HW5, SOLUTIONS

Problem #1

First, for *sales* onto *radio*:

```
Advertising <- read.csv("~/Downloads/Advertising.csv")
lm.obj <- lm(sales ~ radio, data = Advertising)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = sales ~ radio, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

1. Yes, due to a tiny p-value of ≈ 0, there is a statistically significant relationship beteen radio advertisement budget and sales.

2. Code below:

```
confint(lm.obj)
```

```
##                 2.5 %     97.5 %
## (Intercept) 8.2015885 10.4216877
## radio       0.1622443  0.2427472
```

Intercept: With 95% confidence, for markets with 0$ invested into radio advertisement, we expect, on average, between 8201 and 10421 items sold.

Slope: With 95% confidence, per $1,000$$ increase in radio ad budget, we expect to sell between 162 to 242 items more, on average.

3. a. Calculating the single prediction:

```
predict(lm.obj, newdata=data.frame(radio=20))
```

```
##        1
## 13.36155
```

Interpreting: For markets with $20,000$$ radio budget, we expect to sell 13361 items, on average.

b. Calculating the 95% confidence bands:

```
predict(lm.obj, newdata=data.frame(radio=20), int = "c")
```

```
##        fit      lwr      upr
## 1 13.36155 12.75114 13.97197
```

Interpreting: With 95% confidence, for markets with $20,000$$ radio budget, we expect the **average sales** to be between 12751 and 13972 items.

c. Calculating the 95% prediction bands:

```
predict(lm.obj, newdata=data.frame(radio=20), int = "p")
```

```
##        fit      lwr      upr
## 1 13.36155 4.909218 21.81389
```

Interpreting: For 95% of all markets with $20,000$$ radio budget, we expect the **individual sales** to be between 4909 and 21814 items.

Prediction bands are wider because they're trying to capture most of **individual response values** (along with their individual variability) rather than just their averages (which is what confidence bands do).

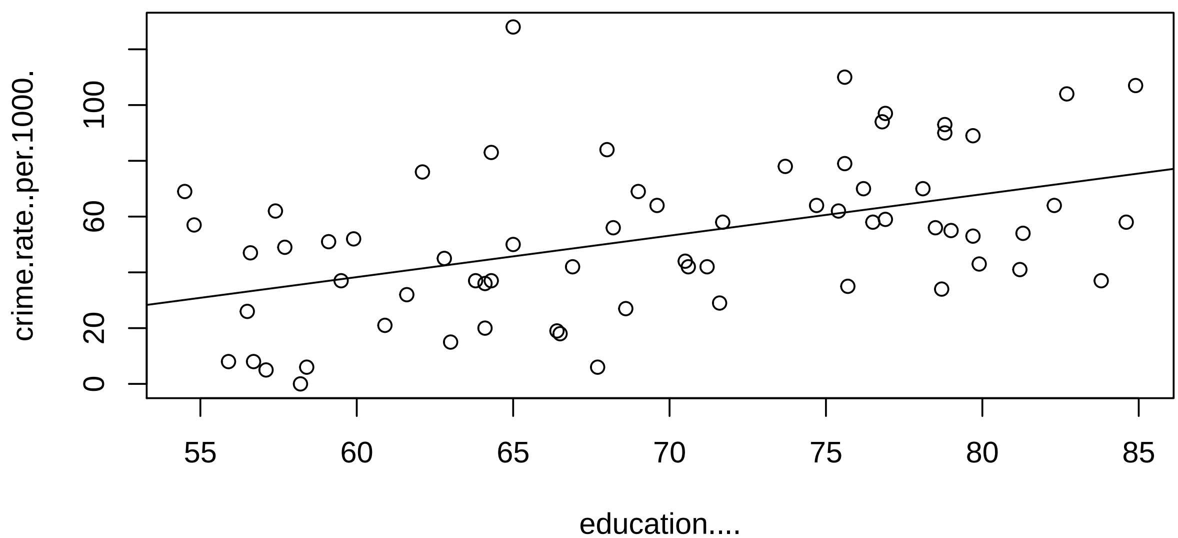
Problem #2

1. a.

$$crime_i = \beta_0 + \beta_1 education_i, \ \epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$$

b. Code below:

```
fl_crime <- read.csv("~/Downloads/fl_crime.csv")
attach(fl_crime)
lm.obj <- lm(crime.rate..per.1000. ~ education....,
             data=fl_crime)
plot(crime.rate..per.1000. ~ education....)
abline(lm.obj)
```




```r
summary(lm.obj)
```

```
##
## Call:
## lm(formula = crime.rate..per.1000. ~ education...., data = fl_crime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.74 -21.36  -4.82  17.42  82.27
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -50.8569    24.4507  -2.080   0.0415 *
## education....   1.4860     0.3491   4.257 6.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.12 on 65 degrees of freedom
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.206
## F-statistic: 18.12 on 1 and 65 DF,  p-value: 6.806e-05
```

Fitted equation:

$$\widehat{crime} = -50 + 1.48 \times education$$

c. Yes, there's a statistically significant relationship between education and crime due to tiny p-value of $6.81e^{-05}$. Per 1-unit increase in education, we expect, on average, a 1.48-unit increase in crime.

2. a.

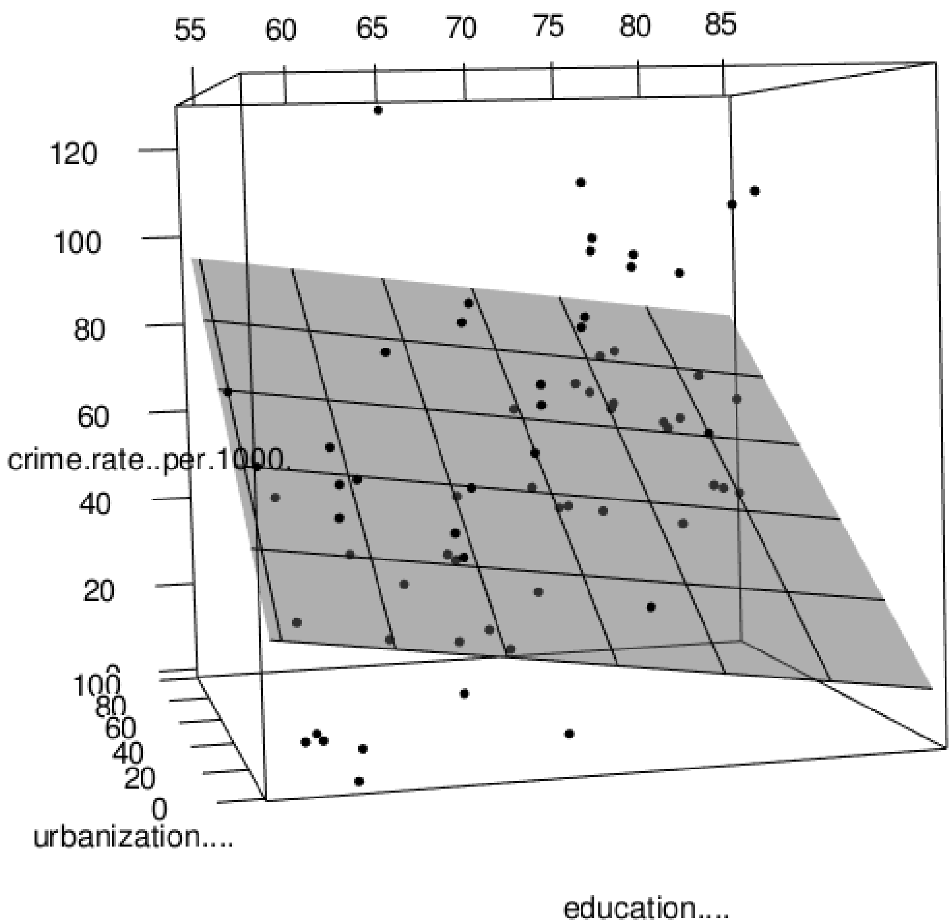
$$crime_i = \beta_0 + \beta_1 education_i + \beta_2 urbanization_i,\ \epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$$

b. Code below:

```
lm2.obj <- lm(crime.rate..per.1000. ~ education.... + urbanization....,
              data=fl_crime)
lm2.obj
```

```
##
## Call:
## lm(formula = crime.rate..per.1000. ~ education.... + urbanization....,
##     data = fl_crime)
##
## Coefficients:
##      (Intercept)     education....  urbanization....
##          59.1181           -0.5834            0.6825
```

```
library(rgl)
plot3d(lm2.obj, size=5)
```



Fitted equation:

$$\widehat{crime} = 59.11 - 0.58 \times education + 0.68 \times urbanization$$

c. Per 1-unit increase in education, **holding** *urbanization* **constant**, we expect, on average, a 0.58-unit decrease in crime. The direction of the relationship changed due to us **controlling for urbanization** this time, hence looking at the effect of education on crime for cities with the **same urbanization level** (comparing "apples to apples", so to speak).

d. Intercept interpretation: For cities with 0 education and 0 urbanization level, the crime rate will be 59.11, on average. Technically, it's not impossible to have such cities, and the crime rate value doesn't look fully unreasonable (e.g. it's not negative), so this interpretation sort of makes sense.

3. a.

$$crime_i = \beta_0 + \beta_1 education_i + \beta_2 urbanization_i + \beta_3 income_i, \; \epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$$

b. Code below:

```
fl_crime <- read.csv("~/Downloads/fl_crime.csv")
lm3.obj <- lm(crime.rate..per.1000. ~ education.... + urbanization.... + income..median..in.1000.,
             data=fl_crime)
summary(lm3.obj)
```

```
##
## Call:
## lm(formula = crime.rate..per.1000. ~ education.... + urbanization.... +
##     income..median..in.1000., data = fl_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.407 -15.080  -6.588  16.178  50.125
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               59.7147    28.5895   2.089   0.0408 *
## education....             -0.4673     0.5544  -0.843   0.4025
## urbanization....           0.6972     0.1291   5.399 1.08e-06 ***
## income..median..in.1000.  -0.3831     0.9405  -0.407   0.6852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.95 on 63 degrees of freedom
## Multiple R-squared:  0.4728, Adjusted R-squared:  0.4477
## F-statistic: 18.83 on 3 and 63 DF,  p-value: 7.823e-09
```

$$\widehat{crime} = 59.11 - 0.46 \times education + 0.69 \times urbanization - 0.38 \times income$$

c. For *education*: $H_0 : \beta_1 = 0, \;\; vs \;\; H_a : \beta_1 \neq 0$, we fail to reject H_0 due to p-value of 0.40. Hence, no statistically significant relationship.

For *urbanization*: $H_0 : \beta_2 = 0, \;\; vs \;\; H_a : \beta_2 \neq 0$, we reject H_0 due to p-value of ≈ 0. Hence, an evidence of statistically significant relationship.

For *income*: $H_0 : \beta_3 = 0, \;\; vs \;\; H_a : \beta_3 \neq 0$, we fail to reject H_0 due to p-value of 0.68. Hence, no statistically significant relationship.

d. $\hat{\beta}_2 = 0.69$: Per 1-unit increase in urbanization, **holding** *education* **and** *income* **constant**, the crime will decrease by 0.69 units, **on average**.

e.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, \quad vs \quad H_a : \{ \text{ at least one } \beta_j \neq 0\}$$

We reject H_0 due to tiny p-value of F-test ($7.823e^{-09}$), concluding model significance.

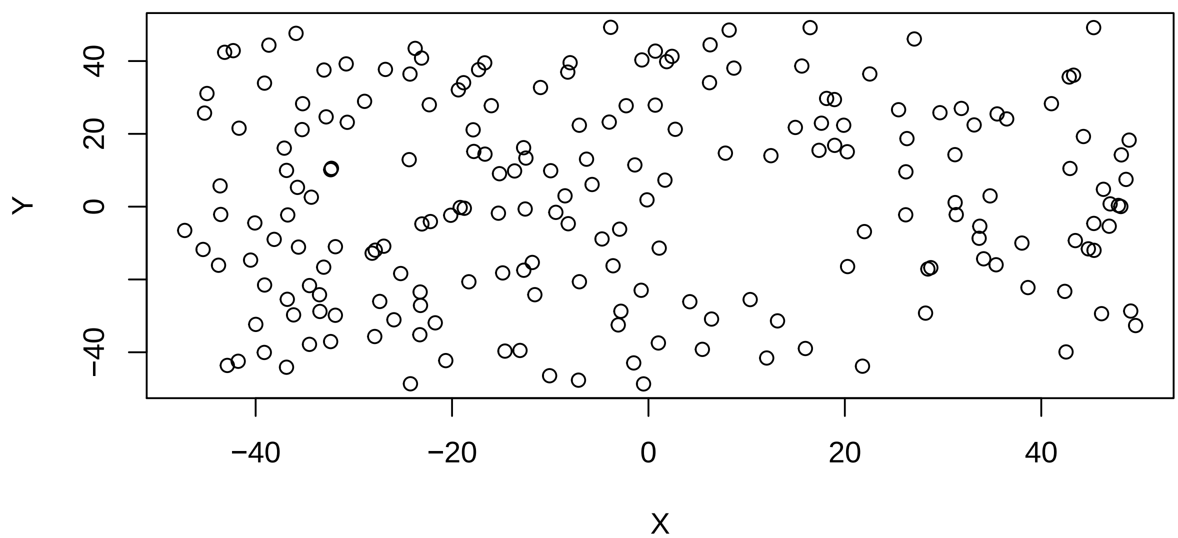
Problem #3 (Why need F-statistic?)

1. Code below:

```
set.seed(1)

# Generating response and predictors in independent fashion.
Y <- runif(200, -50,50)
X <- runif(200, -50,50)

# Basic scatterplot demonstrates lack of relationship between y & x.
plot(Y~X)
```



Clearly, no discernible pattern of a relationship between X and Y, reflecting the fact that they were generated in an independent fashion.

2. Code below:

```
set.seed(1)

n.var <- 50
```

```r
# Generating response and predictors in independent fashion.
Y <- runif(200, -50,50)
X.vars <- matrix(runif(200*n.var, -50,50), ncol=n.var)
```

3. Code below:

```r
lm.obj <- lm(Y ~ X.vars)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = Y ~ X.vars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.90 -19.11  -2.64  17.39  55.33
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.728363   2.271622   1.201   0.2316
## X.vars1      0.107986   0.077077   1.401   0.1633
## X.vars2     -0.008971   0.077798  -0.115   0.9084
## X.vars3      0.072561   0.078807   0.921   0.3587
## X.vars4      0.065204   0.070916   0.919   0.3593
## X.vars5     -0.096428   0.077072  -1.251   0.2128
## X.vars6      0.087861   0.073906   1.189   0.2364
## X.vars7     -0.070926   0.073007  -0.971   0.3329
## X.vars8     -0.051138   0.078321  -0.653   0.5148
## X.vars9      0.116349   0.078287   1.486   0.1393
## X.vars10    -0.117965   0.073597  -1.603   0.1111
## X.vars11     0.023375   0.079396   0.294   0.7689
## X.vars12    -0.020238   0.082243  -0.246   0.8060
## X.vars13     0.104101   0.083755   1.243   0.2159
## X.vars14     0.040169   0.078655   0.511   0.6103
## X.vars15     0.052605   0.075999   0.692   0.4899
## X.vars16     0.014496   0.075069   0.193   0.8471
## X.vars17    -0.066333   0.073135  -0.907   0.3659
## X.vars18    -0.076406   0.076835  -0.994   0.3216
## X.vars19    -0.067140   0.072523  -0.926   0.3561
## X.vars20     0.054361   0.076691   0.709   0.4795
## X.vars21    -0.159874   0.075751  -2.111   0.0365 *
## X.vars22    -0.006714   0.076880  -0.087   0.9305
## X.vars23     0.050658   0.084272   0.601   0.5487
## X.vars24     0.015815   0.076443   0.207   0.8364
## X.vars25    -0.103205   0.082574  -1.250   0.2133
## X.vars26    -0.040809   0.082953  -0.492   0.6235
## X.vars27     0.022297   0.080046   0.279   0.7810
## X.vars28    -0.084455   0.076854  -1.099   0.2736
## X.vars29    -0.023868   0.077868  -0.307   0.7596
## X.vars30    -0.014506   0.075581  -0.192   0.8481
## X.vars31    -0.093968   0.079945  -1.175   0.2417
## X.vars32     0.100147   0.079457   1.260   0.2095
```

```
## X.vars33     0.117234   0.082753   1.417   0.1587
## X.vars34     0.083164   0.078635   1.058   0.2920
## X.vars35    -0.050134   0.076052  -0.659   0.5108
## X.vars36     0.016324   0.075806   0.215   0.8298
## X.vars37     0.054042   0.079170   0.683   0.4959
## X.vars38     0.011276   0.071519   0.158   0.8749
## X.vars39     0.105300   0.089664   1.174   0.2421
## X.vars40    -0.021386   0.073686  -0.290   0.7720
## X.vars41    -0.071170   0.078616  -0.905   0.3668
## X.vars42     0.017666   0.075718   0.233   0.8158
## X.vars43     0.044137   0.080109   0.551   0.5825
## X.vars44     0.029598   0.079206   0.374   0.7092
## X.vars45    -0.047734   0.075744  -0.630   0.5295
## X.vars46    -0.122470   0.079698  -1.537   0.1265
## X.vars47     0.064775   0.076588   0.846   0.3990
## X.vars48    -0.005178   0.076485  -0.068   0.9461
## X.vars49    -0.003645   0.075483  -0.048   0.9616
## X.vars50    -0.048246   0.077576  -0.622   0.5349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.52 on 149 degrees of freedom
## Multiple R-squared:  0.2167, Adjusted R-squared:  -0.04614
## F-statistic: 0.8245 on 50 and 149 DF,  p-value: 0.7825
```

a. There was one t test significant at $\alpha = 0.05$ value. Rejecting H_0 was an incorrect decision, as we know that there is no actual relationship between Y and any of generated $X_1, X_2, \ldots, X_{50}$ variables. Hence, rejecting H_0 leads to us **falsely concluding** that there's a relationship between Y and X_j. Those are Type I errors, because

$$\text{Type I error} = (\text{Reject } H_0 \mid H_0 \text{ is true})$$

b. Instead, we need to conduct ***F*-test**, which deals with the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{50} = 0, \quad vs \quad H_a : \{\text{at least one } \beta_j \neq 0\}$$

F-test (last line of $summary()$ output) was succesfully able to recognize that there not a single predictor that is significantly related to the response. It showed that model is not significant as a whole (p-value of 0.78), hence we fail to reject H_0 of all β_j's $= 0$.