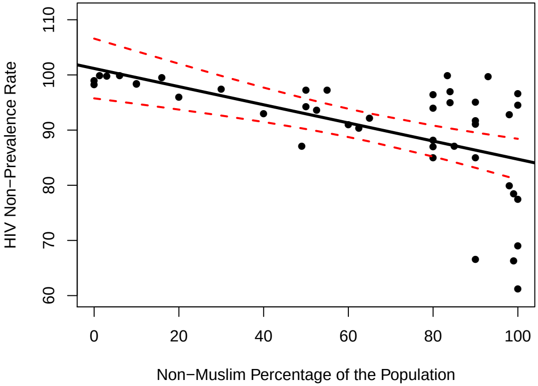
Scatterplot of Rescaled HIV/AIDS Rates on Muslim Population Percentage



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

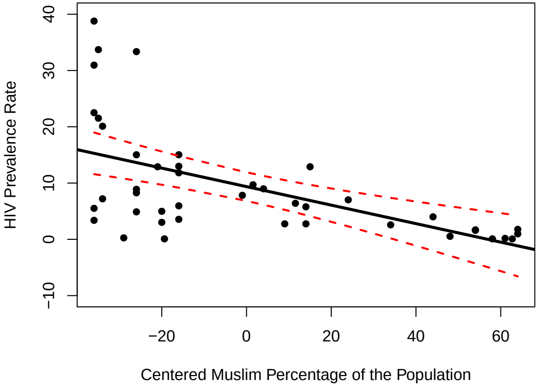


- Adding (subtracting) a positive constant to X shifts the X-axis to the left (right).
- Adding (subtracting) a positive constant to Y shifts the Y-axis downwards (upwards).
- Multiplying X (Y) times a positive constant greater than 1.0 stretches the X (Y) axis.
- Multiplying X (Y) times a positive constant 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

> mean(africa$adrate)
[1] 9.365116
```

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



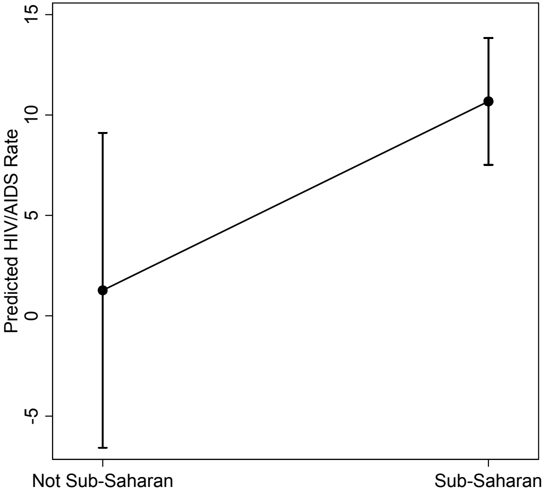
Use: Rescaling X for Interpretability

```
> 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



The results:

```
> fit<-lm(adrate~muslperc, data=africa)
> summary.lm(fit)

Call:
lm(formula = adrate ~ muslperc, 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)  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
```

The table:

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

Variables	Model I
(Constant)	15.28
	(1.83)
Muslim Percentage of the Population	-0.164*
	(0.037)
Adjusted R^2	0.31

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.*

Another Table (using default-y `stargazer`)

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

	Model I
(Constant)	15.28^{***}
	(1.83)
Muslim Percentage of the Population	-0.16^{***}
	(0.04)
Observations	43
R^2	0.33
Adjusted R^2	0.31
Residual Std. Error	8.28 (df = 41)
F Statistic	19.83^{***} (df = 1; 41)
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.*
- *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.*

“Multivariate” Regression

The Model

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

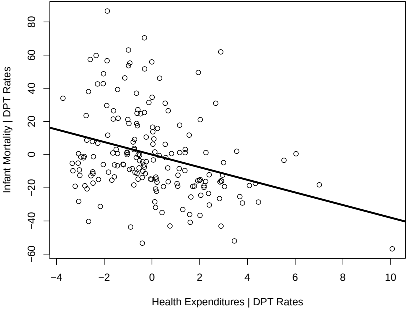
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + u_i$$

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

Diversion: "Added Variable Plots"

- Regress Y on X_1 and save the residuals $\hat{u}_i$,
- Regress X_2 on X_1 and save the residuals (call these $\hat{e}_i$),
- Plot $\hat{u}_i$ (conventionally on the y-axis) vs. $\hat{e}_i$ (conventionally on the x-axis).

Example: Infant Mortality and Health Expenditures Given DPT Immunization Rates



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:

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

> *"Do not compute the least squares estimates using $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$!"*
>
> – *Weisberg (p. 61)*

Most software uses:

$$\mathbf{X} = \mathbf{QR}$$

where $\mathbf{Q}$ is orthogonal ($\mathbf{Q}'\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is upper-triangular.

Why??? See e.g. here, or section 3.19, here

1. Expectation-Zero Disturbances

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

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:

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

3. "Fixed" **X**...

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

4. **X** is full column rank.

Means:

- no exact linear relationship among **X**, and
- $K < N$.

5. Normal Disturbances

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

Unbiasedness:

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

Substitute OLS $\hat{\boldsymbol{\beta}}$:

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

and so:

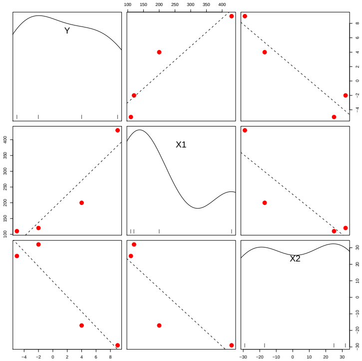
$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

By $\mathrm{Cov}(\mathbf{X}, \mathbf{u}) = \mathbf{0}$, we have $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

A Toy Example

$$\mathbf{Y} = \begin{bmatrix} 4 \\ -2 \\ 9 \\ -5 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 200 & -17 \\ 1 & 120 & 32 \\ 1 & 430 & -29 \\ 1 & 110 & 25 \end{bmatrix}$$



Example, continued

So:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 860 & 11 \\ 860 & 251400 & -9280 \\ 11 & -9280 & 2779 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 3.2453 & -0.0132 & -0.05694 \\ -0.0132 & 0.000058 & 0.0002468 \\ -0.0569 & 0.000247 & 0.001409 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 6 \\ 3880 \\ 518 \end{bmatrix}$$

So:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{bmatrix} 3.2453 & -0.0132 & -0.05694 \\ -0.0132 & 0.000058 & 0.0002468 \\ -0.0569 & 0.000247 & 0.001409 \end{bmatrix} \begin{bmatrix} 6 \\ 3880 \\ 518 \end{bmatrix} \\ &= \begin{bmatrix} -2.264 \\ 0.0190 \\ -0.1141 \end{bmatrix} \end{aligned}$$

Minimal Example: Correlation

```
Y<-c(4,-2,9,-5)
X1<-c(200,120,430,110)
X2<-c(-17,32,-29,25)
data<-cbind(Y,X1,X2)
```

```
cor(data)
```

```
         Y      X1      X2
Y   1.0000  0.9285 -0.9425
X1  0.9285  1.0000 -0.8613
X2 -0.9425 -0.8613  1.0000
```

```
fit<-lm(Y~X1+X2)
summary(fit)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     1      2      3      4
 0.531  1.639 -0.201 -1.970

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.2643     4.7284   -0.48     0.72
X1            0.0190     0.0200    0.95     0.52
X2           -0.1141     0.0985   -1.16     0.45

Residual standard error: 2.62 on 1 degrees of freedom
Multiple R-Squared: 0.941,Adjusted R-squared: 0.823
F-statistic: 7.99 on 2 and 1 DF,  p-value: 0.243
```

- Pick some $\mathbf{H}_A : \boldsymbol{\beta} = \boldsymbol{\beta}_A$
- Estimate $\hat{\boldsymbol{\beta}}$
- Determine distribution of $\hat{\boldsymbol{\beta}}$ under $\mathbf{H}_A$
- Form a *test statistic* $\hat{\mathbf{S}} = h(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$
- Assess $\Pr(\hat{\mathbf{S}}|\mathbf{H}_A)$

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:

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

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

Taking expectations:

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

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

Empirical estimate:

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

Yields:

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

Single Coefficient Hypothesis Tests

We know that:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

In practice, using $\hat{\sigma}^2$ means

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim t_{N-K}$$

<u>Procedure</u>:

- Choose a value of β_k that you want to test (say, $\beta_k = 0$),
- Calculate the t-statistic for the coefficient associated with X_k, which is:

$$\hat{t}_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\mathbf{V}(\hat{\beta}_k)}}}$$

- Compare $\hat{t}_k$ to a t distribution with $N - K$ degrees of freedom.

E.g.: $\quad \mathsf{H}_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$

or: $\quad \mathsf{H}_0 : \beta_3 = \beta_6 = 0$

Generally: *Linear restrictions*:

$$\underset{q\times k}{\mathbf{R}}\,\underset{k\times 1}{\boldsymbol{\beta}} = \underset{q\times 1}{\mathbf{r}}$$

E.g.:

$$\beta_2 = -2 \iff (0\ 1\ 0)\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = -2$$

The F-test

Recall:

$$\textbf{TSS} \quad = \quad \textbf{MSS} \quad + \quad \textbf{RSS}$$

Consider:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{Ui}$$

and the restriction:

$$\mathrm{H}_a : \beta_2 = \beta_4 = 0.$$

Restricted model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + 0 X_{2i} + \beta_3 X_{3i} + 0 X_{4i} + u_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + u_{Ri} \end{aligned}$$

F-tests: Sums of Squared Residuals

"Unrestricted":

$$\text{RSS}_U \equiv \hat{\mathbf{u}}_U' \hat{\mathbf{u}}_U = \sum_{i=1}^{N} \hat{u}_{Ui}^2$$

"Restricted":

$$\text{RSS}_R \equiv \hat{\mathbf{u}}_R' \hat{\mathbf{u}}_R = \sum_{i=1}^{N} \hat{u}_{Ri}^2$$

F-statistic:

$$\mathbf{F} = \frac{(\mathrm{RSS}_R - \mathrm{RSS}_U)/q}{\mathrm{RSS}_U/(N-K)}$$

$$= \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/N-K}$$

Testing:

$$\mathbf{F} \sim F_{q,N-K}$$

Consider:

$$
\begin{aligned}
\mathrm{H}_b: \quad & \beta_1 + \beta_4 = 1 \\
& \beta_1 = 1 - \beta_4
\end{aligned}
$$

Implies:

$$
\begin{aligned}
Y_i &= \beta_0 + (1 - \beta_4)X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{R'i} \\
&= \beta_0 + X_{1i} - \beta_4 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{R'i} \\
&= \beta_0 + X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{4i} - X_{1i}) + u_{R'i}
\end{aligned}
$$

implying restricted model:

$$
Y_i - X_{1i} = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{4i} - X_{1i}) + u_{R'i}
$$

$$F = \frac{(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)' \hat{\mathbf{V}}_q^{-1} (\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)}{q\hat{\sigma}^2}$$

Implies:

$$\Pr\left[\frac{(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)' \hat{\mathbf{V}}_q^{-1} (\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)}{q\hat{\sigma}^2} \leq F_{q,N-K}\right] = 1 - \alpha. \tag{1}$$

$\rightarrow$ "confidence region" of all points satisfying:

$$(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)' \hat{\mathbf{V}}_q^{-1} (\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H) \leq q\hat{\sigma}^2 F_{q,N-K}.$$

Prediction:

$$\hat{Y}_j = \mathbf{X}_j \hat{\boldsymbol{\beta}}$$

Variance:

$$\widehat{\mathbf{V}(\hat{Y}_j)} = \hat{\sigma}^2[1 + \mathbf{X}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j']$$

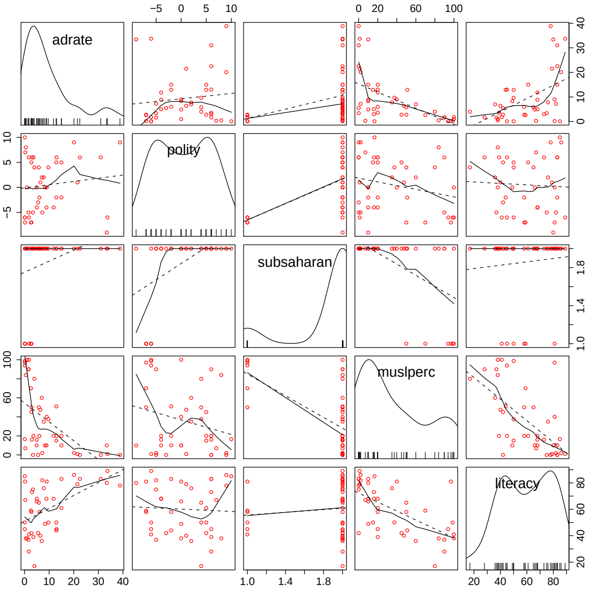
Standard error:

$$\widehat{\text{s.e.}(\hat{Y}_j)} = \sqrt{\hat{\sigma}^2[1 + \mathbf{X}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j']}$$

Example: Africa Data

```
> library(RCurl)
> temp<-getURL("https://raw.githubusercontent.com/PrisonRodeo/PLSC503-2022-git/master/Data/africa2001.csv")
> Data<-read.csv(text=temp, header=TRUE)
> Data<-with(Data, data.frame(adrate,polity,
+           subsaharan=as.numeric(as.factor(subsaharan))-1,
+           muslperc,literacy))

> summary(Data)
     adrate          polity         subsaharan       muslperc        literacy   
 Min.   : 0.10   Min.   :-9.000   Min.   :0.000   Min.   :  0.0   Min.   :17.0  
 1st Qu.: 2.70   1st Qu.:-4.500   1st Qu.:1.000   1st Qu.: 10.0   1st Qu.:43.0  
 Median : 6.00   Median : 0.000   Median :1.000   Median : 20.0   Median :61.0  
 Mean   : 9.37   Mean   : 0.512   Mean   :0.861   Mean   : 36.0   Mean   :60.1  
 3rd Qu.:12.90   3rd Qu.: 5.500   3rd Qu.:1.000   3rd Qu.: 55.5   3rd Qu.:78.5  
 Max.   :38.80   Max.   :10.000   Max.   :1.000   Max.   :100.0   Max.   :89.0  

> 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
-15.4681  -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 $H_0 : \beta_{\texttt{polity}} = \beta_{\texttt{subsaharan}} = 0$:

```
> 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
```

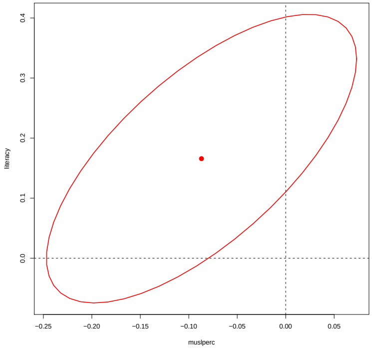
Test $H_0 : \beta_{\texttt{muslperc}} = 0.1$:

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

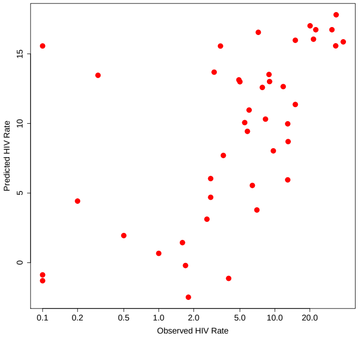
```
> linearHypothesis(model,"literacy=muslperc")

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
```

```
> confidenceEllipse(model=model,which.coef=c(4,5),
                    xlab="Muslim Percentage",ylab="Literacy")
> abline(h=0,v=0,lty=2)
```



```
> hats<-fitted(model)
> # Or, alternatively:
> fitted<-predict(model,se.fit=TRUE, interval=c("confidence"))
> scatterplot(model$fitted~adrate,log="x",smooth=FALSE,boxplots=FALSE,
  reg.line=FALSE,xlab="Observed HIV Rate",ylab="Predicted HIV Rate",
  pch=16,cex=2)
```

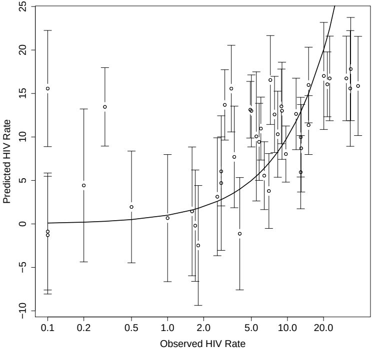
Predicted and Actual HIV/AIDS Rates (X-Axis Logged)



An Even More Useful Plot

```
> library(plotrix)
> plotCI(Data$adrate,model$fitted,uiw=(1.96*(fitted$se.fit)),
         log="x",xlab="Observed HIV Rate",ylab="Predicted HIV Rate")
> lines(lowess(Data$adrate,Data$adrate),lwd=2)
```

Predicted and Actual HIV/AIDS Rates, with 95% C.I.s

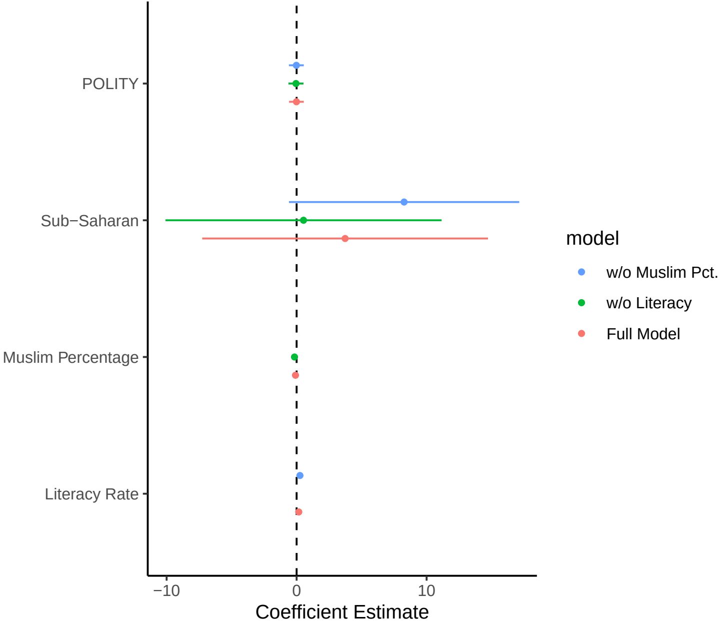


Presentation: A (De)Fault-y Table

```
> M1<-lm(adrate~polity+subsaharan+muslperc+literacy,data=Data)
> M2<-lm(adrate~polity+subsaharan+muslperc,data=Data)
> M3<-lm(adrate~polity+subsaharan+literacy,data=Data)
>
> stargazer(M1,M2,M3)
```

	Dependent variable:		
	adrate		
	(1)	(2)	(3)
polity	−0.014	−0.051	−0.020
	(0.280)	(0.286)	(0.283)
subsaharan	3.730	0.530	8.268*
	(5.431)	(5.252)	(4.379)
muslperc	−0.087	−0.163***	
	(0.063)	(0.047)	
literacy	0.166*		0.256***
	(0.094)		(0.069)
Constant	−0.669	14.800**	−13.120**
	(10.410)	(5.701)	(5.298)
Observations	43	43	43
R^2	0.377	0.326	0.346
Adjusted R^2	0.312	0.275	0.295
Residual Std. Error	8.264 (df = 38)	8.483 (df = 39)	8.361 (df = 39)
F Statistic	5.751*** (df = 4; 38)	6.302*** (df = 3; 39)	6.870*** (df = 3; 39)
Note:			*p<0.1; **p<0.05; ***p<0.01

A Dot-Whisker Plot



Gelman (2008 *Statistics in Medicine*)

Suggestion: Rescale *all* non-binary predictors by **dividing them by two times their standard deviation**.

- Creates a "common scale" for every predictor.
- More specifically: Scales continuous predictors to be comparable to binary (0/1) ones.
- $\hat{\beta}_X$ now represents the change in $E(Y)$ associated with a change in X of two standard deviations (for example, from one s.d. below the mean to one s.d. above the mean).

Note that:

- People don't *routinely* (or even generally) do this. But...
- ...it can be very useful when you have predictor variables that are measured on very different "natural" scales.

A (Better?) Dot-Whisker Plot