(BST210 Lab1)Simple Linear Regression: Basic Theory and Application

Note: this document shows solutions in `R`. Code for `STATA` and `SAS` solutions are available in `Lab1.do` and `Lab1.sas`, respectively.

With Simple Linear Regression, we are looking at models of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where the Y_i are our outcomes and the X_i are our explanatory variables (covariates).

1

What assumptions about our data and model do we make?

- The mean of Y is a linear function of X
- The variability of Y about its mean value is equal for all x values
- The distribution of Y about its mean is normally distributed
- All responses are independent

2

Read in 'lab1.csv' with your preferred software, do a one-side T test on X with $H_0 : \mu = 3.5$ and $H_1 : \mu > 3.5$. What is the P-value? Do we reject the null?

```r
# Read in lab1.csv
dat_lab1 = read.csv("lab1.csv",header = TRUE)
# Do one-side T test with alternative H1: mu > 3.5
t.test(dat_lab1$x, mu = 3.5, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  dat_lab1$x
## t = -5.6926, df = 299, p-value = 1
## alternative hypothesis: true mean is greater than 3.5
## 95 percent confidence interval:
##  2.653724      Inf
## sample estimates:
## mean of x
##  2.843894
```

p-value = 1, thus we do not reject the null hypothesis that $\mu = 3.5$.

3

Fit a regression model, what are the 95% CIs for β_0 and β_1?

```r
# Fit a regression model
lm_lab1 = lm(y ~ x, data = dat_lab1)
# Get the 95\% Confidence Intervals for beta
confint(lm_lab1)
```

```
##                2.5 %   97.5 %
## (Intercept) 1.660227 2.060428
## x           2.980157 3.095399
```

4

How to interpret the coefficients?

```r
summary(lm_lab1)
```

```
## 
## Call:
## lm(formula = y ~ x, data = dat_lab1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.8311 -0.6538  0.0051  0.6698  3.2273 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.86033    0.10168    18.3   <2e-16 ***
## x            3.03778    0.02928   103.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.011 on 298 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.973 
## F-statistic: 1.076e+04 on 1 and 298 DF,  p-value: < 2.2e-16
```

- The mean of Y is 1.86 when $X = 0$
- The average of Y increase 3.04 for every one unit increase in X

*Background of Least Squares Estimation

Notation

Let n denote the number of observations.

$\boldsymbol{Y} := (Y_1, \cdots, Y_n)^\top$.

$X := (X_1, \cdots, X_n)^\top$.

$\mathbf{1} = (1, \cdots, 1)^\top$.

$\boldsymbol{X} := (\mathbf{1}, X)$.

$\boldsymbol{\beta} := (\beta_0, \beta_1)^\top$

To minimize $(\boldsymbol{Y} - \boldsymbol{X\beta})^\top(\boldsymbol{Y} - \boldsymbol{X\beta})$, we will let the first derivative equal to 0.

$$
\frac{d(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -2\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = 0
$$
$$
\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}
$$

```
# Implement above estimation
X = cbind(1, dat_lab1$x)
Y = dat_lab1$y
beta_lab1 = solve(t(X)%*%X)%*%t(X)%*%Y
beta_lab1
```

```
##          [,1]
## [1,] 1.860328
## [2,] 3.037778
```

5

What is your prediction of Y if $X = 5$?

$$
\hat{Y}_{n+1} = 1.86033 + 3.03778 * 5 = 17.05.
$$

```
# Predict Y if X = 5
pred_5 = predict(lm_lab1, newdata = data.frame(x = 5))
pred_5
```

```
##        1
## 17.04922
```

```
round(pred_5, 2)
```

```
##     1
## 17.05
```

```
# For more information of predict
?predict
```