PLSC 503 – Spring 2018
Bivariate Regression, I

January 25, 2018

Regression...

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

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

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 $\hat{\beta}_1$

Regression (continued)

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

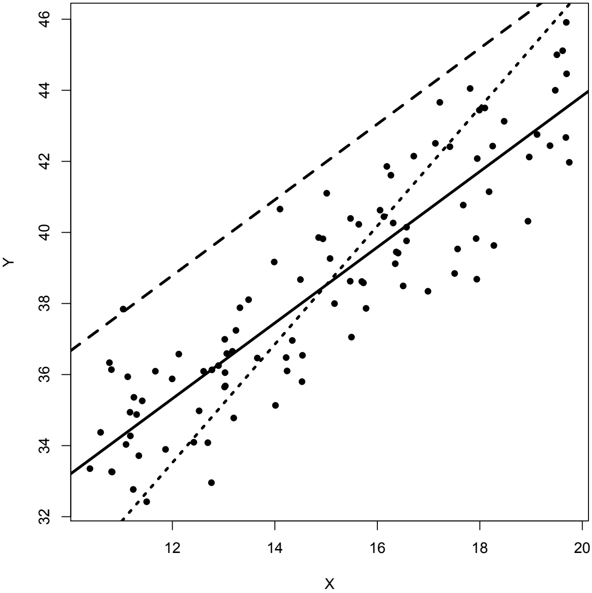
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3}$$

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)



Ordinary Least Squares

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

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

OLS (continued)

Solving yields:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \qquad (5)\\ &= \frac{\text{Covariance of } X \text{ and } Y}{\text{Variance of } X}\end{aligned}$$

and

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

Infant Mortality Data

```
> url <- getURL("https://raw.githubusercontent.com/PrisonRodeo/
    PLSC503-Spring-2016-git/master/Data/CountryData2000.csv")
> Data <- read.csv(text = url) # read the "votes" 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
```

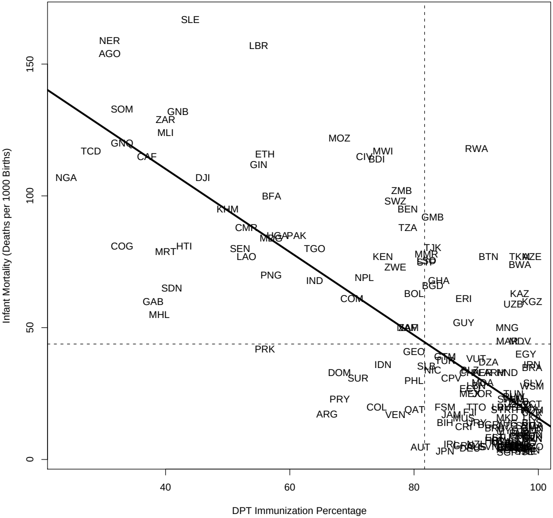
OLS Regression

```
> 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
```

Analysis of Variance

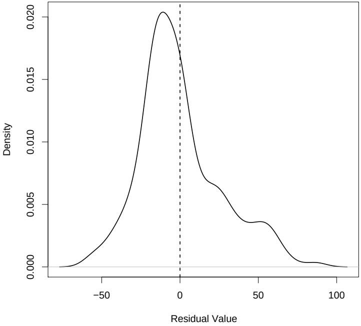
```
> anova(IMDPT)
Analysis of Variance Table

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.

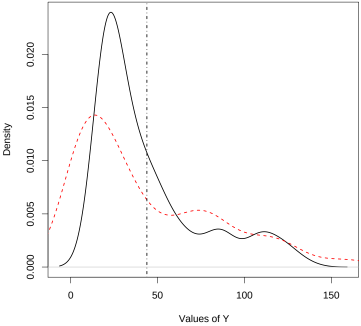
```
> # Residuals (u):
> Data$IMDPTres <- with(Data, residuals(IMDPT))
> describe(Data$IMDPTres)

  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

```
> # 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
```

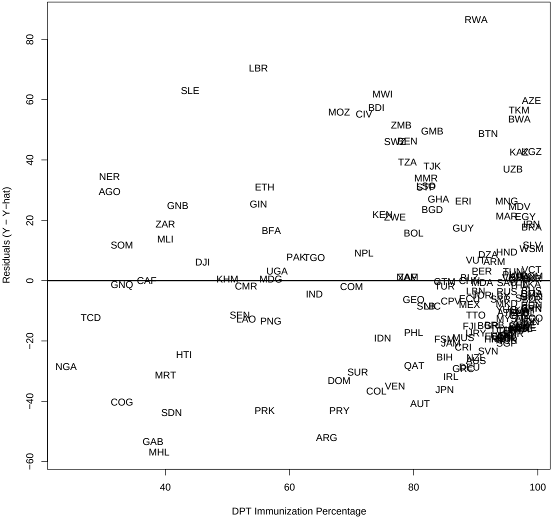
Figure: Density Plot: Actual (Y) and Fitted Values ($\hat{Y}$)



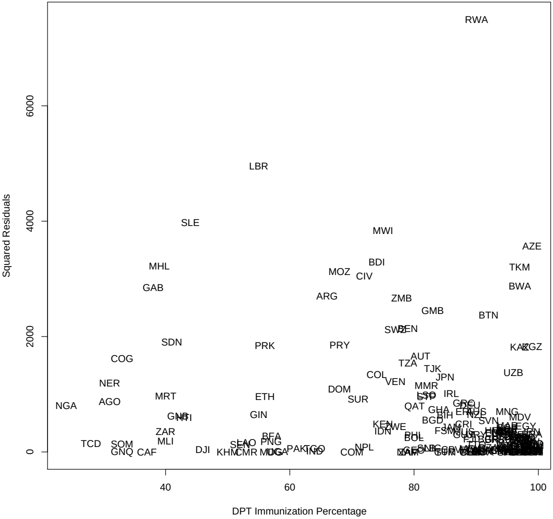
Some Correlations

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

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

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

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

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

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

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



Squared Residuals vs. DPT Percentage


Squared Residuals
DPT Immunization Percentage
RWA
LBR
SLE
MWI
AZE
BDI
MHL
MOZ
CIV
TKM
GAB
BWA
ARG
ZMB
GMB
BTN
SWZ
BEN
SDN
PRK
PRY
KAZ
KGZ
COG
AUT
TZA
TJK
UZB
JPN
COL
NER
VEN
MMR
DOM
LSO
STP
IRL
MRT
ETH
SUR
NGA
AGO
QAT
GRC
DEU
GHA
ERI
TUS
NZL
MNG
GNB
HTI
GIN
BIH
MDV
BGD
SVN
KEN
ZWE
CRI
EGY
IDN
JAM
FSM
ZAR
BFA
PHL
BOL
MLI
TCD
SOM
LAO
PNG
SEN
NPL
DJI
PAK
TGO
GNQ
CAF
KHM
CMR
MDG
UGA
IND
COM
ZAF
MAR
CPV