LM Homework 2 B

Joshua Ingram

2/19/2020

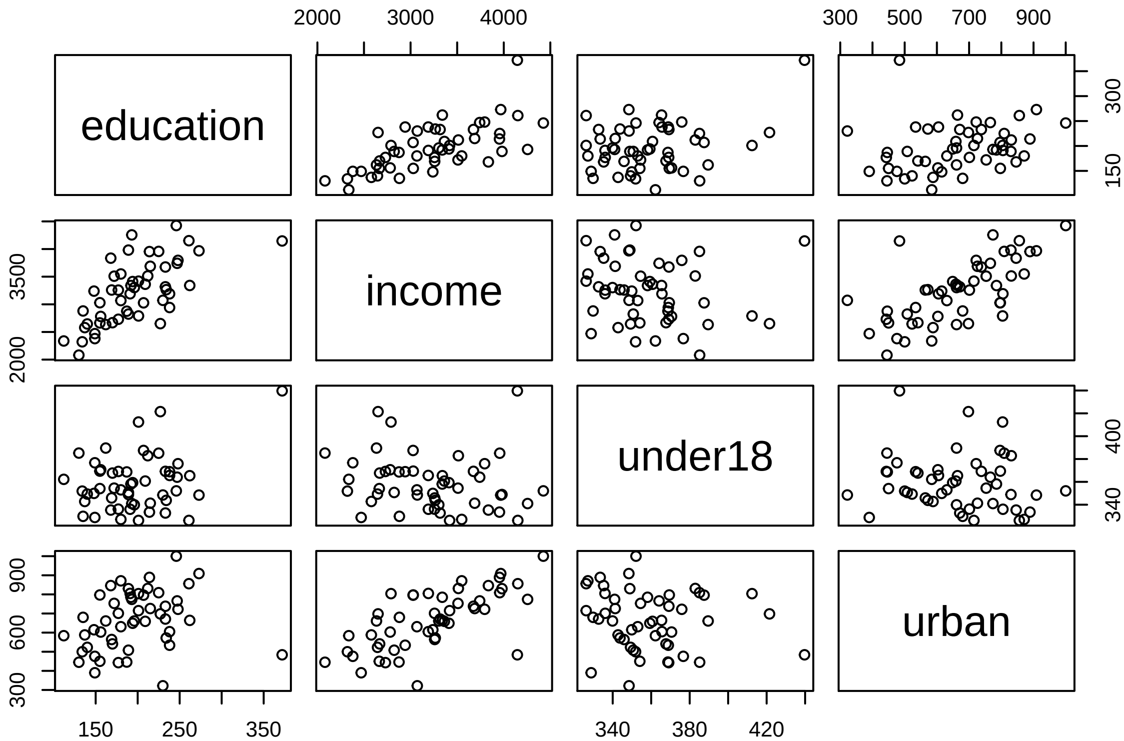
Problem 1

1.

```r
head(Anscombe)
```

```
##    education income under18 urban
## ME       189   2824   350.7   508
## NH       169   3259   345.9   564
## VT       230   3072   348.5   322
## MA       168   3835   335.3   846
## RI       180   3549   327.1   871
## CT       193   4256   341.0   774
```

```r
plot(Anscombe)
```




```r
cor(Anscombe)
```

```
##           education     income    under18      urban
## education 1.0000000  0.6675773  0.3114855  0.2633238
## income    0.6675773  1.0000000 -0.1623600  0.6854580
## under18   0.3114855 -0.1623600  1.0000000 -0.1386334
## urban     0.2633238  0.6854580 -0.1386334  1.0000000
```

income seems to be the variable that has the strongest relationship with education spendings.

2.

```r
lm_edu1 <- lm(education ~ income, data = Anscombe)
summary(lm_edu1)
```

```
##
## Call:
## lm(formula = education ~ income, data = Anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.077 -21.868  -4.617  17.523 124.701
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.710031  28.873840   0.613    0.542
## income       0.055376   0.008823   6.276 8.76e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.94 on 49 degrees of freedom
## Multiple R-squared:  0.4457, Adjusted R-squared:  0.4343
## F-statistic: 39.39 on 1 and 49 DF,  p-value: 8.762e-08
```

Equation: $\hat{education_i} = 17.71 + 0.0554(income_i)$

Interpretations:

slope: On average, we predict that there will be a \$0.055 increase in per-capita education expenditures for every one dollar increase in per-capita income.

intercept: When per-capita income is 0, we predict, on average, that the per-capita education expenditures will be \$17.71. (not sure about the full context of this data, but perhaps this is the per-capita income for a given county, so if a county has close to 0 per-capita income, they may still receive state/federal assistance? That's why I interpretted)

RSE: 34.94

interpretation: On average, our estimates for education expenditures (per-capita) are off by 34.94 dollars

R^2 : 0.4457

intperpretation: 44.57% of the variance in education ependitures per capita is explained by our model

3.

a.

$\hat{education_i} = \hat{\beta}_0 + \hat{\beta}_1(income_i) + \hat{\beta}_2(under18_i) + \hat{\beta}_3(urban_i) + \epsilon_i$

b.

Find that: $\alpha = \bar{y} - \beta_1\bar{x_1} - \beta_2\bar{x_2} - \beta_3\bar{x_3}$

$\frac{\delta}{\delta\alpha}\Sigma(y_i - (\alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3))^2$

$= \Sigma\frac{\delta}{\delta\alpha}(y_i - (\alpha + \beta_1x_1 - +\beta_2x_2 + \beta_3x_3))^2$

$= \Sigma 2(y_i - (\alpha + -\beta_1x_1 - \beta_2x_2 - \beta_3x_3)(-1)$

$= \Sigma - 2(y_i - (\alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3) = 0$

$\Sigma(y_i - (\alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3) = 0$

$\Sigma y_i - \Sigma(\alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3) = 0$

$\Sigma y_i - \Sigma\alpha - \Sigma\beta_1x_1 - \Sigma\beta_2x_2 - \Sigma\beta_3x_3) = 0$

$\Sigma y_i - n\alpha - \Sigma\beta_1x_1 - \Sigma\beta_2x_2 - \Sigma\beta_3x_3) = 0$

$\Sigma y_i - \Sigma\beta_1x_1 - \Sigma\beta_2x_2 - \Sigma\beta_3x_3) = n\alpha$

$\alpha = \bar{y} - \beta_1\bar{x_1} - \beta_2\bar{x_2} - \beta_3\bar{x_3}$

c.

```r
lm_edu2 <- lm(education ~ income + under18 + urban, data = Anscombe)
summary(lm_edu2)
```

```
##
## Call:
## lm(formula = education ~ income + under18 + urban, data = Anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.240 -15.738  -1.156  15.883  51.380
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.868e+02  6.492e+01  -4.418 5.82e-05 ***
## income       8.065e-02  9.299e-03   8.674 2.56e-11 ***
## under18      8.173e-01  1.598e-01   5.115 5.69e-06 ***
## urban       -1.058e-01  3.428e-02  -3.086  0.00339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.69 on 47 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6698
## F-statistic: 34.81 on 3 and 47 DF,  p-value: 5.337e-12
```

fitted equation: $\hat{education_i} = -0.02868 + 0.08065(income_i) + 0.8173(under18_i) - .1058(urban_i)$

intercept: doesn't make sense to interpret

income: on average and holding all other variables constant, we predict there will be a 0.08065 dollar increase in education expenditure per capita for every one dollar increase in income per capita

under18: on average and holding all other variables constant, we predict there will be a 0.8173 dollar increase in education expenditure per capita for every 1 person increase in the number of people under 18 per 1000.

urban: on average and holding all other variables constant, we predict there will be a .1058 dollar decrease in education expenditure per capita for every one dollar increase in the number of urban per 1000

RSE: 26.69 (On average, our estimates for education expenditures (per-capita) are off by 26.69 dollars)

R^2 : 0.6896 (68.96% of the variance in education expenditures per capita is explained by our mdoel)

d.

```r
r2 <- summary(lm_edu2)$r.squared
rse <- summary(lm_edu2)$sigma

fitted <- as.vector(lm_edu2$fitted.values)
response <- Anscombe[,1]
residuals <- response - fitted
mean <- mean(response)

rse2 <- sqrt(sum(residuals^2) / (51-4))
rse2
```

```
## [1] 26.69343
```

```
r2_2 <- 1 - sum(residuals^2)/sum((response - mean)^2)
r2_2
```

```
## [1] 0.6896288
```

Both the output from my "hard-code" and summary are the same ### e.

We could not compare the practical effects of the three explanatory variables because they are not in terms of the same units. We have to standardize them to be in terms of standard deviations to objetively compare the effects of each variable.

```
scaled_edu <- data.frame(scale(Anscombe))
lm_edu3 <- lm(education ~ income + under18 + urban, data = scaled_edu)
summary(lm_edu3)
```

```
##
## Call:
## lm(formula = education ~ income + under18 + urban, data = scaled_edu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29675 -0.33879 -0.02489  0.34191  1.10602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.084e-16  8.046e-02   0.000  1.00000
## income       9.723e-01  1.121e-01   8.674 2.56e-11 ***
## under18      4.216e-01  8.242e-02   5.115 5.69e-06 ***
## urban       -3.447e-01  1.117e-01  -3.086  0.00339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5746 on 47 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6698
## F-statistic: 34.81 on 3 and 47 DF,  p-value: 5.337e-12
```

It seems that income has the greatest practical effect on education expenditures.

Problem 2

```
lm_pres1 <- lm(prestige ~ education, data = Prestige)
lm_pres2 <- lm(prestige ~ education + income, data = Prestige)
lm_pres3 <- lm(prestige ~ education + income + women, data = Prestige)

summary(lm_pres1)$r.squared
```

```
## [1] 0.7228007
```

```r
summary(lm_pres2)$r.squared
```

```
## [1] 0.7980008
```

```r
summary(lm_pres3)$r.squared
```

```
## [1] 0.7981775
```

Note: I'll be referring to lm_pres1 (prestige ~ education) as model 1, lm_pres2 (prestige ~ education + income) as model 2, and lm_pres3 (prestige ~ education + income + women) as model 3 from here on

1.

model 1 R^2 : 0.7228 (72.28% of the variance in prestige is explained by model 1)

model 2 R^2 : 0.798 (79.8% of the variance in prestige is explained by model 2)

model 3 R^2 : 0.7982 (79.82% of the variance in prestige is explained by model 3)

Our R^2 increases as we add more predictors. However, the increase from model 2 to model 3 (adding the women variable) is very small

2.

$R^2 = \frac{TSS-RSS}{TSS}$

(**Note:** I want to clarify that I did seek help online to see what a more formal version of this proof looked like, as I did not fully understand it initially. However, I tried to perform this task at my own level on my own after getting a general idea of how this is done… not sure if I still fully understand this? (At least, at a mathematical level))

Given null model:

$\hat{y}' = \alpha'$

TSS is the RSS for the null model: $\Sigma(y_i - \bar{y})^2$

So, as we add more explanatory variables, the least squares regression algorithm seeks to find the equation for $\hat{y}$ that minimizes the RSE and maximizes R^2… meaning each additonal β's that are found has to be ones that minimizes the RSS such that it is at least as "good" as the TSS (the RSS for the null model)…

For a non-null model, say:

$\hat{y} = \alpha + \beta_1 x_1$

The RSS: $\Sigma(y_i - \hat{y})^2$ is the same or smaller than the TSS since the RSS is minimized… meaning that the β estimate is at least 0

i.e.

$\hat{y} = \alpha + (0)x_1$, which is the same as the null model…

So as the number of explanatory variables increase, the same logic applies where RSS gets smaller or stays the same (but cannot be worse than the null model or the one before since the alogrithm seeks to minimize)

Problem 3

Prove $E[y_i] = \alpha + \beta x_i$

$E[Y_i] = E[\alpha + \beta x_i + \epsilon_i]$

$= E[\epsilon_i] + E[\alpha + \beta x_i]$

$= 0 + \alpha + \beta x_i$

Prove $Var[y_i] = \sigma^2$

$Var[Y_i] = Var[\alpha + \beta x_i + \epsilon_i]$

$= Var[\alpha + \beta x_i] + Var[\epsilon_i]$

$= 0 + \sigma^2$

Prove $Y - i$ follows a normal distribution with a mean μ

Given the expected value and variance above and that ϵ follows a normal distribution:

$Y_i \ N(\alpha + \beta x_i, \sigma^2)$

It only sees a shift in the mean, but the normality remains.

Prove Independence

Given $P(\epsilon_i = e_i, \epsilon_j = e_j) = P(\epsilon_i = e_i) * P(\epsilon_j = e_j)$

$P(Y_i = y_i, Y_j = y_j)$

$= P(\alpha + \beta x_i + \epsilon_i = y_i, \alpha + \beta x_j + \epsilon_j = y_j)$

$= P(\epsilon_i = y_i - \alpha - \beta x_i, \epsilon_j = y_j - \alpha - \beta x_j)$

$= P(\epsilon_i = y_i - \alpha - \beta x_i) * P(\epsilon_j = y_j - \alpha - \beta x_j)$

$= P(\alpha + \beta x_i + \epsilon_i = y_i) * P(\alpha + \beta x_j + \epsilon_j = y_j)$

$= P(Y_i = y_i) * P(Y_j = y_j)$

Thus:

$Y_i = \mu + \epsilon_i, \epsilon \ _{i.i.d.} N(0, \sigma^2)$

leads to: $Y_i \ _i N(\mu, \sigma^2)$

(Wouldn't Y_i not be identically distributed since it will have different meaens for different x, but it will still be independently distributed?)

Problem 4

Lab 1

FL Crime Data

Exploration and Data cleaning

```
# - Explore the data set.
head(fl_crime)
```

```
##     county crime.rate..per.1000. education.... urbanization....
## 1  Alachua                   104          82.7             73.2
## 2    Baker                    20          64.1             21.5
## 3      Bay                    64          74.7             85.0
## 4 Bradford                    50          65.0             23.2
## 5  Brevard                    64          82.3             91.9
## 6  Broward                    94          76.8             98.9
##   income..median..in.1000.
## 1                     22.1
## 2                     25.8
## 3                     24.7
## 4                     24.6
## 5                     30.5
## 6                     30.6
```

```r
# - Clean the column names.
ls(fl_crime)
```

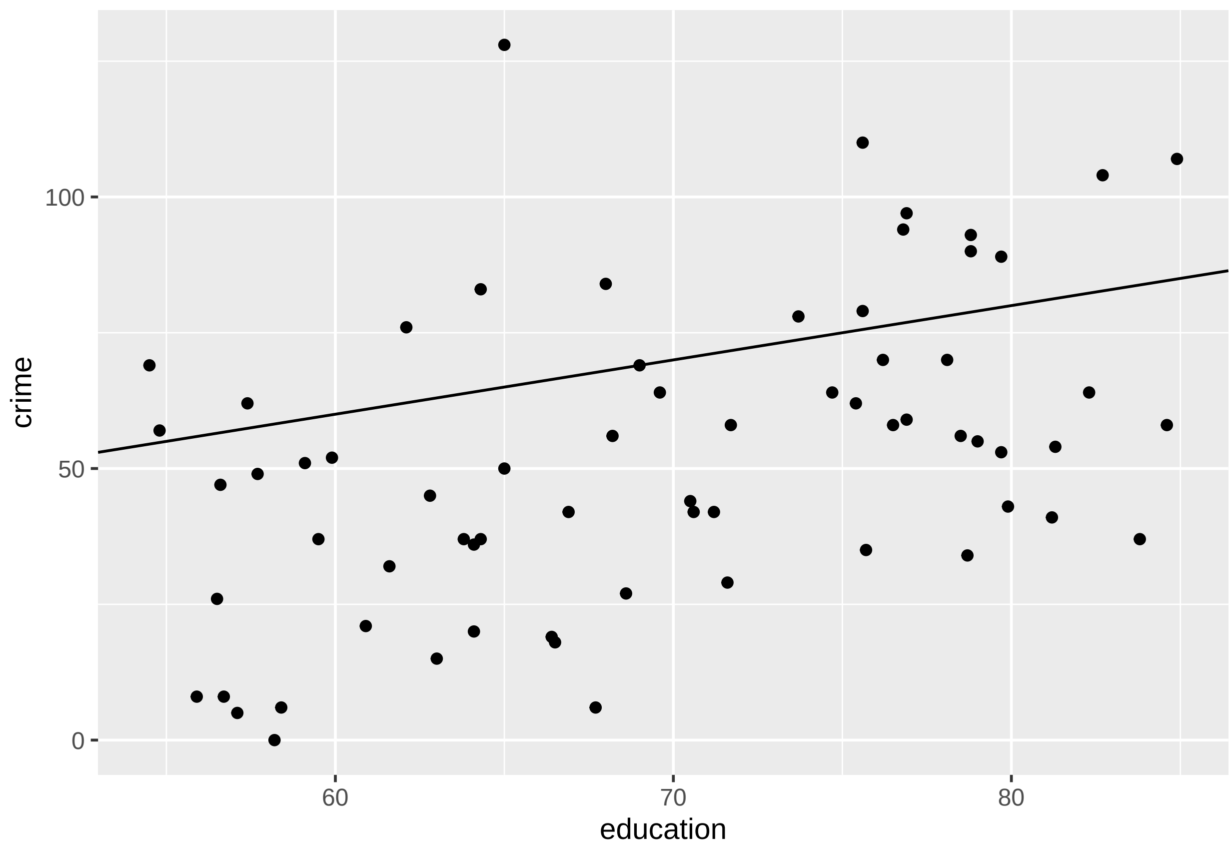
```
## [1] "county"                      "crime.rate..per.1000."
## [3] "education...."               "income..median..in.1000."
## [5] "urbanization...."
```

```r
colnames(fl_crime) <- c("county", "crime", "education", "urbanization", "income")
head(fl_crime)
```

```
##     county crime education urbanization income
## 1  Alachua   104      82.7         73.2   22.1
## 2    Baker    20      64.1         21.5   25.8
## 3      Bay    64      74.7         85.0   24.7
## 4 Bradford    50      65.0         23.2   24.6
## 5  Brevard    64      82.3         91.9   30.5
## 6  Broward    94      76.8         98.9   30.6
```

Relationship between crime and education

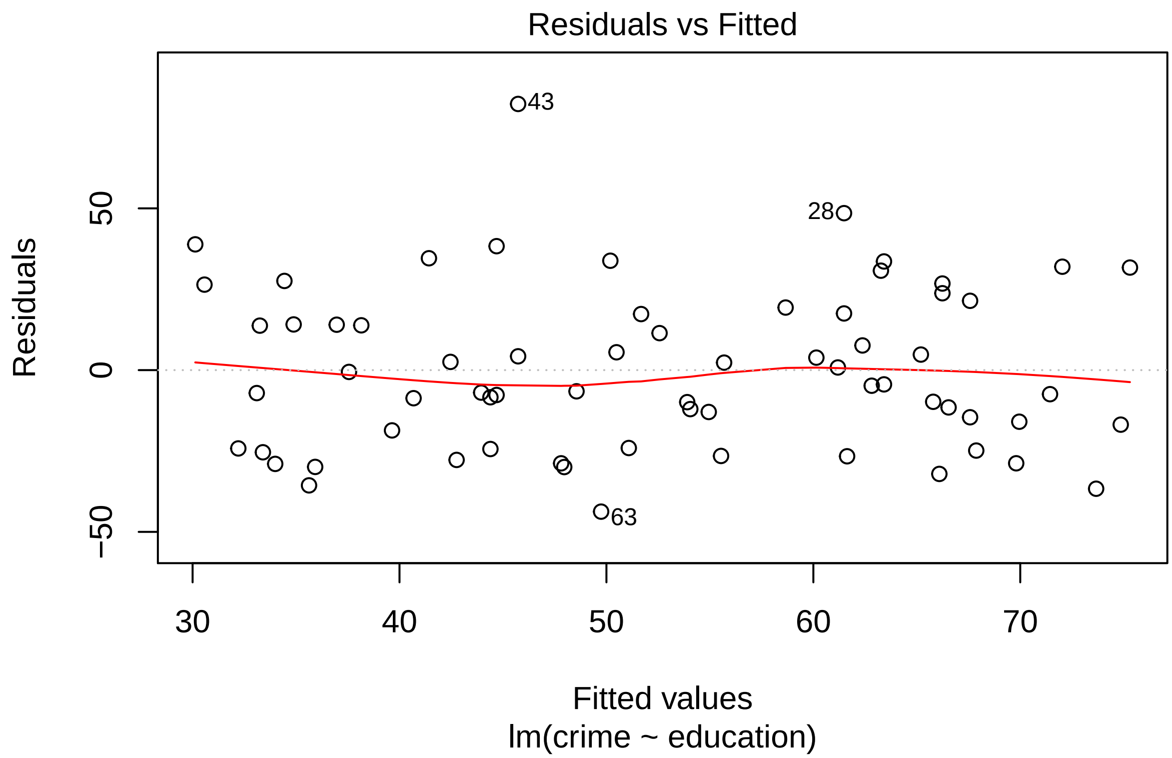
```r
ggplot(data=fl_crime,
       aes(x=education, y=crime)) + geom_point() + geom_abline()
```



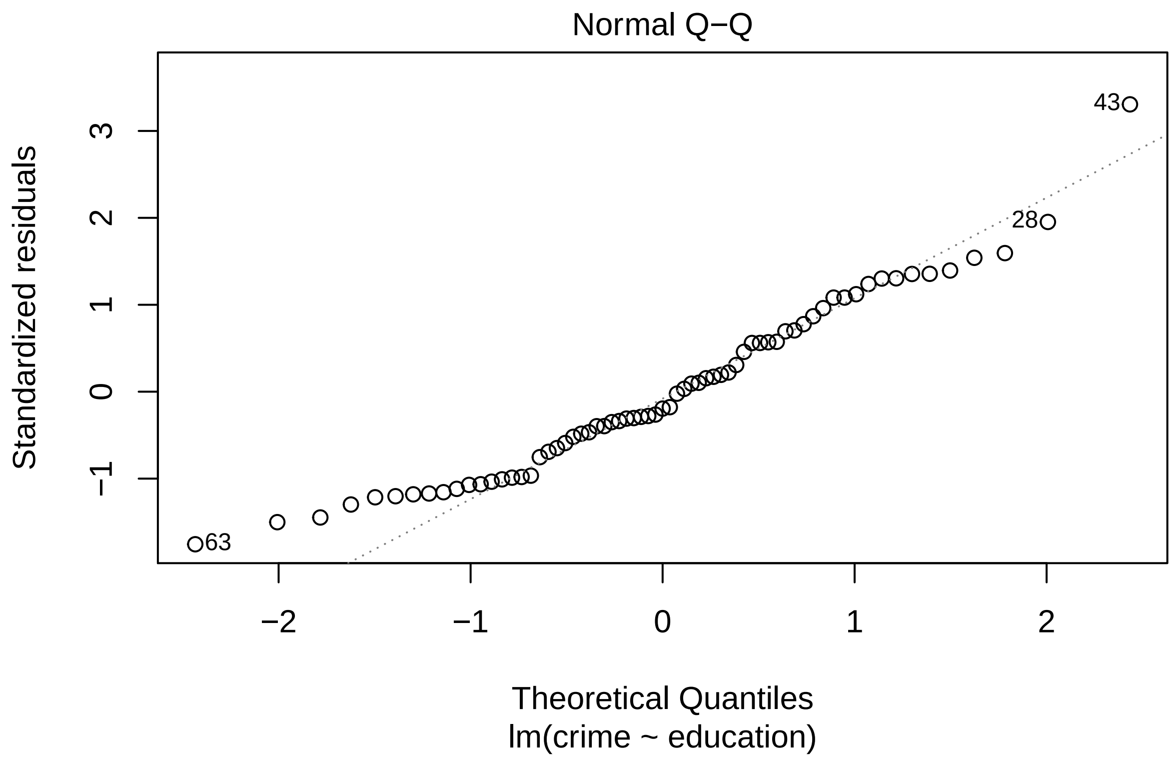

```
lm_obj_crime <- lm(crime ~ education, data = fl_crime)
summary(lm_obj_crime)
```

```
##
## 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
```

```
plot(lm_obj_crime, 1)
```




```
plot(lm_obj_crime, 2)
```



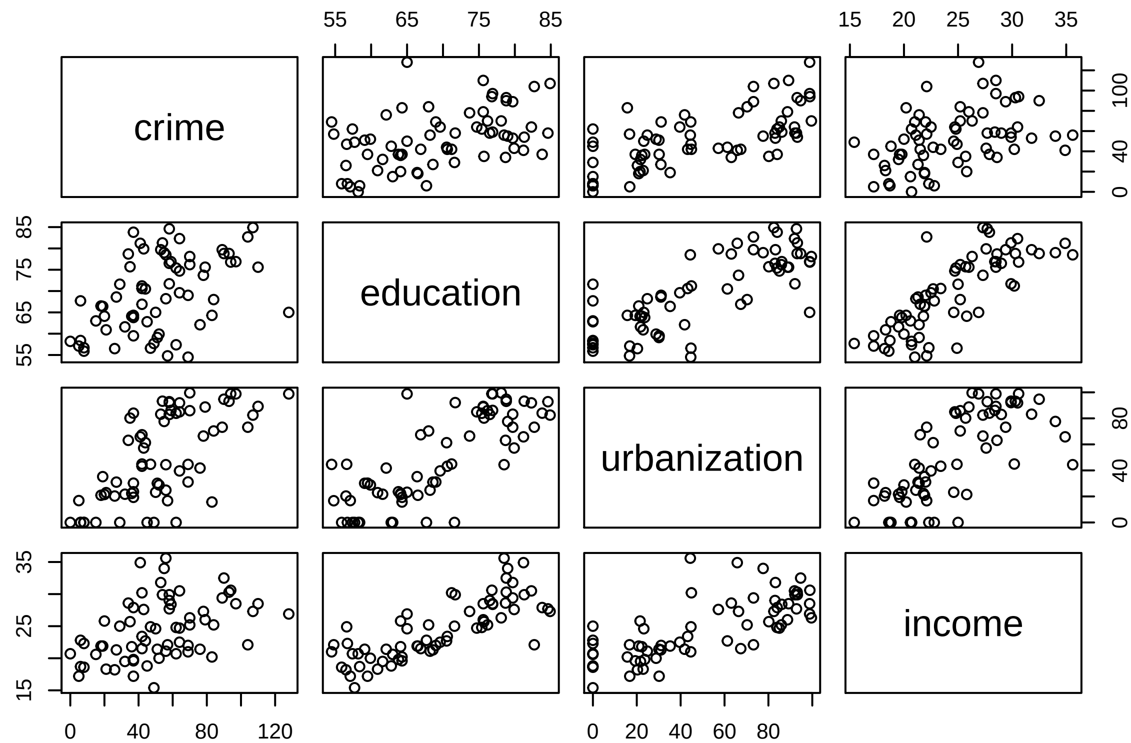
Interpretations

Intercept: Doesn't make sense to interpret the intercept because it is negative.

education: On average, for a one percentage point increase in education we expect to see a 1.486 unit increase in crime rate per 1000 people.

Lurking Variables

```
plot(fl_crime[,-1])
```




```r
cor(fl_crime[,-1])
```

```
##                  crime education urbanization    income
## crime        1.0000000 0.4669119    0.6773678 0.4337503
## education    0.4669119 1.0000000    0.7907190 0.7926215
## urbanization 0.6773678 0.7907190    1.0000000 0.7306983
## income       0.4337503 0.7926215    0.7306983 1.0000000
```

Urbanization could be this lurking variable as it has a higher correlation between crime and education.

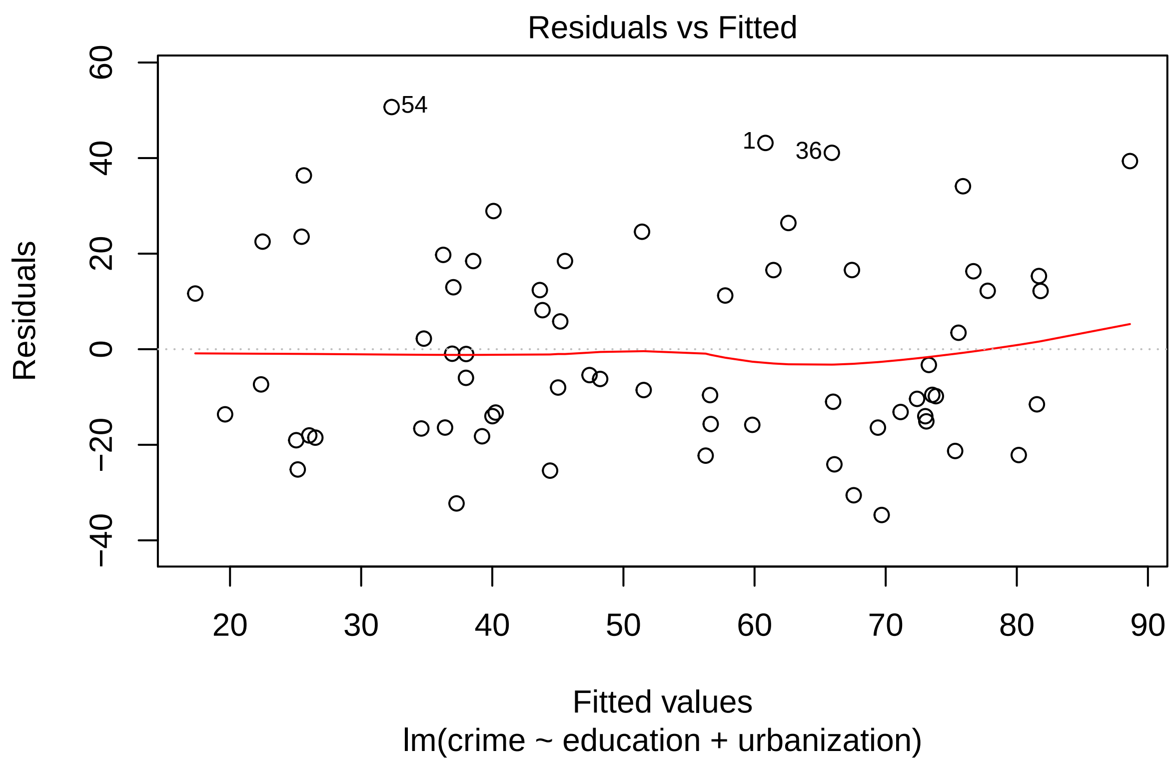
Adding urbanization as a new variable

```r
lm_obj_crime2 <- lm(crime ~ education + urbanization, data = fl_crime)
summary(lm_obj_crime2)
```

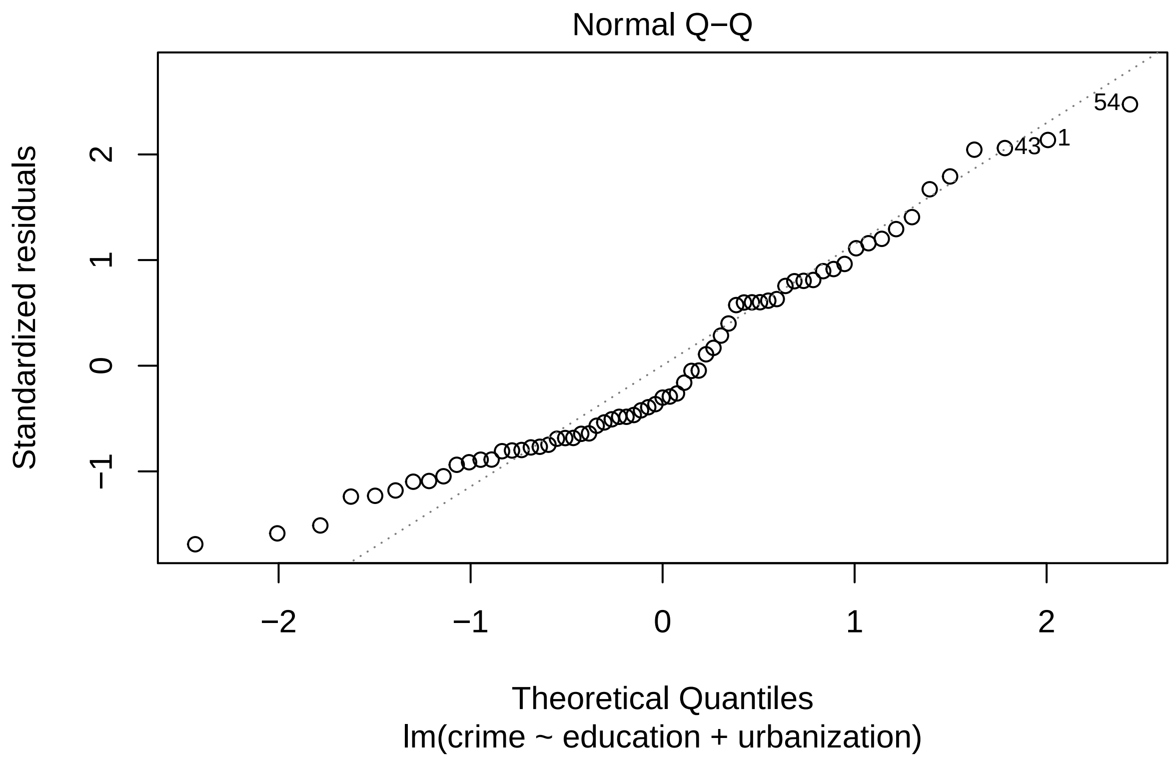
```
##
## Call:
## lm(formula = crime ~ education + urbanization, data = fl_crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.693 -15.742  -6.226  15.812  50.678
##
```

```
## 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
```

```r
plot(lm_obj_crime2, 1)
```




```r
plot(lm_obj_crime2, 2)
```



Interactive plot

(commented out due to markdown error)

```
#plot3d(lm_obj_crime2, size=5, col=1, data = fl_crime)

# - Add the lines showing the residuals (see "Lecture_2.R")

#plot3d(lm_obj_crime2, size=5, col=1, data = fl_crime)

#segments3d(rep(education, each=2), rep(urbanization, each=2),
#           z=matrix(t(cbind(crime,predict(lm_obj_crime2))), #nc=1),
#           add=T,
#           lwd=2,
#           col=2)
```

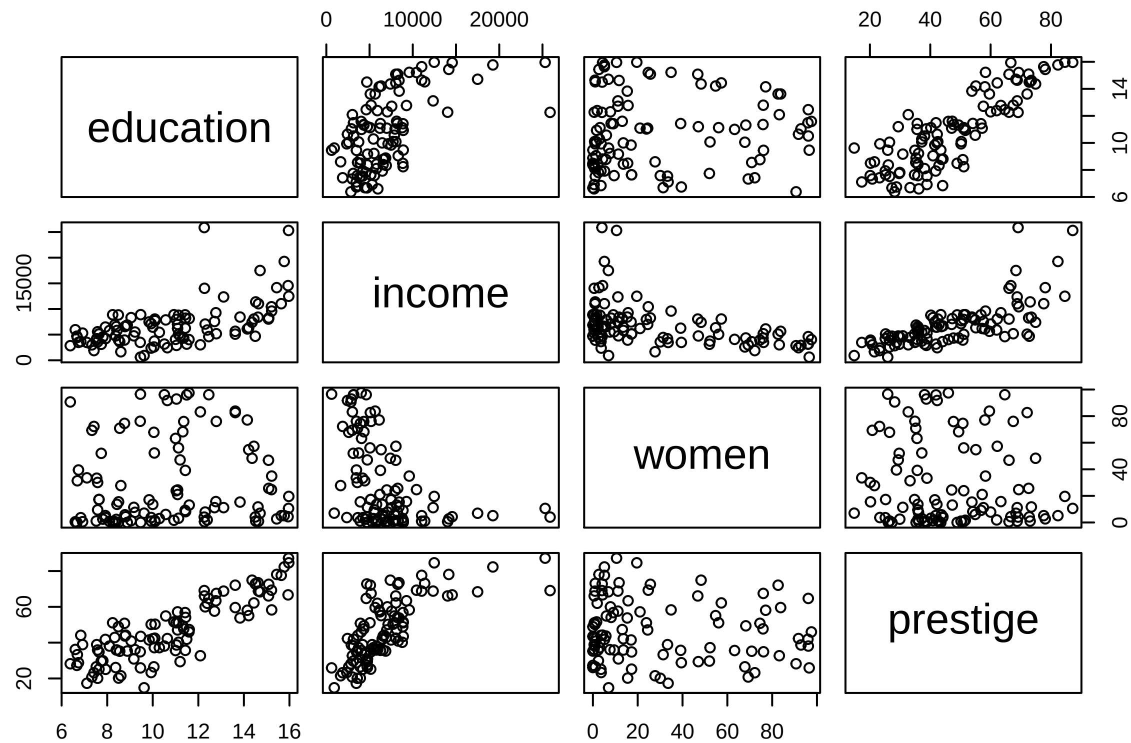
Prestige Data

Data Exploration

```
head(Prestige)
```

```
##                        education income women prestige census type
## GOV.ADMINISTRATORS         13.11  12351 11.16     68.8   1113 prof
## GENERAL.MANAGERS           12.26  25879  4.02     69.1   1130 prof
## ACCOUNTANTS                12.77   9271 15.70     63.4   1171 prof
## PURCHASING.OFFICERS        11.42   8865  9.11     56.8   1175 prof
## CHEMISTS                   14.62   8403 11.68     73.5   2111 prof
## PHYSICISTS                 15.64  11030  5.13     77.6   2113 prof
```

```
plot(Prestige[,c(-5, -6)])
```




```
cor(Prestige[,c(-5, -6)])
```

```
##            education     income       women   prestige
## education 1.00000000  0.5775802  0.06185286  0.8501769
## income    0.57758023  1.0000000 -0.44105927  0.7149057
## women     0.06185286 -0.4410593  1.00000000 -0.1183342
## prestige  0.85017689  0.7149057 -0.11833419  1.0000000
```

Census seems to be the odd one out considering the way the values are defined… Otherwise, education and income have the two most pronounced relationships with prestige. The third would be women, though it doesn't follow very closely to the other variables.

Multiple Linear Regression Model

```
lm_obj_prest <- lm(prestige ~ education + income + women, data = Prestige)
summary(lm_obj_prest)
```

```
##
## Call:
## lm(formula = prestige ~ education + income + women, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
## education    4.1866373  0.3887013  10.771  < 2e-16 ***
## income       0.0013136  0.0002778   4.729 7.58e-06 ***
## women       -0.0089052  0.0304071  -0.293   0.7702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

$pres\hat{t}ige_i = -6.794 + 4.187(education_i) + 0.00131(income_i) - 0.0089(women_i)$

Interpretations:

intercept - doesn't make much sense to interpret

education - on average and holding all other variables constant, for every one year increase in education, we predict the prestige score to increase by 4.187 points.

income - on average and holding all other variables constant, for every one dollar increase in income, we predict the prestige score to increase by 0.0013 points

women - on average and holding all other variables constant, for every one percentage point increase in the number of incumbents who are women, we predict the prestige score to decrease by 0.0089 points

R^2 - 79.82% of the variance in prestige is explained by our model

RSE - on average, our predictions are off by 7.846 Ineo-Porter prestige score points

Standardize slopes

```
scaled_Prestige <- data.frame(scale(Prestige[,c(-5,-6)]))
lm_obj_prest2 <- lm(prestige ~ education + income + women, data = scaled_Prestige)
coef(lm_obj_prest2)
```

```
##   (Intercept)     education        income         women
## -1.396196e-17  6.639551e-01  3.241757e-01 -1.642104e-02
```

Education has the greatest practical effect on prestige, followed by income.