Statistics Assignment #7

Naeem Chowdhury

6/16/2020

##1. Setup ### options Set up global options

libraries

Load in needed libraries

2. File management

Create variables for directories

3. Importing Data

Problem #1

Part 1

For the *FL_crime.csv* data, proceed to fit

1. For simple linear regression $crime \sim education$,

```
a. Write down the __full modeling equation__, with all __error assumptions__.
b. Fit the model, provide the __fitted equation__. Provide a plot of the fitted line. Is there a statis
```

1a.

$$crime = \beta_0 + \beta_1 \cdot education + \epsilon, \quad \epsilon \sim_{iid} N(0, \sigma^2).$$

1b.

```r
lm.obj <- lm(crime ~ education, fl_crime)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = crime ~ 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
```

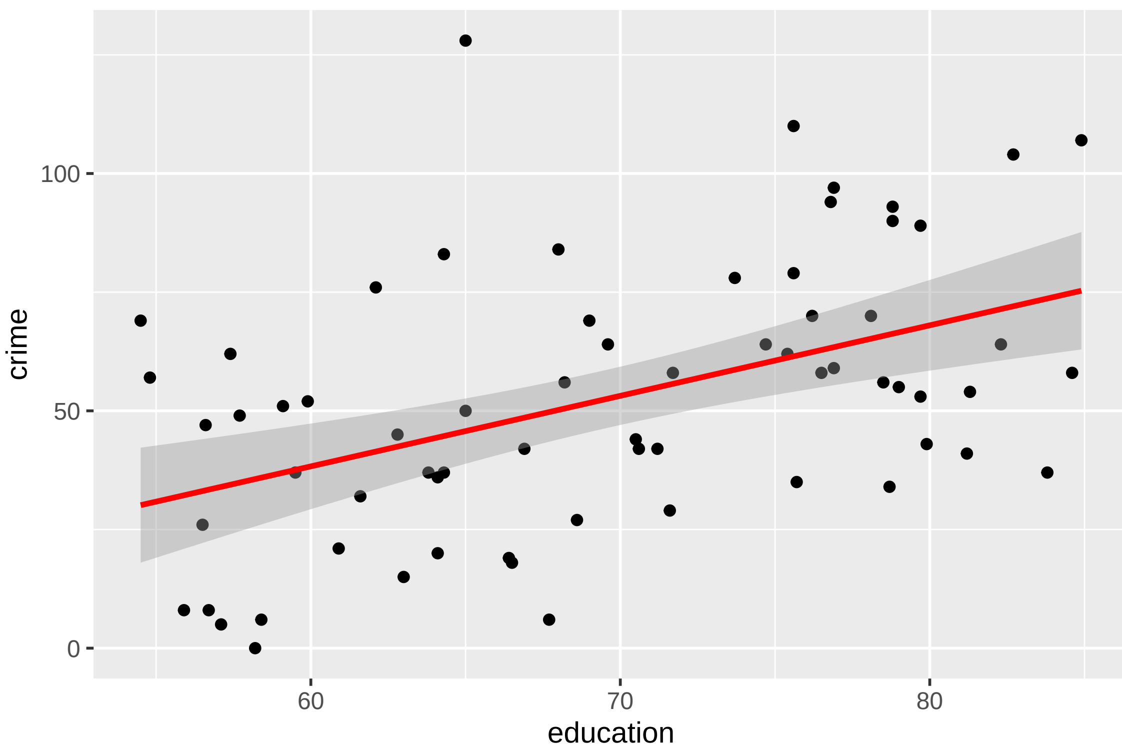
The fitted equation is thus,

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i},$$

where $\hat{y_i}$ *crime*, x_1 is *education*, $\hat{\beta}_0$ is -50.86, and $\hat{\beta}_1$ is 1.49.

$$crime_i = -50.86 + 1.49 \cdot education_i$$

```r
ggplot(fl_crime, aes(x = education, y = crime)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")
```



There is indeed a statistically significant relationship, since for confidence level 90% we have $p \approx 0.000068 < 0.05$ for the slope of the linear regression. For every 2% increase in education, we can expect that the number of crimes per 1000 will increase by about 3, on average.

Part 2

2. For multiple linear regression *crime* ~ *education* + *urbanization*, a. Write down the **full modeling equation**, with all **error assumptions**. b. Fit the model, provide the **fitted equation**. Provide a plot of the **fitted plane**. Describe the relationship between crime and education now. Why did it change compared to part 1? What statistical phenomena did we encounter in part 1 that led to such non-sensical interpretation?

2a.

$$crime = \beta_0 + \beta_1 \cdot education + \beta_2 \cdot urbanization + \epsilon, \quad \epsilon \sim_{iid} N(0, \sigma^2).$$

2b.

```r
lm.obj <- lm(crime ~ education + urbanization, fl_crime)

summary(lm.obj)
```

```
##
## Call:
## lm(formula = crime ~ education + urbanization, data = fl_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.693 -15.742  -6.226  15.812  50.678
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.1181    28.3653   2.084   0.0411 *
## education     -0.5834     0.4725  -1.235   0.2214
## urbanization   0.6825     0.1232   5.539 6.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.82 on 64 degrees of freedom
## Multiple R-squared:  0.4714, Adjusted R-squared:  0.4549
## F-statistic: 28.54 on 2 and 64 DF,  p-value: 1.379e-09
```

The fitted equation is thus,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i},$$

where $\hat{y}_i$ *crime*, x_1 is *education*, x_2 is *urbanization*, $\hat{\beta}_0$ is 59.12, $\hat{\beta}_1$ is -0.58, and $\hat{\beta}_2$ is 0.68.

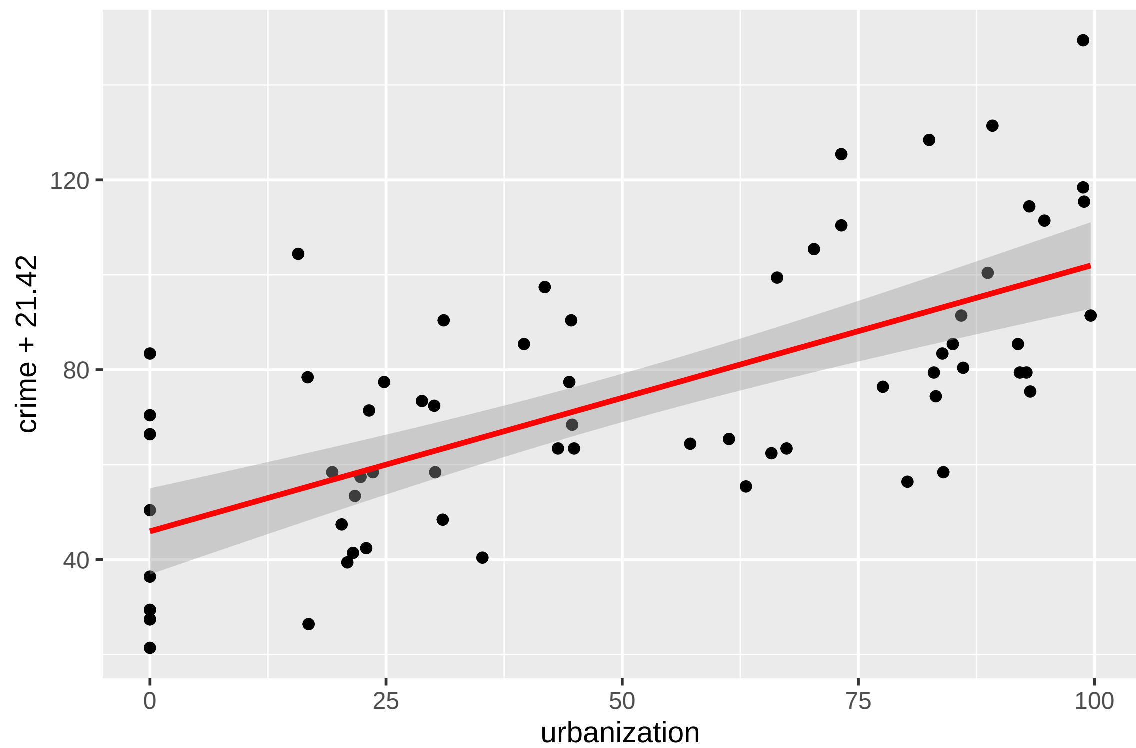
$$crime_i = 59.12 - 0.58 \cdot education_i + 0.68 \cdot urbanization_i$$

The resulting p-value from hypothesis testing is sufficiently low to conclude that β_2 coefficient for *urbanization* is statistically significant. However, the *education* coefficient β_1 now has $p \approx 0.22 > 0.05$, meaning we fail to reject the null hypothesis $H_0 := \beta_1 = 0$.

The change in relationship between *crime* and *education* seems to have occured increase *education* may be correlated with increase in another variable, such as *urbanization*.

We can plot as a plane accordingly.

```r
ggplot(fl_crime, aes(x = urbanization, y = crime + 21.42)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")
```




```r
plot3d(lm.obj, size = 5)

# segments3d(rep(TV, each=2),
#            rep(radio, each=2),
#            z=matrix(t(cbind(sales,predict(lm.obj))), nc=1),
#            add=T,
#            lwd=2,
#            col=2)
```

Part 3

3. For multiple linear gression *crime* ~ *education* + *urbanization* + *income*, proceed to

```
a. Write down the __full modeling equation__, with all __error assumptions__.
b. Show that $y_i \sim N(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}, \ \ \sigma^2) $
c. Having fitted the model from $(a)$, provide the __fitted equation__.
d. Write down the hypotheses ( in terms of parameters of the model in part (a)) and make conclusions fo
e. Interpret the effect of the only statistically significant predictor from part (d).
f. Formulate the hypotheses (in terms of parameters of the model in part (a)) for testing the overall m
```

3a, 3b.

We first show that $E[y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$.

Consider,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon.$$

Then,

$$E[y_i] = E[\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon.].$$

By the linearity of expectation,

$$E[y_i] = E[\beta_0] + E[\beta_1 x_{1,i}] + E[\beta_2 x_{2,i}] + E[\beta_3 x_{3,i}] + E[\epsilon].$$

Since β_i are constants,

$$E[y_i] = \beta_0 + \beta_1 E[x_{1,i}] + \beta_2 E[x_{2,i}] + \beta_3 E[x_{3,i}] + E[\epsilon].$$

And since $x_[i, j]$ are all fixed values,

$$E[y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + E[\epsilon].$$

Finally, since $\epsilon \sim N(0, \sigma^2)$, $E[\epsilon] = 0$.

Thus,

$$E[y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i},$$

as we wished to show.

Next, we show that

$$V[y_i] = \sigma^2.$$

$$V[y_i] = V[\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon].$$

We know by the linearity of variance that,

$$= V[\beta_0] + V[\beta_1 x_{1,i}] + V[\beta_2 x_{2,i}] + V[\beta_3 x_{3,i}] + V[\epsilon].$$

But $x_[i, j]$ and β_i do not vary. So,

$$V[y_i] = V[\epsilon]$$

.

Finally, since $\epsilon \sim N(0, \sigma^2)$, $V[\epsilon] = \sigma^2$.

Thus,

$$V[y_i] = \sigma^2,$$

as required.

To see that $y_i \sim N(0, \sigma^2)$, we note that y_i is just ϵ shifted by $\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i}$.

3c.

```r
lm.obj <- lm(crime ~ education + urbanization + income, fl_crime)

summary(lm.obj)
```

```
##
## Call:
## lm(formula = crime ~ education + urbanization + income, 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        -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
```

The fitted equation is thus,

$$\hat{y}_i = 59.71 - 0.47 \cdot education + 0.70 \cdot urbanization - 0.38 \cdot income$$

3d.

The hypotheses differ depending on the variable in question. For example,

For **urbanization**,

$$H_0 := (\beta_3 = 0)$$

and

$$H_1 := (\beta_3 \neq 0).$$

The others follow similarly.

3e.

For $\hat{\beta}_3 = 0.70$, we interpret that when holding *education* and *income* constant, we expect a 0.70 unit increase in *crime* per unit increase in *urbanization*, on average.

3f.

To test the overall significance of the model, we would instead use the hypotheses:

$$H_0 := (\beta_i = 0, \quad \forall i)$$

$$H_a := (\exists i \ st. \ \beta_i \neq 0).$$

Since for the F-test, $p = 7.89 \times 10^{-9} < 0.05$, we reject H_0, and thus the model is significant.

Problem #2 (Why need F-statistic?)

Part 1

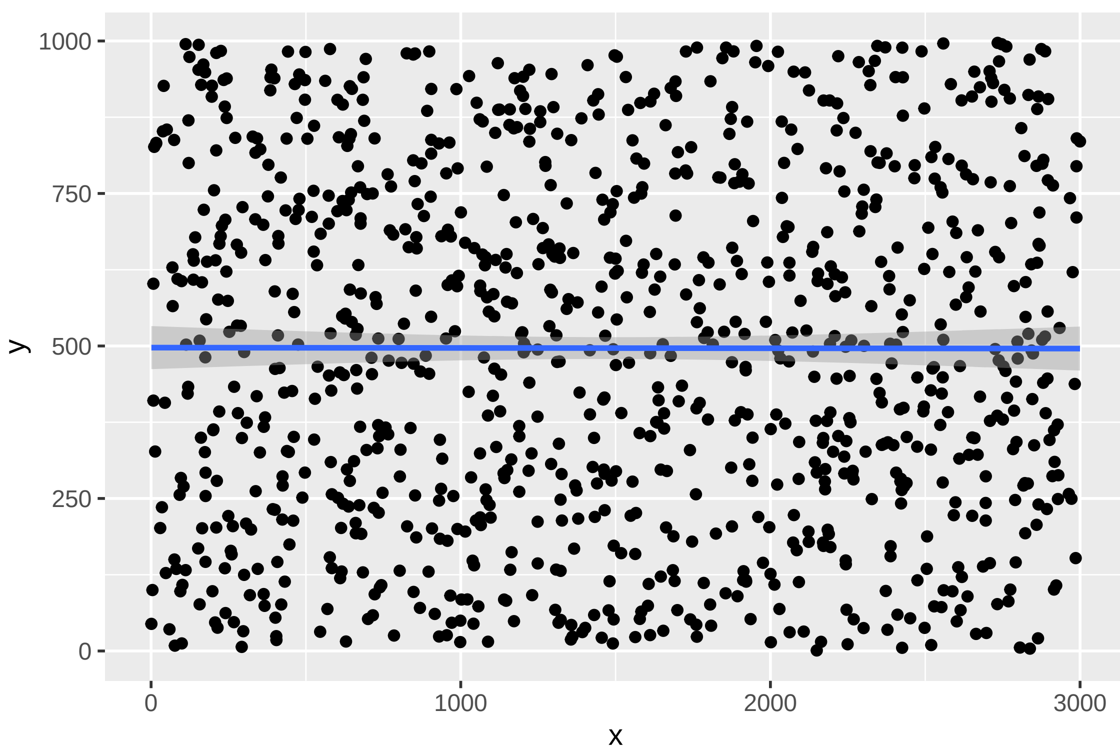
1. Generate a data example where you have a response variable y and a predictor variable x that are *unrelated* to each other (make sure to use a **random** generation mechanism). How would you do that? How would you demonstrate that they're unrelated (think of basic visualizations)?

Solution 1.

```
y <- runif(1000, min = 0, max = 1000)
x <- runif(1000, min = 0, max = 3000)

rand.df <- data.frame(x,y)

ggplot(rand.df, aes(x=x, y=y))+
  geom_point()+
  stat_smooth(method = 'lm')
```



In order to make sure the response and predictors were *unrelated*, I genderated them both by sampling from a random uniform distribution. To demonstrate they are unrelated, we can simply create a simple linear regression model and show that the slope of the model is practically 0, as shown.

Part 2

2. Having settled on a method of generating such unrelated variables in part 1, proceed to:

```
a. Generate response variable $y$ (e.g. of length 200)
b. Generate 50 predictor variables $x$ according to your method from part 1. __Record them__.
```

2a. and 2b.

```r
y <- y <- runif(200, min = 0, max = 1000)

df <- data.frame(y)


# Make 50 columns of random uniformly distributed values
for(i in 1:50){
  x <- runif(200, min = -1000, max = 1000)
  df <- cbind(df, x)
}

# Give the columns of the dataframe unique names.
var_names <- sprintf("X%s", 0:50)
var_names[1] <- "Y"
# var_names

colnames(df) <- var_names
# df
```

Part 3

3. Fit a **multiple** linear regression model, regression response y from part 2(a) on all 50 x's you've generated in part 2(b).

```
a. Report the \# of individual $t$-tests that resulted into a significant $p$-value, hence rejecting $H
b. Given that the individual siginifant $t$-test aren't necessarily indicative of at least one predicto
```

For reference, use in-class demo (slide #42).

3a.

```r
lm.obj <- lm(Y~., df)
# Take only the p value portion of the summary
p_values <- summary(lm.obj)$coefficients
# Find which column has the p values
# p_values[,4]

# List of variables which have a value that implies statistical significance
significant_variables <- p_values[,4][p_values[,4] < 0.05]
```

There are 3 p-values and associated terms which lead to the rejection of the null hypothesis under the t-test with 95% confidence level. Those variables are y-intercept β_0, and two coefficients β_i and β_j.

This doesn't really make sense, since I generated all variables independent of one another, with uniform random distributions.

However, upon closer inspection, we remember that this corresponds to a Type I error. We rejected

$$H_0 := (\beta_i = 0,\ \forall i \in \mathbb{N}),$$

even though by construction H_0 is true. This gives concrete evidence to the interpretation of the $\alpha = 0.05$ value, which gives a 5% rate of Type I errors.

3b.

The appropriate testing procedure would be the F-test. With the F-test, we can assess whether at least one predictor has a strong relationship with the response variable. A low F-statistic would imply that the ratio of variance esxplained by our model to unexplained variance is close to 0, thus the perceived relationship we found in the previous t-test is due to unexplained variance.

```r
# F-test
summary(lm.obj)
```

```
##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -566.44 -185.03    2.44  195.26  499.59
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 507.512219  23.066475  22.002   <2e-16 ***
## X1           -0.068794   0.042100  -1.634   0.1044
## X2           -0.063551   0.039432  -1.612   0.1091
## X3           -0.010919   0.042297  -0.258   0.7966
## X4            0.034378   0.040759   0.843   0.4003
## X5           -0.011580   0.037629  -0.308   0.7587
## X6           -0.097066   0.038799  -2.502   0.0134 *
## X7           -0.026441   0.041091  -0.643   0.5209
## X8           -0.004496   0.043819  -0.103   0.9184
## X9            0.003432   0.040185   0.085   0.9321
## X10          -0.036064   0.041152  -0.876   0.3822
## X11           0.018698   0.041993   0.445   0.6568
## X12          -0.024384   0.038957  -0.626   0.5323
## X13           0.006963   0.041003   0.170   0.8654
## X14           0.079009   0.040329   1.959   0.0520 .
## X15          -0.015421   0.042547  -0.362   0.7175
## X16           0.001437   0.039445   0.036   0.9710
## X17          -0.013932   0.044459  -0.313   0.7544
## X18          -0.099949   0.040160  -2.489   0.0139 *
## X19          -0.016361   0.040828  -0.401   0.6892
## X20          -0.007456   0.040386  -0.185   0.8538
## X21           0.058214   0.039792   1.463   0.1456
## X22          -0.003501   0.042574  -0.082   0.9346
## X23           0.025608   0.040469   0.633   0.5278
## X24           0.041787   0.044497   0.939   0.3492
## X25           0.034912   0.042367   0.824   0.4112
## X26           0.046596   0.037651   1.238   0.2178
## X27           0.047781   0.041778   1.144   0.2546
## X28          -0.057331   0.046503  -1.233   0.2196
```

```
## X29          0.034667  0.041841   0.829   0.4087
## X30         -0.009884  0.040835  -0.242   0.8091
## X31         -0.071464  0.041606  -1.718   0.0879 .
## X32          0.028543  0.041425   0.689   0.4919
## X33         -0.042509  0.044586  -0.953   0.3419
## X34         -0.028940  0.043535  -0.665   0.5072
## X35         -0.041293  0.038671  -1.068   0.2873
## X36          0.035366  0.041652   0.849   0.3972
## X37         -0.031467  0.040595  -0.775   0.4395
## X38          0.071125  0.041044   1.733   0.0852 .
## X39          0.042053  0.038569   1.090   0.2773
## X40          0.026927  0.039516   0.681   0.4967
## X41          0.045732  0.039711   1.152   0.2513
## X42          0.008610  0.039942   0.216   0.8296
## X43         -0.050413  0.039914  -1.263   0.2085
## X44          0.011321  0.041006   0.276   0.7829
## X45          0.021509  0.043623   0.493   0.6227
## X46          0.028465  0.040706   0.699   0.4855
## X47         -0.029728  0.041305  -0.720   0.4728
## X48         -0.065695  0.038236  -1.718   0.0879 .
## X49          0.039969  0.041113   0.972   0.3325
## X50         -0.037086  0.040080  -0.925   0.3563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 291.4 on 149 degrees of freedom
## Multiple R-squared:  0.2638, Adjusted R-squared:  0.01674
## F-statistic: 1.068 on 50 and 149 DF,  p-value: 0.3735
```

With such an incredibly small value for the F-statistic, we should be able to say that none of the predictors have a strong relationship with the response variable. We fail to reject the null hypothesis for confidence level 90%,

$$H_0 := (\beta_0, \beta_1, ..., \beta_5 0) = 0$$

Since $p = 0.73 > 0.05$.

Thus we cannot be sure that all of the β_i's are not 0.

Problem #3

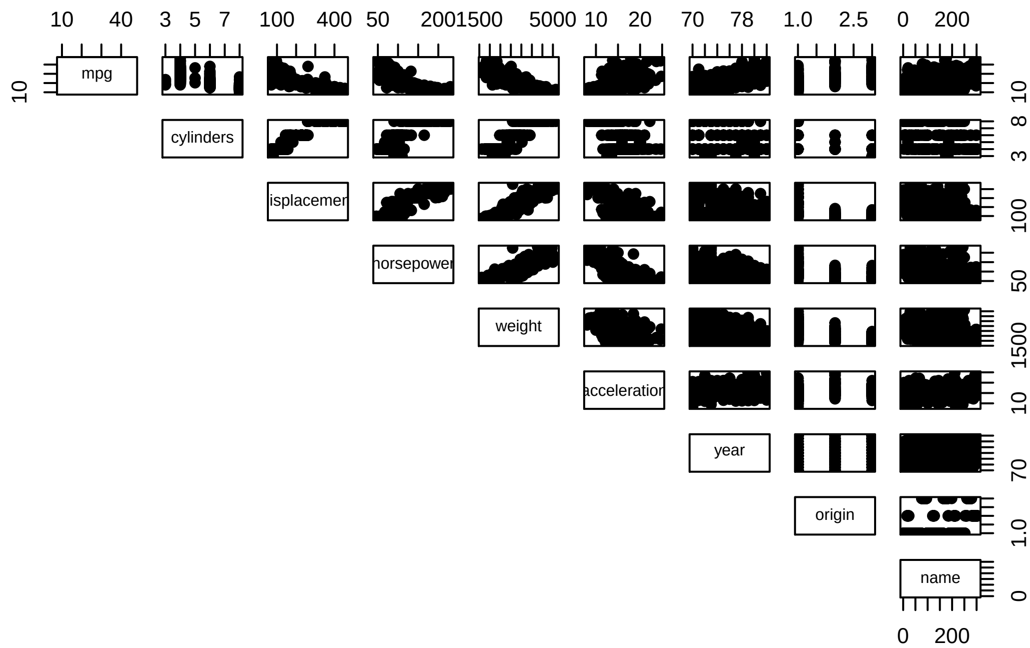
Part 1

This question involves the use of multiple linear regression on the *Auto* data set of *ISLR* library.

1. Produce a scatterplot matrix which includes all of the variables in the data set. Which variables appear to have a strong linear relationship with our intended response variable - miles per gallon (*mpg*)?

1.

```
pairs(Auto, pch = 19, lower.panel = NULL)
```



Dis-
placement, horsepower, and weight all seem to have strong linear relationships with *mpg*.

Solution 1 end.

Part 2

2. Compute the matrix of correlations between the variables using the function *cor()*, to confirm your observation from part 1. Which predictors have strongest linear relationship with *mpg*?

2.

```
auto.continuous <- Auto %>%
  select(-year, -origin, -name)


cor(auto.continuous)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
##              acceleration
## mpg             0.4233285
## cylinders      -0.5046834
## displacement   -0.5438005
## horsepower     -0.6891955
## weight         -0.4168392
## acceleration    1.0000000
```

As claimed, *displacement*, *horsepower*, and *weight* are all have the strongest linear relationship with *mpg*.

Solution 2 end.

Part 3

3. Pick one predictor variable that you feel to have the strongest **linear** relationship with *mpg*, and preform a simple linear regression. Use the *summary()* fnction to print the results.

```
a. Is there a statistically significant relationship between the predictor and the response? Provide th
b. What is the predicted _mpg_ associated with the median value of you predictor's range? Interpret tha
c. Provide and interpret both metrics for the qualify of model fit.
```

3a.

```
lm.obj <- lm(mpg ~ weight, Auto)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9736  -2.7556  -0.3358   2.1379  16.5194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.216524   0.798673   57.87   <2e-16 ***
## weight      -0.007647   0.000258  -29.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.333 on 390 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6918
## F-statistic: 878.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

Since $p < 0.5$ for the coefficient of *weight*, we reject the null hypothesis and conclude that the relationship between *weight* and *mpg* is statistically significant.

3b.

The equation of the fitted line is,

$$mpg = 46.2165 - 0.007647 \cdot weight$$

We can first calculate the median,

```
median <- median(Auto$weight)
median
```

```
## [1] 2803.5
```

and finally plug it into our fitted equation.

```
46.2165 -0.007647*median
```

```
## [1] 24.77814
```

Thus, for the median value of *weight*, the predicted *mpg* of a car is 24.78 MPG.

3c.

For our model, we have,

$$R^2 = 0.69,$$

so that we can say 69 of the variance is explained by the model.

We also have that,

$$RSE = 4.33$$

and given that our predictor variable is *mpg*, this is actually fairly significant.

The model is decent, but there is much that is left unexplained by the model, and it is can be seen in the RSE.

Solution 3 end.

Part 4

4. Use the *lm()* function to perform a multiple linear regression with *mpg* as the response and all other variables (except *name*) as the predictors. Use the *summary()* function to print the results.

a. Formulate the H_0 and H_a hypotheses (using parameter notation) for testing whether the overall model is significant. Which part of *summary()* output corresponds to this test? Is the model significant?
b. Which predictors appear to have a statistically significant relationship to the response? Just list them.
c. Interpret effects of the **two** most statistically significant predictors. Compare the interpretation here, with the one given in part 3(a) - what's the crucial difference?
d. Report and interpret the 95% confidence intervals for *weight* and *year* effects.
e. Report and interpret both quality-of-fit metrics.

Solution 4.

```r
auto.continuous <- Auto %>% select(-name)

lm.obj <- lm(mpg~., auto.continuous)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = auto.continuous)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
```

```
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

4a.

For this problem, we use the F-test to test whether the overall model is significant. Our hypotheses are:

$$H_0 := (\beta_i = 0, \quad \forall i \in \mathbb{N}$$

$$H_a := (\exists i \in \mathbb{N} \quad st. \quad \beta_i \neq 0)$$

The last line of the summary output, the F-statistic and its corresponding p value, refers to the results of this test. Since $p < 0.05$ we can conclude that for a 90% confidence level, we reject the null hypothesis. That is, our model is significant, there is at least one coefficient which is non-zero.

4b.

The predictors *displacement*, *weight*, *year*, and *origin* all seem to have a statistically significant relationship with the response.

4c.

Under the multiple linear regression model containing all coefficients, $\beta_{year} = 0.7507$ and $\beta_{weight} = -0.0065$.

That is, when all other variables are fixed, for every 4 unit increase in *year*, we can expect about a 3 unit increase in *mpg*, on average. And when all other variables are fixed, for every 1000 unit increase in *weight*, we can expect a 6.5 unit decrease in *mpg*.

This result is relatively similar to the result we had in problem 3a, but earlier we concluded that *year* did not have a very strong relationship with *mpg* when initially looking at the correlation matrix.

Crucially, the intercept of the fitted line is incredibly dependent on the other values in the model.

4d.

```r
confint(lm.obj,c('weight', 'year'), level = 0.95)
```

```
##               2.5 %       97.5 %
## weight -0.007756074 -0.005192013
## year    0.650551315  0.850994041
```

We are 95% confident that the true model values for the coefficients lie between the intervals shown above.

4e.

Lastly, we interpret the quality of fit metrics.