FTAP Homework 5

Zachary Fogelson

July 16, 2015

Problem 1

a

$$s_{fit} = s\sqrt(\frac{1}{n} + \frac{(X_f - \bar{X})}{(n-1)s_X^2})$$

$$Sd = 20.76$$

$$\bar{X} = 11.04966$$

$$n = 5336$$

$$S_x^2 = 7.817402$$

$$s_{fit} = 20.76 * \sqrt(\frac{1}{5336} + \frac{(10 - 11.0496)^2}{5335 * 7.817402})$$

$$s_{fit} = 0.303531$$

$$P(Y|X = 10) = 13.295$$

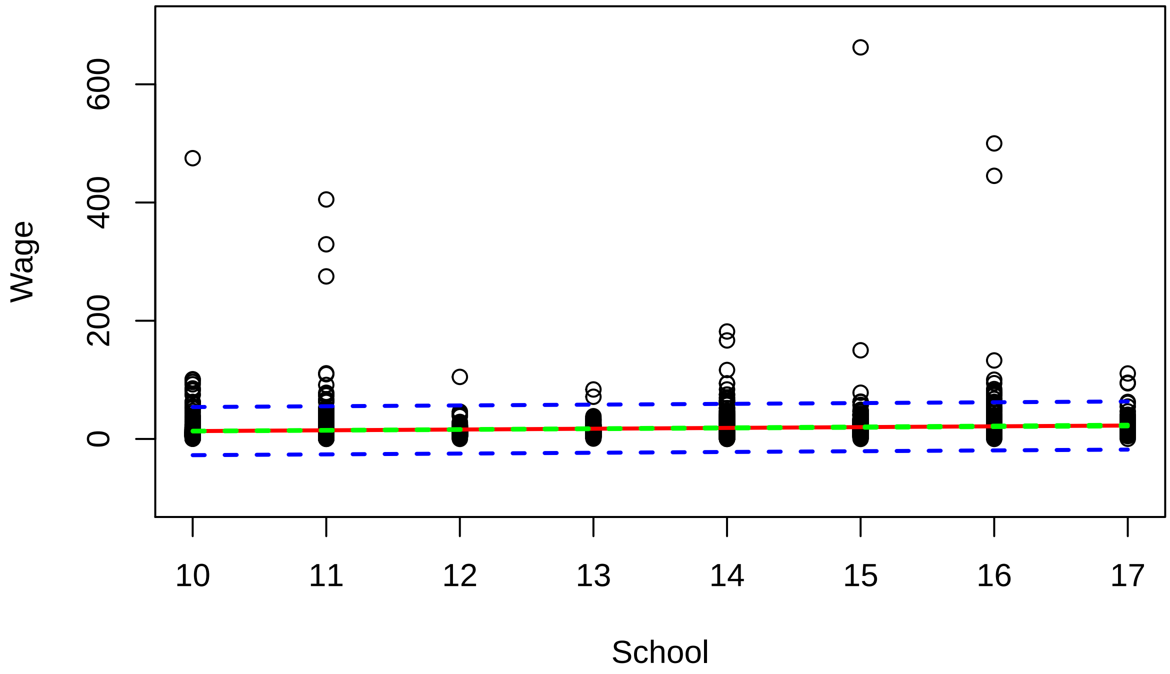
$$CI = 13.295 + / - 2 * 0.303531 = (12.7, 13.9)$$

```
wgs <- read.xls(xls = "censuswage.xls")
lm.out <- lm(Wage ~ School, wgs)
predict(lm.out,data.frame(School=c(10)), interval = "confidence")
```

```
##       fit      lwr      upr
## 1 13.2955 12.70028 13.89072
```

b

```
plot(Wage ~ School, wgs[wgs$School >= 10, ], ylim=c(-100, 700))
pred.int <- predict(lm.out, newdata = data.frame(School=c(10:17)), interval = "predict")
conf.int <- predict(lm.out, newdata = data.frame(School=c(10:17)), interval = "confidence")
lines(c(10:17), pred.int[,1], col="red", lwd = 2)
lines(c(10:17), pred.int[,2], col="blue",  type="l", lty=2, lwd = 2)
lines(c(10:17), pred.int[,3], col="blue",  type="l", lty=2, lwd = 2)
lines(c(10:17), conf.int[,2], col="green",  type="l", lty=2, lwd = 2)
lines(c(10:17), conf.int[,3], col="green",  type="l", lty=2, lwd = 2)
```



Problem 2

```r
ceo <- read.xls(xls = "ceosalary.xls")
lm.out <- lm(salary ~ comten + ceoten + sales, ceo)
print(lm.out)  # Point Estimates
```

```
##
## Call:
## lm(formula = salary ~ comten + ceoten + sales, data = ceo)
##
## Coefficients:
## (Intercept)       comten       ceoten        sales
##   674.17896     -3.05712     15.62693      0.03858
```

```r
summary(lm.out)$coefficients[,"Std. Error"] # Std Errors (You can also see this from sqrt(diag(vcov(lm.
```

```
##  (Intercept)       comten       ceoten        sales
## 89.434836362  3.504800534  6.006818282  0.006732074
```

```r
confint(lm.out)  # Confidence intervals
```

```
##                     2.5 %       97.5 %
## (Intercept) 497.65504252 850.70287558
## comten       -9.97479541   3.86055426
## ceoten        3.77084090  27.48301240
## sales         0.02529341   0.05186856
```

```
summary(lm.out)$coefficients[,"Pr(>|t|)"]
```

```
##  (Intercept)       comten       ceoten        sales
## 2.562366e-12 3.842719e-01 1.008491e-02 4.351557e-08
```

a

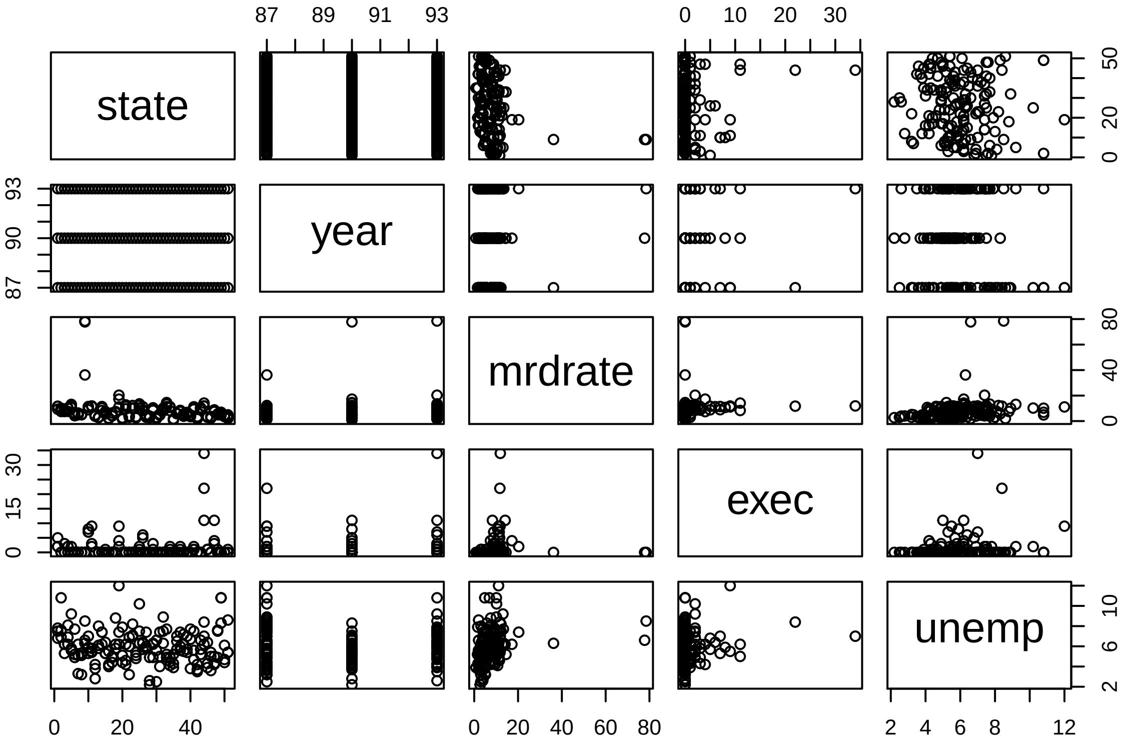
Because the confidence interval for the slope coefficient of company tenure on ceo salary includes 0. We could hypothesise that company tenure is unrelated to ceo salary. However, company tenure is probably highly correlated with ceo tenure. Therefore, the large covariance between the two variables may be causing the counter intuative relationship between company tenure and salary.

b

Because the confidence interval for sales does not include 0 and because the p-value for the statistical significance for sales as a determininent of sales is far below the 99% confidence level we can be confident that theory two is incorrect.

Probelm 3

```
redrum <- read.xls(xls = "murder.xls")
plot(redrum)
```



a

```
redrum$CapPun <- as.numeric(redrum$exec > 0)
lm.out <- lm(mrdrate ~ unemp + CapPun, redrum)
lm.out
```

```
##
## Call:
## lm(formula = mrdrate ~ unemp + CapPun, data = redrum)
##
## Coefficients:
## (Intercept)        unemp       CapPun
##       0.116        1.253        1.753
```

b

The coefficient on the capital punishment variable means that for a given level of unemployment, whether a community has capital punishment or not is associated with 1.753% higher murder rate.

c

```
summary(lm.out)
```

```
##
## Call:
## lm(formula = mrdrate ~ unemp + CapPun, data = redrum)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.892 -3.549 -1.159  0.933 69.414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1160     2.6812   0.043  0.96554
## unemp         1.2531     0.4357   2.876  0.00462 **
## CapPun        1.7530     1.6480   1.064  0.28917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.949 on 150 degrees of freedom
## Multiple R-squared:  0.06479,    Adjusted R-squared:  0.05232
## F-statistic: 5.196 on 2 and 150 DF,  p-value: 0.00658
```

Null Hypothesis: $H_0 : \beta_2 = 0$

Alternative: $H_1 : \beta_2 \neq 0$

Because the p-value for β_2 is >5% we fail to reject the null hypothesis that whether or not a community has capital punishment influences the murder rate.

Controlling for the effects of unemployment, the presence of capital punishment we cannot say that capital punishment is correlated to the murder rate.

d

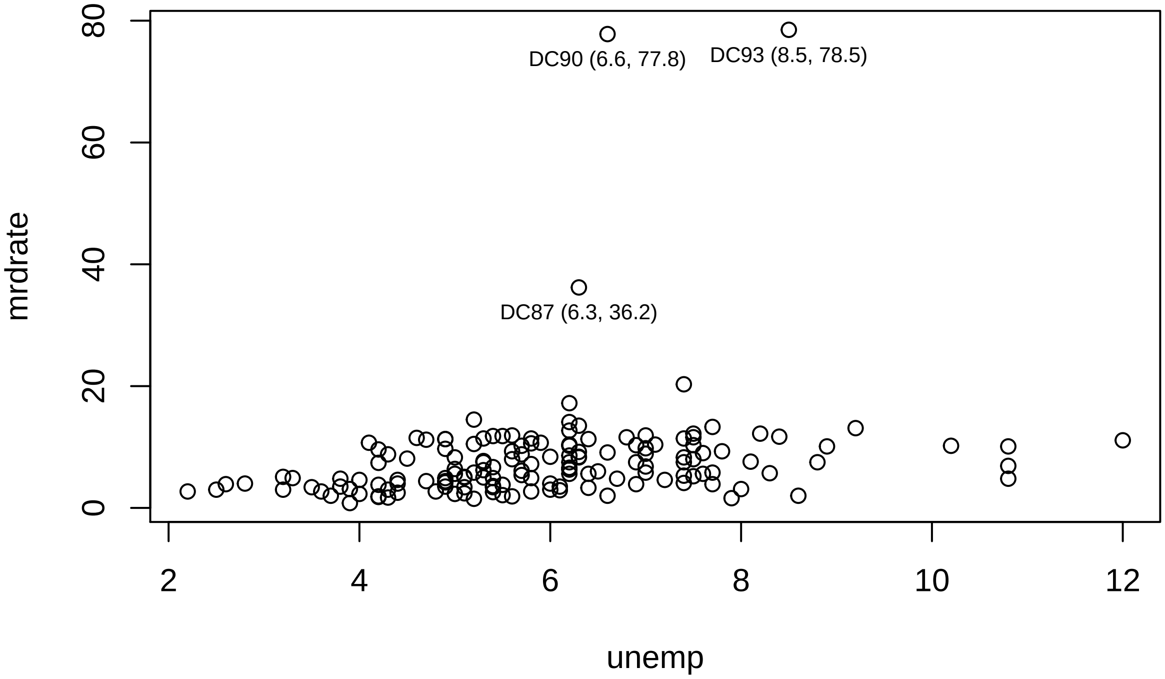
```
predict(lm.out, newdata = data.frame(unemp=c(6,6), CapPun=c(1,0)), interval = "predict")
```

```
##        fit        lwr      upr
## 1 9.387374  -8.512036 27.28678
## 2 7.634422 -10.127569 25.39641
```

These intervals are very large! Both intervals include 0 also both intervals include values with are both >2x and <x/2 the fit (spot estimate) value.

e

```
plot(mrdrate ~ unemp, redrum)
dcUnemp <- redrum[redrum$state == 9,]$unemp
dcMrdrate <- redrum[redrum$state == 9,]$mrdrate
text(x = dcUnemp, y = dcMrdrate, labels = c("DC87 (6.3, 36.2)", "DC90 (6.6, 77.8)", "DC93 (8.5, 78.5)")
```



```
redrumNoDC <- redrum[redrum$state != 9,]
lm.out <- lm(mrdrate ~ unemp + CapPun, redrumNoDC)
lm.out
```

```
##
## Call:
## lm(formula = mrdrate ~ unemp + CapPun, data = redrumNoDC)
##
## Coefficients:
## (Intercept)        unemp       CapPun
##      2.1320       0.6443       3.5957
```

Based on the new regression which excludes Washington D.C. for an given level of unemployment, whether or not a state has capital punishment is associated with a 3.6% higher murder rate.

```
summary(lm.out)
```

```
##
## Call:
## lm(formula = mrdrate ~ unemp + CapPun, data = redrumNoDC)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.5515 -2.0081 -0.3782  1.4475  9.8042 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   2.1320     0.9315   2.289   0.0235 *  
## unemp         0.6443     0.1522   4.235 4.01e-05 ***
## CapPun        3.5957     0.5729   6.276 3.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.095 on 147 degrees of freedom
## Multiple R-squared:  0.3106, Adjusted R-squared:  0.3012 
## F-statistic: 33.11 on 2 and 147 DF,  p-value: 1.344e-12
```

Null Hypothesis: $H_0 : \beta_2 = 0$

Alternative: $H_1 : \beta_2 \neq 0$

Because the p-value for the β_2 coefficient is less than .001% we are able to be more than 99% confident in rejecting the null hypothesis.

Discussion:

We know that the t-statistic for multiple varaiable regressions depend the standard error of the coefficient we are examining.

$$t = \frac{b_j - \beta_j}{s_{b_j}}$$

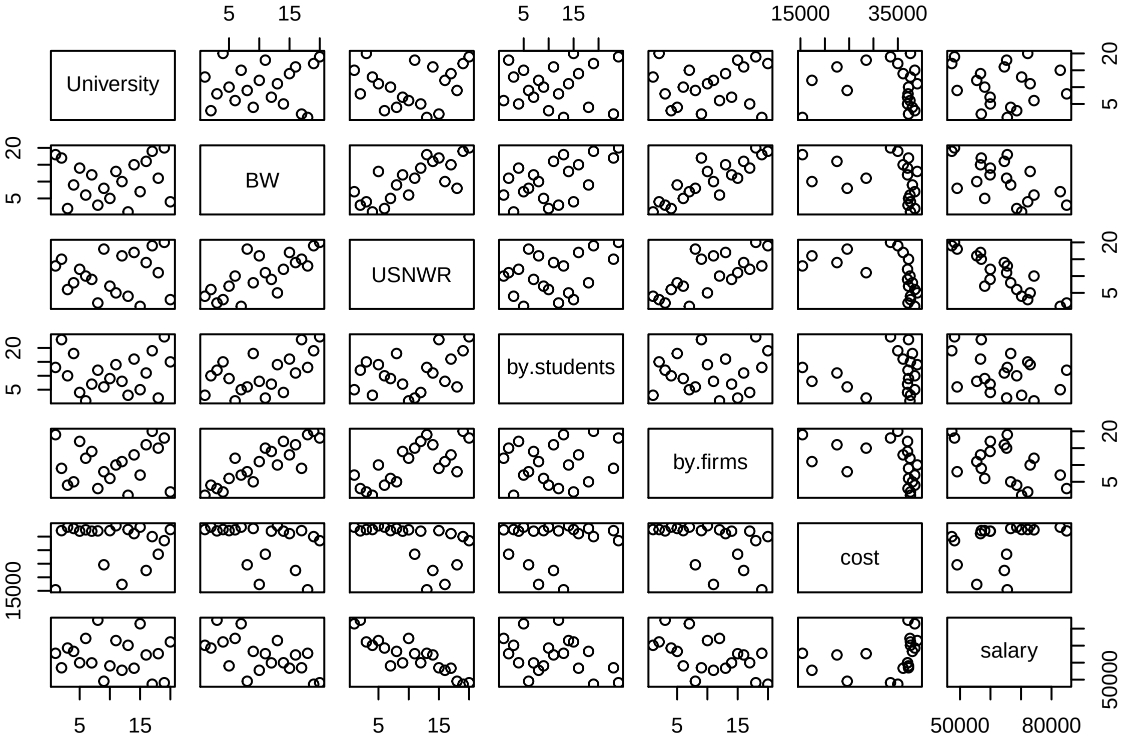
Furthermore, we know that the standard error for a given β depends on σ^2.

(https://stats.stackexchange.com/questions/44838/how-are-the-standard-errors-of-coefficients-calculated-in-a-regression)

Therefore, by removing the observations for DC which have extremely large residuals we have dramatically reduced the observed variance in σ^2 and, we manufacture much smaller t-values.

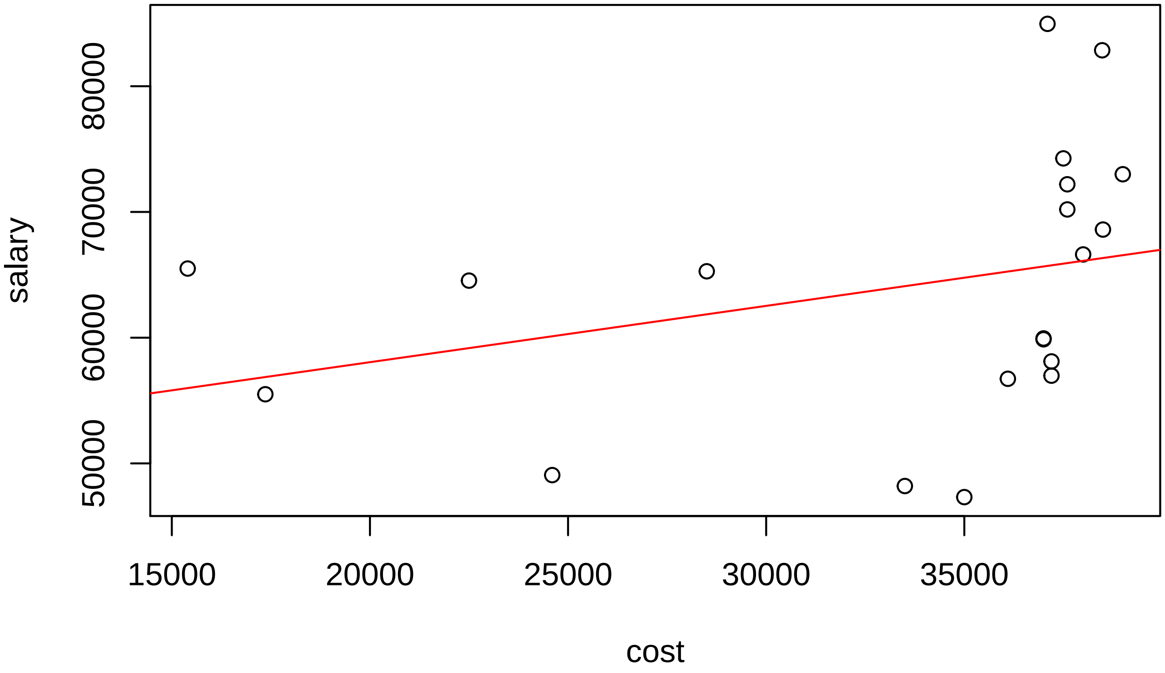
Problem 4

```r
rank <- read.xls(xls = "rank.xls")
plot(rank)
```



a

```
lm.out <- lm(salary ~ cost, rank)
plot(salary ~ cost, rank)
abline(lm.out, col = "red")
```



```
summary(lm.out)
```

```
##
```

```
## Call:
## lm(formula = salary ~ cost, data = rank)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17448  -7865   1387   6311  19252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.909e+04  1.088e+04   4.514 0.000268 ***
## cost        4.478e-01  3.196e-01   1.401 0.178176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10290 on 18 degrees of freedom
## Multiple R-squared:  0.09834,    Adjusted R-squared:  0.04825
## F-statistic: 1.963 on 1 and 18 DF,  p-value: 0.1782
```

```r
cor(rank$salary, rank$cost)
```

```
## [1] 0.3135941
```

The correlation between cost and entry salary is $\approx .314$.

b

Based on the scatter plot of salary versus cost shown above, business school costs are clearly left skewed. There are a large number of schools whose tuition are between 35 and 40 thousand dollars a semester but there individual schools which charge as little as 15 thousand a semester.

c rent According to the regression, for each additional dollar we spend to go to business school we will make an addition $.32 for our future entry salary. Based on the slope estimate alone I don't believe that anything can be concluded, we do not know whether the slope is significant. It is possible that we would get this slope even through there is not a linear relationship between cost and entry salary.

d

Yes, I believe we need to be very careful in interpretting these results because the scatter plot looks highly non-linear.