PLSC 503 – Spring 2024 Introduction + Bivariate Regression Review

January 8, 2024

"Multivariate Analysis for Political Research"

- "Regression" course

- Texts: Weisberg (2014) /

- Class materials: https://github.com/PrisonRodeo/PLSC503-2024-git

- Preceptor: Natham Morse

- Software: $R >$ Stata $>$ others

- Grading: Ten homework assignments (@ 50 points), plus a final project (500 points)

Will: "Regression":

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

Won't: Multivariate regression:

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

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

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

Won't: Classification:

- Cluster Analysis
- Classification and Regression Trees $\rightarrow$ Random Forests, etc.
- Pattern Recognition
- Machine Learning, Support Vector Machines, etc.

Regression

"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>...

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

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

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

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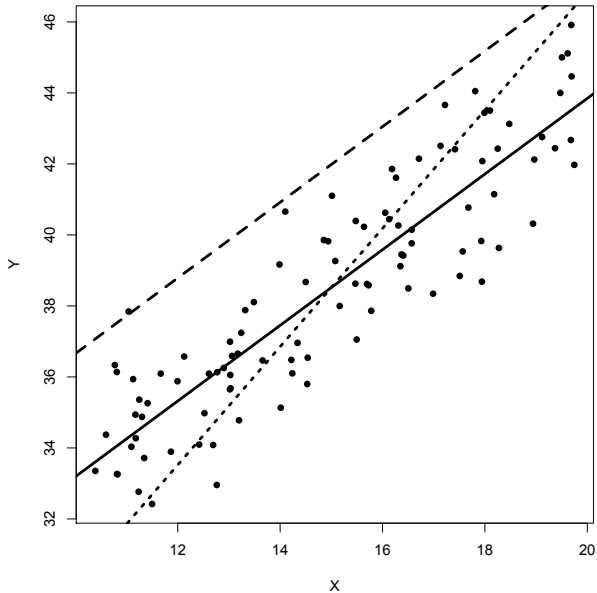
If we have $\hat{\beta}_0$ and $\hat{\beta}_1$, then:

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

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 $\hat{S} = \sum_{i=1}^{N} \hat{u}_i^2$.

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

Differentiate:

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

and

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

Yields:

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

and

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

Solving yields:

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

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

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (6)$$

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

TSS = **MSS** + **RSS**
("Total") ("Estimated," or "Model") ("Residual")

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
```
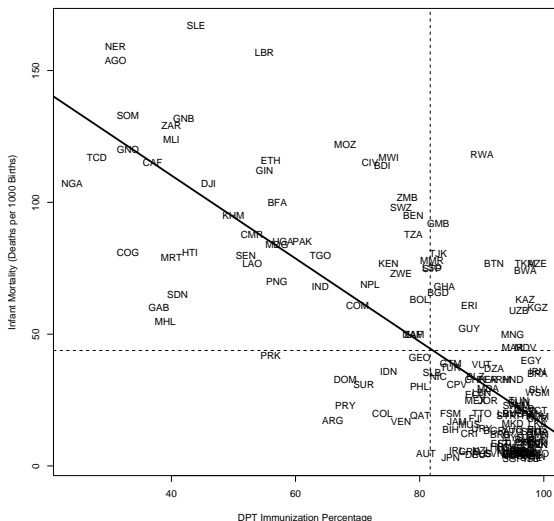
Scatterplot: Infant Mortality and DPT Immunization Rates

```
> 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
```

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

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

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

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

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

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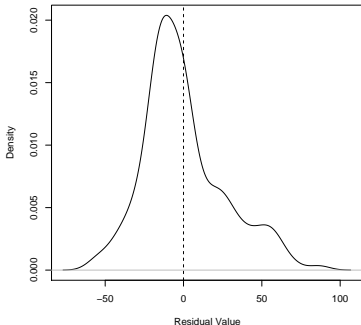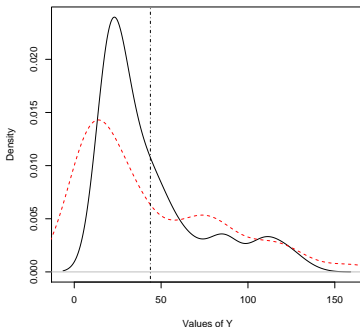
```
> # Fitted Values:
> Data$IMDPThat<-fitted.values(IMDPT)
> describe(Data$IMDPThat)

  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

Corr(Y, X):

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

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

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

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

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

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

```
> with(Data, cor(IMDPThat,infantmortalityperK,use="complete.obs"))
[1] 0.7632
```
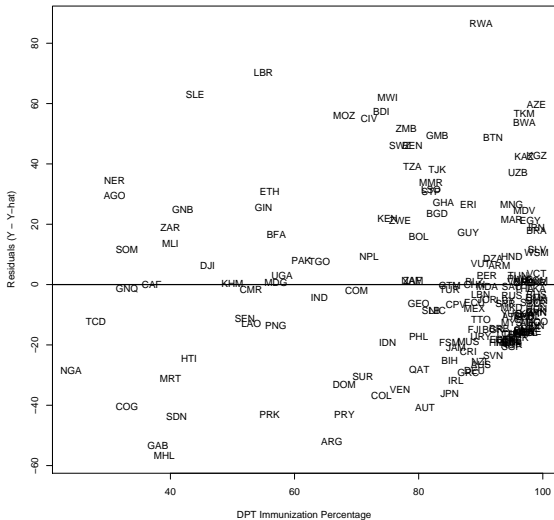
Corr($\hat{Y}, X$):

```
> with(Data, cor(IMDPThat,DPTpct,use="complete.obs"))
[1] -1
```
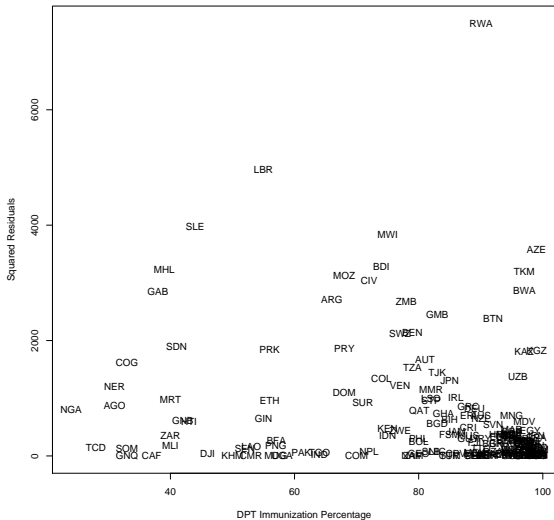
Corr($\hat{u}, \hat{Y}$):

```
> 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 $Y_i = \mu + u_i$.

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

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

meaning:

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

so:

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

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 $\text{Var}(\hat{\beta}_1) \propto -\sum(X_i - \bar{X})^2$

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

Gauss-Markov (continued)

Alternative (non-LS) estimator:

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

Unbiasedness requires:

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

Variance:

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

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

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

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:

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

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

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:

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

```
> 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
```

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

```
> vcov(IMDPT)

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

95 percent c.i.s:

```
> confint(IMDPT)

              2.5 %   97.5 %
(Intercept) 156.523  190.032
DPTpct       -1.775   -1.377
```
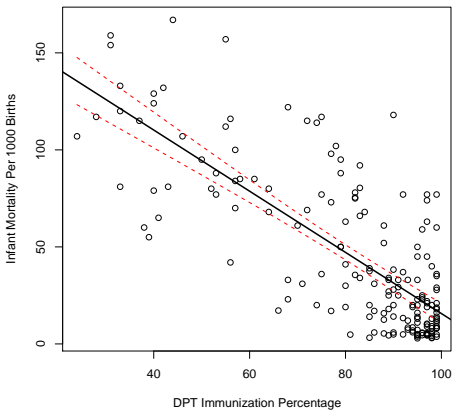
```
> SEs<-predict(IMDPT,interval="confidence")
> SEs
       fit   lwr   upr
1    25.10 20.53 29.68
3    17.22 12.05 22.40
4    23.53 18.84 28.21
.
.
.
<rows omitted>
.
.
189  21.95 17.15 26.75
190  39.29 35.36 43.23
191  17.22 12.05 22.40
```

A Plot, With Confidence Intervals

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

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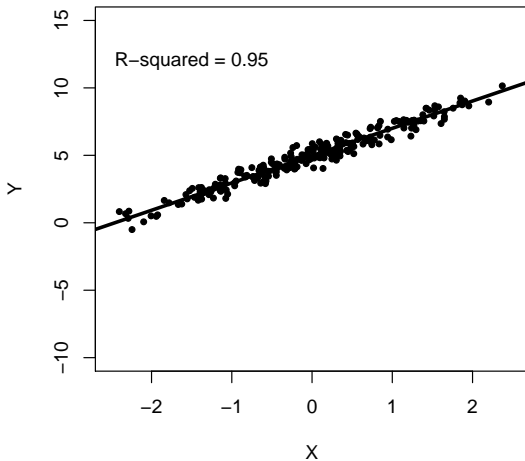
Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.95$)

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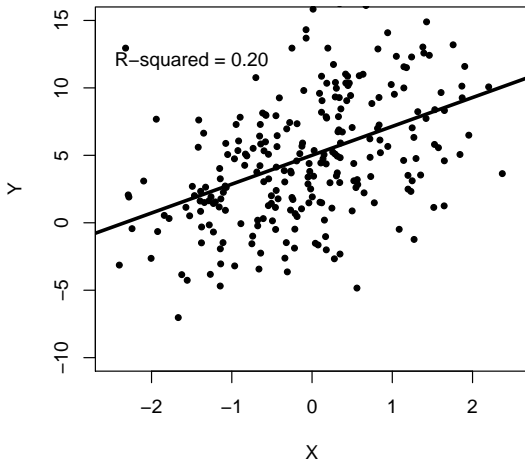
Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.20$)

R–squared = 0.20

Luskin: Population analogue "P^2":

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

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

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

and standard error:

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

Adjusted R^2

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

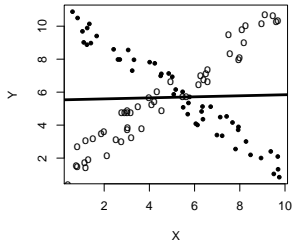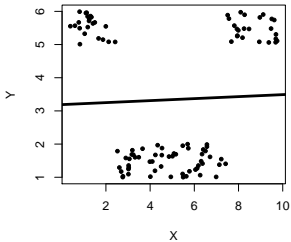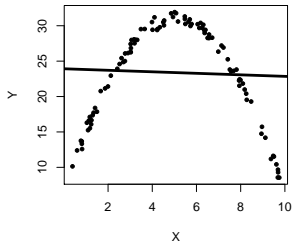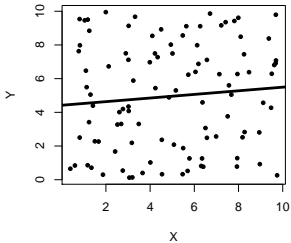where $c = 1$ if there is a constant in the model and $c = 0$ otherwise.

$R^2_{adj.}$:

- $R^2_{adj.} \to R^2$ as $N \to \infty$
- $R^2_{adj.}$ 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

- Graphical methods

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

The results:

```
> summary.lm(IMDPT)

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

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
  (14 observations deleted due to missingness)
Multiple R-squared:  0.582,Adjusted R-squared:  0.58
F-statistic:  244 on 1 and 175 DF,  p-value: <2e-16
```

A Better Table (using default-y `stargazer`)

Table: OLS Regression Model of Infant Mortality Rates, 2000

	Model I
(Constant)	173.30***
	(8.49)
DPT Immunization Rate	−1.58***
	(0.10)
Observations	177
R^2	0.58
Adjusted R^2	0.58
Residual Std. Error	26.19 (df = 175)
F Statistic	244.10*** (df = 1; 175)
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.*