PLSC 503 – Spring 2022
Bivariate Regression

January 26, 2022

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

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

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*

Regression (continued)

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

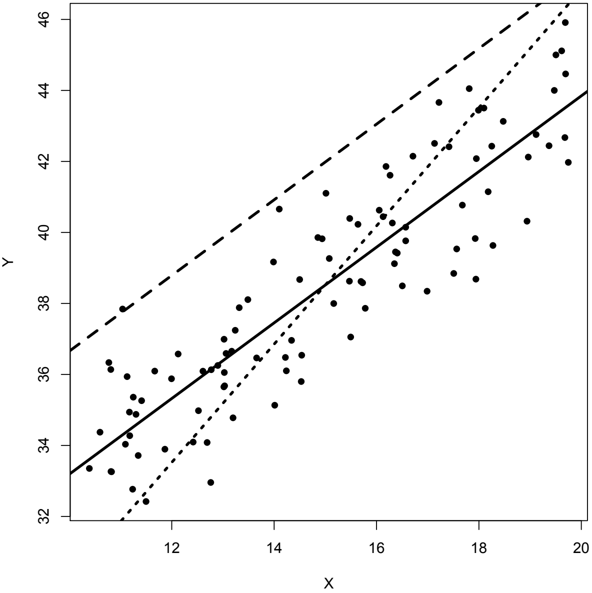
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3}$$

and

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

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

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

Yields:

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

and

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

OLS (continued)

Solving yields:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \tag{5}$$

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

and

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

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\hat{Y} + \hat{u}) \\
&= \mathrm{Var}(\hat{Y}) + \mathrm{Var}(\hat{u}) + 2\,\mathrm{Cov}(\hat{Y}, \hat{u}) \\
&= \underbrace{\mathrm{Var}(\hat{Y})}_{\text{“Systematic”}} + \underbrace{\mathrm{Var}(\hat{u})}_{\text{“Stochastic”}}
\end{aligned}
$$

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

The World's Simplest Regression

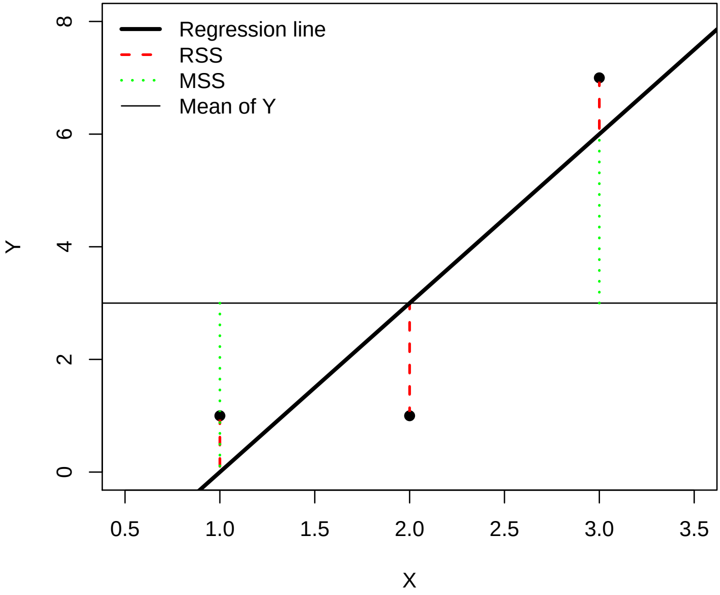
Data:

```
X Y
1 1
2 1
3 7
```

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
	1	1	-1	-2	1	4	2
	2	1	0	-2	0	4	0
	3	7	1	4	1	16	4
$\sum_{i=1}^{3}(\cdot) =$	6	9	0	0	2	24	6

- $\hat{\beta}_1 = \frac{6}{2} = 3$
- $\hat{\beta}_0 = 3 - (3 \times 2) = -3$

The World's Simplest Regression



The World's Simplest Regression

```
> X<-c(1,2,3)
> Y<-c(1,1,7)
> WSR<-lm(Y~X)
> summary(WSR)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.000      3.742  -0.802    0.570
X              3.000      1.732   1.732    0.333

Residual standard error: 2.449 on 1 degrees of freedom
Multiple R-squared:   0.75,Adjusted R-squared:    0.5
F-statistic:     3 on 1 and 1 DF,  p-value: 0.3333
```

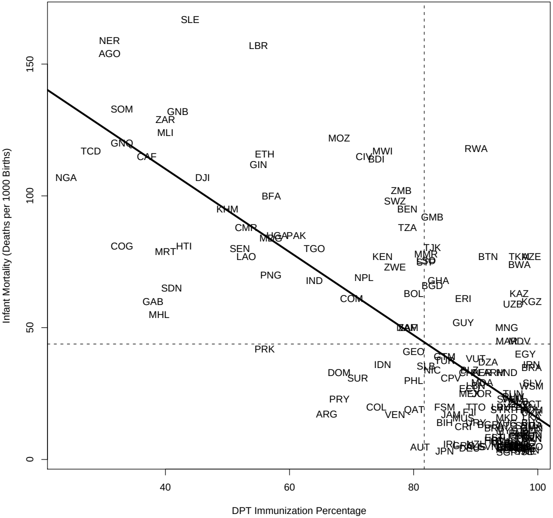
Running Example: Infant Mortality

```
> url <- getURL("https://raw.githubusercontent.com/PrisonRodeo/
    PLSC503-2022-git/master/Data/CountryData2000.csv")
> Data <- read.csv(text = url) # read the "countries" data
> rm(url)
>
> # Summary statistics
>
> # install.packages("psych") <- Install psych package, if necessary
> library(psych)

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

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

OLS Regression

```
> IMDPT<-lm(infantmortalityperK~DPTpct,data=Data,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
```

Scatterplot: Infant Mortality and DPT Immunization Rates



```
> 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
```

$$\begin{aligned} TSS &= \text{total variability in } Y \text{ around its mean} \\ &= \sum (Y_i - \bar{Y})^2 \\ &= 167423 + 120033 \\ &= \mathbf{287456} \end{aligned}$$

$$\begin{aligned} MSS(\equiv \texttt{DPTpct}) &= \text{model ("explained" or "regression") sum of squares} \\ &= \sum (\hat{Y}_i - \bar{Y})^2 \\ &= \mathbf{167423} \end{aligned}$$

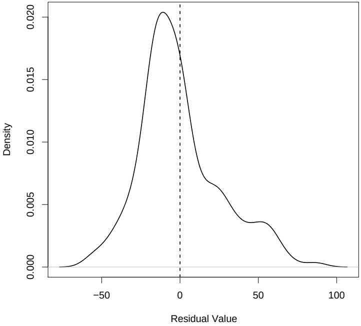
$$\begin{aligned} RSS(\equiv \texttt{Residuals}) &= \text{residual ("unexplained" or "error") sum of squares} \\ &= \sum \hat{u}_i^2 \\ &= \mathbf{120033} \end{aligned}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{RSS}{N-k} \\ &= \frac{\sum \hat{u}_i^2}{N-2} \\ &= \frac{120033}{175} \\ &= \mathbf{686} \end{aligned}$$

$$\begin{aligned} \hat{\sigma} &= \text{"SEE" (the standard error of the estimate, or } \texttt{Residual standard error}) \\ &= \sqrt{\hat{\sigma}^2} \\ &= \sqrt{686} \\ &= \mathbf{26.2} \end{aligned}$$

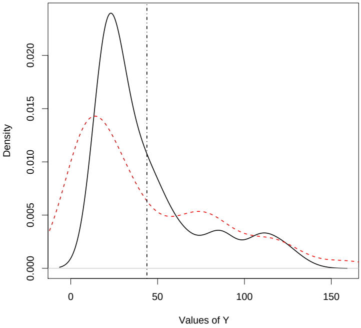
Fitted Values, Residuals, etc.

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

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



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

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

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



Some Correlations

$\text{Corr}(Y, X)$:

```
> with(Data, cor(infantmortalityperK,DPTpct,use="complete.obs"))
[1] -0.7632
```

$\text{Corr}(\hat{u}, Y)$:

```
> with(Data, cor(IMDPTres,infantmortalityperK,use="complete.obs"))
[1] 0.6462
```

$\text{Corr}(\hat{u}, X)$:

```
> with(Data, cor(IMDPTres,DPTpct,use="complete.obs"))
[1] 9.573e-17
```

$\text{Corr}(\hat{Y}, Y)$:

```
> with(Data, cor(IMDPThat,infantmortalityperK,use="complete.obs"))
[1] 0.7632
```

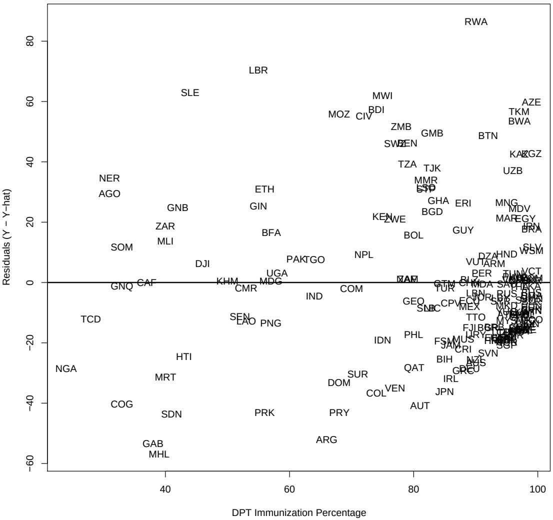
$\text{Corr}(\hat{Y}, X)$:

```
> with(Data, cor(IMDPThat,DPTpct,use="complete.obs"))
[1] -1
```

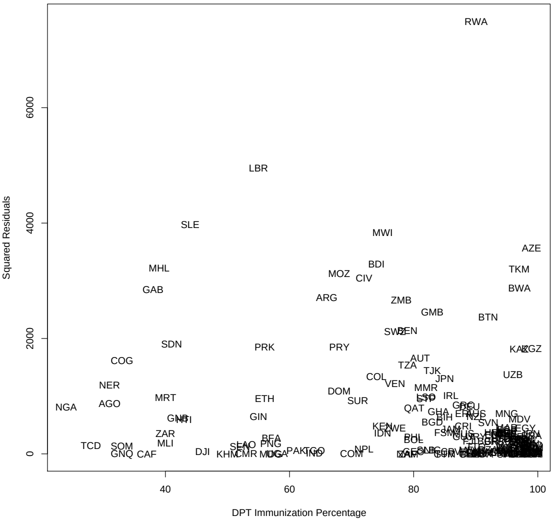
$\text{Corr}(\hat{u}, \hat{Y})$:

```
> with(Data, cor(IMDPTres,IMDPThat,use="complete.obs"))
[1] 5.302e-17
```

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



Squared Residuals vs. DPT Percentage



Inference and Model Fit

The key point:

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables.

Due to (*inter alia*):

- **Sampling variability**: Random samples from a population $\rightarrow$ slightly different $\hat{\beta}_0$s and $\hat{\beta}_1$s.
- **Random variability in X**: In cases where X is also a random variable...
- **Intrinsic variability in Y**: Because $Y_i = \mu + u_i$.

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

meaning:

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

so:

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

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

Similarly:

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

and :

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

Important Things

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

For:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Rewrite:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \left[\frac{\sum_{i=1}^{N}(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right] Y_i.$$

Define "weights" k:

$$\hat{\beta}_1 = \sum k_i Y_i$$

with $k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$.

Alternative (non-LS) estimator:

$$\tilde{\beta}_1 = \sum w_i Y_i$$

Unbiasedness requires:

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_1) &= \sum w_i \mathrm{E}(Y_i) \\ &= \sum w_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i X_i \end{aligned}$$

Variance:

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}_1) &= \mathrm{Var}\left(\sum w_i Y_i\right) \\
&= \sigma^2 \sum w_i^2 \\
&= \sigma^2 \sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} + \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 \\
&= \sigma^2 \sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 + \sigma^2 \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]
\end{aligned}
$$

Because $\sigma^2\left[\frac{1}{\sum(X_i-\bar{X})^2}\right]$ is a constant, min[Var($\tilde{\beta}_1$)] minimizes

$$\sum\left[w_i - \frac{X_i-\bar{X}}{\sum(X_i-\bar{X})^2}\right]^2.$$

Minimized at:

$$w_i = \frac{X_i-\bar{X}}{\sum(X_i-\bar{X})^2}.$$

implying:

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

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

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

and

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

Which means:

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

$\sigma^2 =$???

Solution: use

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

Yields:

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

and

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

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

implies:

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

Predictions and Variance

Point prediction:

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

Y_k is unbiased:

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

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

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

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:

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

Back to the Example

```
> summary(IMDPT)

Call:
lm(formula = infantmortalityperK ~ DPTpct, data = IMdata)

Residuals:
   Min     1Q Median     3Q    Max
 -56.8  -16.3   -5.1   11.8   86.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  173.277      8.489    20.4   <2e-16 ***
DPTpct        -1.576      0.101   -15.6   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 26.2 on 175 degrees of freedom
Multiple R-Squared: 0.582,Adjusted R-squared: 0.58
F-statistic:  244 on 1 and 175 DF,  p-value: <2e-16
```

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

```
> vcov(IMDPT)

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

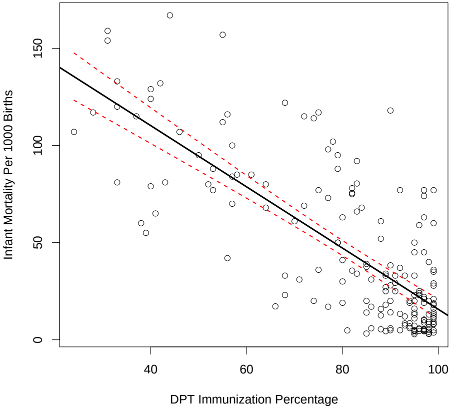
95 percent c.i.s:

```
> confint(IMDPT)

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

```
> SEs<-predict(IMDPT,interval="confidence")
> SEs
      fit   lwr   upr
1    25.10 20.53 29.68
3    17.22 12.05 22.40
4    23.53 18.84 28.21
.
.
<rows omitted>
.
.
189  21.95 17.15 26.75
190  39.29 35.36 43.23
191  17.22 12.05 22.40
```

A Plot, With Confidence Intervals

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



Model Fit

- The closeness of the mapping between model-based values of Y and actual values of Y...
- Can be *in-sample* or *out-of-sample* ($\rightarrow$ "overfitting")
- Is (in part) a function of *model specification* (choice of predictors, functional form, interactions, etc.)
- Related (but not identical) to prediction / predictive ability

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

R^2 Is...

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

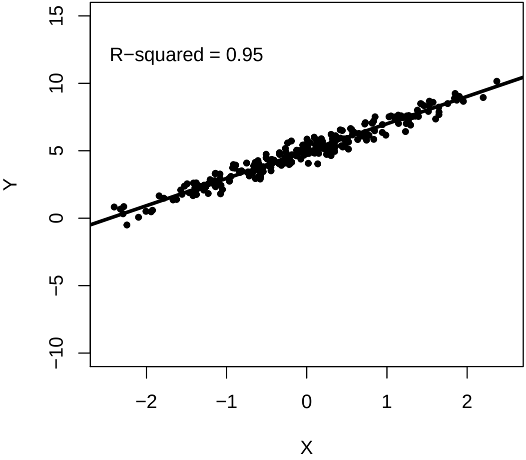
For a single X,

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

A (Simulated) Example

```
seed <- 7222009
set.seed(seed)
> X<-rnorm(250)
> Y1<-5+2*X+rnorm(250,mean=0,sd=sqrt(0.2))
> Y2<-5+2*X+rnorm(250,mean=0,sd=sqrt(20))
> fit<-lm(Y1~X)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.97712    0.02846  174.86   <2e-16 ***
X            2.02529    0.02785   72.73   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4491 on 248 degrees of freedom
Multiple R-squared:  0.9552,Adjusted R-squared:  0.955
F-statistic:  5290 on 1 and 248 DF,  p-value: < 2.2e-16
```

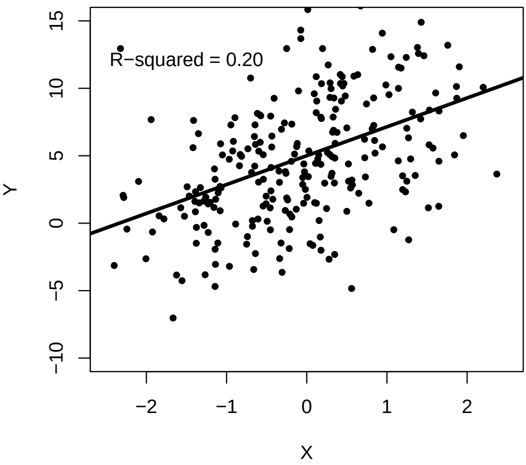
Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.95$)



Same Slope/Intercept, Different R^2

```
> fit2<-lm(Y2~X)
> summary(fit2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0048     0.2757  18.151  < 2e-16 ***
X             2.1402     0.2697   7.934 7.29e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.351 on 248 degrees of freedom
Multiple R-squared:  0.2024,Adjusted R-squared:  0.1992
F-statistic: 62.95 on 1 and 248 DF,  p-value: 7.288e-14
```

Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.20$)



R^2 is Also an *Estimate*...

Luskin: Population analogue "P^2":

$$P^2 = 1 - \frac{\sigma^2}{\sigma_Y^2}$$

Then $\hat{P}^2 = R^2$ has variance:

$$\widehat{\mathrm{Var}(R^2)} = \frac{4R^2(1-R^2)^2(N-k)^2}{(N^2-1)(N+3)}$$

and standard error:

$$\widehat{\mathrm{s.e.}(R^2)} = \sqrt{\frac{4R^2(1-R^2)^2(N-k)^2}{(N^2-1)(N+3)}}.$$

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

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:

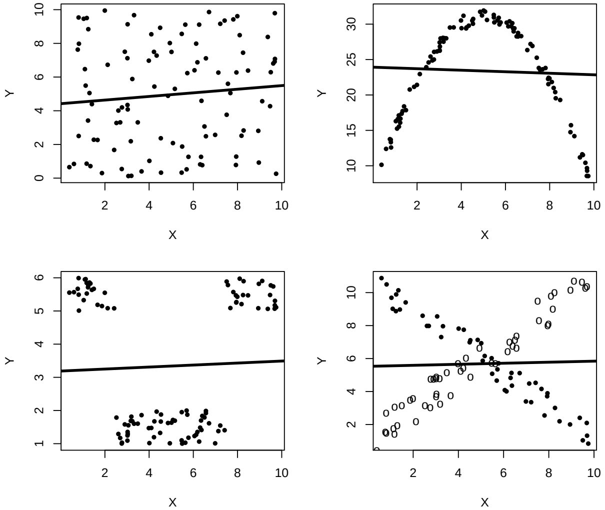
$$\text{SEE} = \sqrt{\frac{\text{RSS}}{N - k}}$$

- F-tests (later...)

- ROC / AUC (later...)

- Graphical methods

Caution: Different Ways to get $R^2 \approx 0$



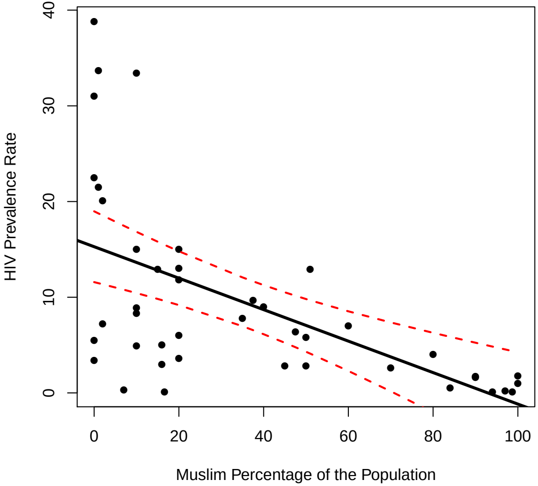
Stupid Regression Tricks

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          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
```

```
> 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
```

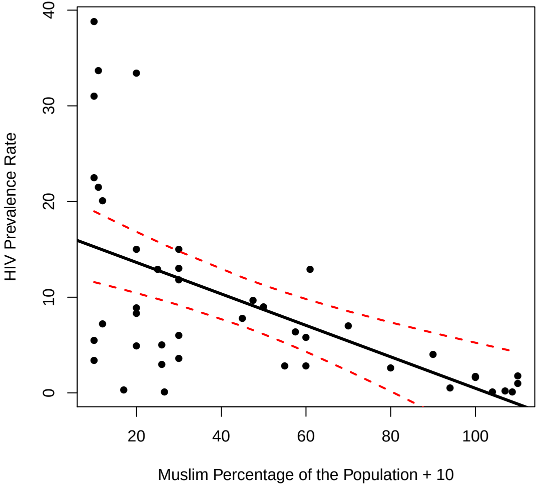
Scatterplot of HIV/AIDS Rates on Muslim Population Percentage, Africa 2001



Adding a Constant to X

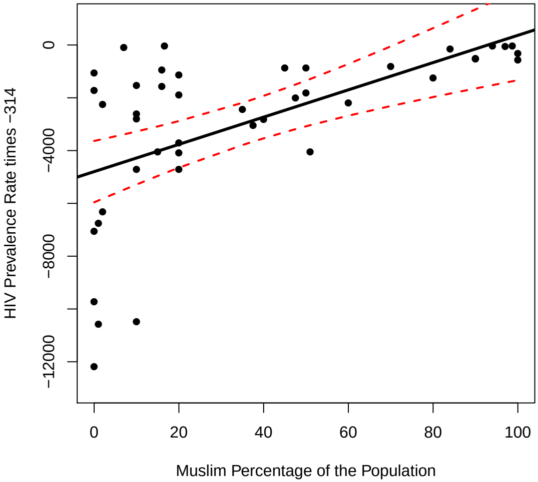
```
> africa$muslplusten<-africa$muslperc+10
> fit2<-with(africa, lm(adrate~muslplusten,data=africa))
> summary(fit2)

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



```
> 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
```

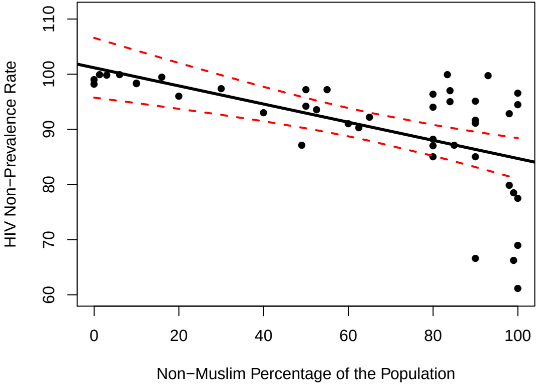
Scatterplot of Rescaled HIV/AIDS Rates on Muslim Population Percentage



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

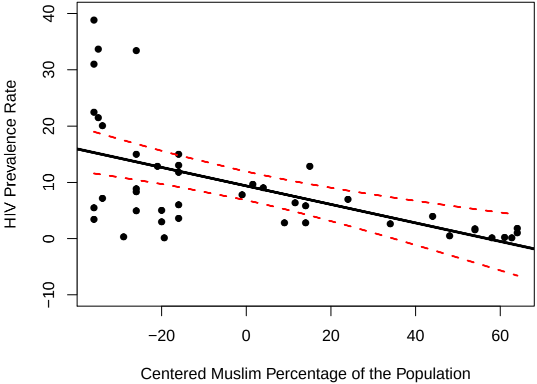


- Adding (subtracting) a positive constant to X shifts the X-axis to the left (right).
- Adding (subtracting) a positive constant to Y shifts the Y-axis downwards (upwards).
- Multiplying X (Y) times a positive constant greater than 1.0 stretches the X (Y) axis.
- Multiplying X (Y) times a positive constant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



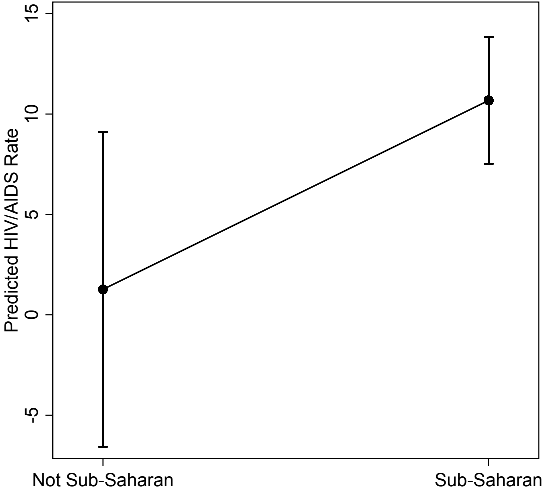
Use: Rescaling X for Interpretability

```
> 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.*