Quantitative Methods in Public Policy Assignment 2

Before moving into the body of this assignment, I will install requisite packages and save our data of interest as an object. I also edit a few options to expand the size of resulting figures. Throughout the assignment, I alternative between commentary and relevant code. For convenience the code is compiled at the end of this document as well.

```
# install requisite packages
# install.packages(tidyverse)
# install.packages(rvest)
# install.packages(RColorBrewer)
# install.packages(knitr)
# install.packages(here)

# save packages to library
library(tidyverse)
library(rvest)
library(RColorBrewer)
library(knitr)
library(here)

# the data analyzed in this assignment concerns the correlation
# between years of education and earnings of men who were
# aged between 41 and 50 years in 1980.
# The data can be downloaded via this link:
# https://press.princeton.edu/student-resources/thinking-clearly-with-data.


# save the data as an object
schooling_data <- read.csv(here("Files for Data Exercises/SchoolingEarnings.csv"))
  # note here() is taken from the here package.
  # It makes code reproducible across different users and
  # computer systems by creating a path from the working directory to the file's location.
  # For instance, running
  # "here("Files for Data Exercises/SchoolingEarnings.csv")"
  # in my console returns
  # "/Users/zanderarnao/Desktop/github projects/Quantitative-Methods/Files for Data
  # Exercises/SchoolingEarnings.csv".
```

With the data saved and options initialized, we now move into the assignment questions.

5.1) Run a regression with earnings as the dependent variable and schooling as the sole independent variable. Interpret the coefficients.

Below I run an ordinary least squares (OlS) regression on the schooling data, resulting in the line of best fit visualized in the scatter plot. Here schooling serves as the independent variable and earnings as the dependent. I provide a table with the relevant regression parameters: the regression coefficient (beta) and the intercept (alpha).

```r
# fit a regression line to the schooling data
regression <- lm(data = schooling_data, earnings ~ schooling)
summary(regression)
```

```
##
## Call:
## lm(formula = earnings ~ schooling, data = schooling_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6431 -1.5019 -0.9109  1.6397  3.7587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.79853    0.84655   10.39 2.82e-09 ***
## schooling    1.16185    0.07241   16.05 1.67e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.009 on 19 degrees of freedom
## Multiple R-squared:  0.9313, Adjusted R-squared:  0.9276
## F-statistic: 257.4 on 1 and 19 DF,  p-value: 1.675e-12
```

The regression coefficient and intercept are given in the table above. Beta is 1.16185, which is the slope of the regression line. Its sign is positive, which tells us that there is a positive correlation between years of schooling and earnings, i.e., that, on average, an increase in one variable is associated with an increase in the other.

The specific magnitude of the correlation 1.16185, which means that for the middle aged men in our data, each additional year of schooling corresponds, on average, to an increase in income by approximately $1,161.85 (in 1980 dollars).

Alpha is 8.79853, which tells us that the men in this cohort who received no years of schooling earned, on average, about $8,798.53 (in 1980 dollars).

Now with the specific meaning of the regression parameters for the schooling data, we briefly discuss why we would run an OLS regression.

5.2) Suppose you wanted a parsimonious way to predict earnings using only years of schooling. What would you do?

For a parsimonious way to predict earnings using only years of schooling, I would run an ordinary least squares (OLS) regression of earnings on years of schooling. That is, years of schooling would serve as the explanatory variable and earnings the outcome of interest. Years of schooling would “explain” variation in earnings seen in the data, and the OSL regression would mathematically model this relationship.

Provided that the relationship between the variables is approximately linear, this statistical test would give us a line that, based on years of schooling, summarizes change in earnings on average. OLS accomplishes this by finding the line which minimizes the total squared errors, i.e, the squared sum of the all distances between the observed and predicted values of earnings for every year of schooling.

OLS regression is useful because it is a parsimonious way to communicate the relationship between two or more variables. Alternatively, one could list all the years of schooling and their associated average earnings and draw conclusions about their relationship, but this method would be time-consuming and heavy with tedious information.

By contrast, OLS regression allows one to study this relationship without detailed reference to the observed data. OLS regression provides a mathematical model which gives highly useful information, including the direction (positive, negative, or none) of the relationship and a sense of its magnitude. This information can have a number of uses, including for description, prediction, and causal assessment of the relationship between earnings and years of schooling. Using an OLS regression model, we could, for example, make an out of sample prediction about a group woman of a similar age in 1980 and their earnings or a group of men and their earnings in 2020. There would be a fair amount of error, but the estimate could be useful.

It is also worth briefly comparing OLS regression to higher order regressions. Higher order regressions add additional explanatory variables to the process (e.g., college major as an additional explainer of income), and while this can be useful (particularly when the relationship relationship is non-linear), it adds dimensionality and complexity that reduces the parsimony of the summary. A regression equation with ten explanatory variables is likely not much more useful than listing or visualizing the data and assessing its

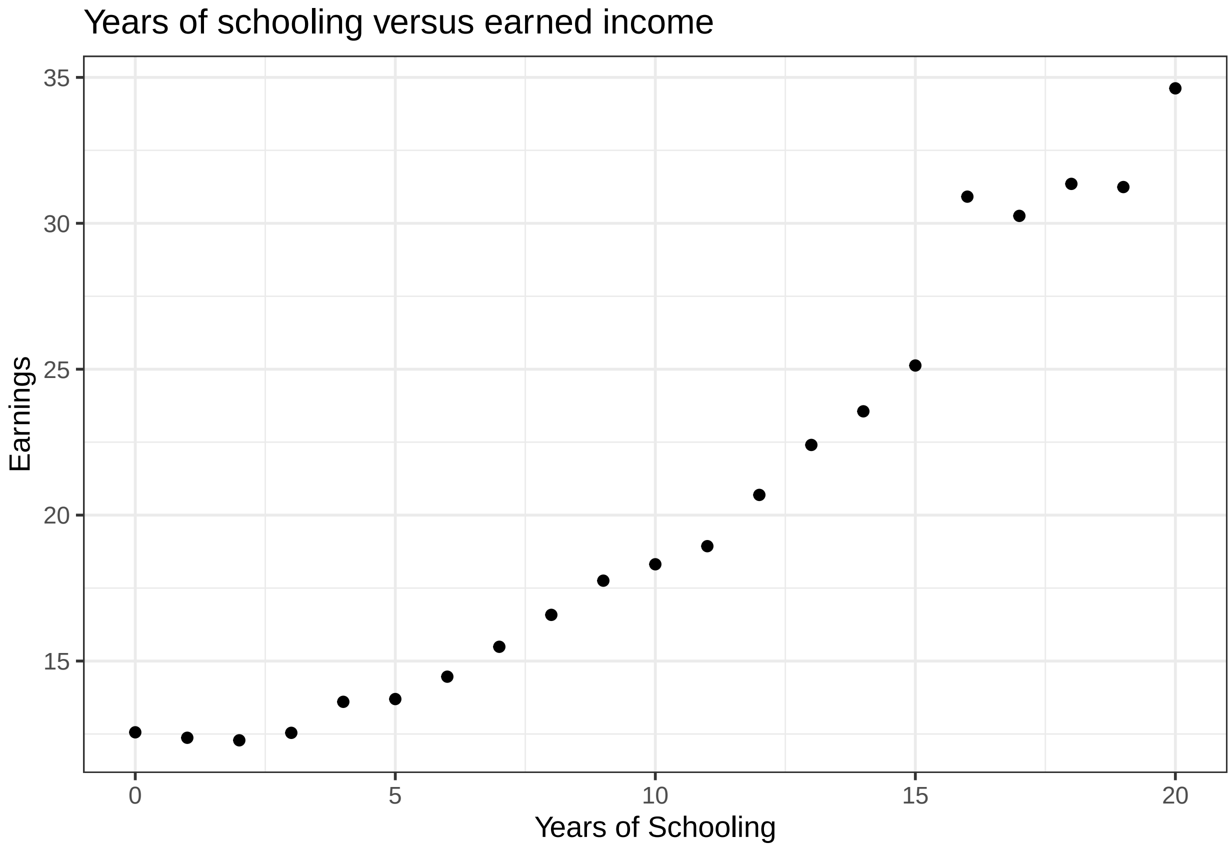
We can apply it here to find a parsimonious summary of the relationship between year of schooling and earnings. The specific meaning of identified regression coefficient and intercept are elaborated above. We not dig into whether the relationship between years of schooling and earnings is actually linear now.

5.3) Let's dig more deeply into whether the relationship between earning and schooling is approximately linear.

a) Start by making a scatter plot. Then plot the predicted values from your regression along with the raw data points, as we did in chapter 2. Does the regression line look like it's fitting the data well?

First we show a scatter plot of the data.

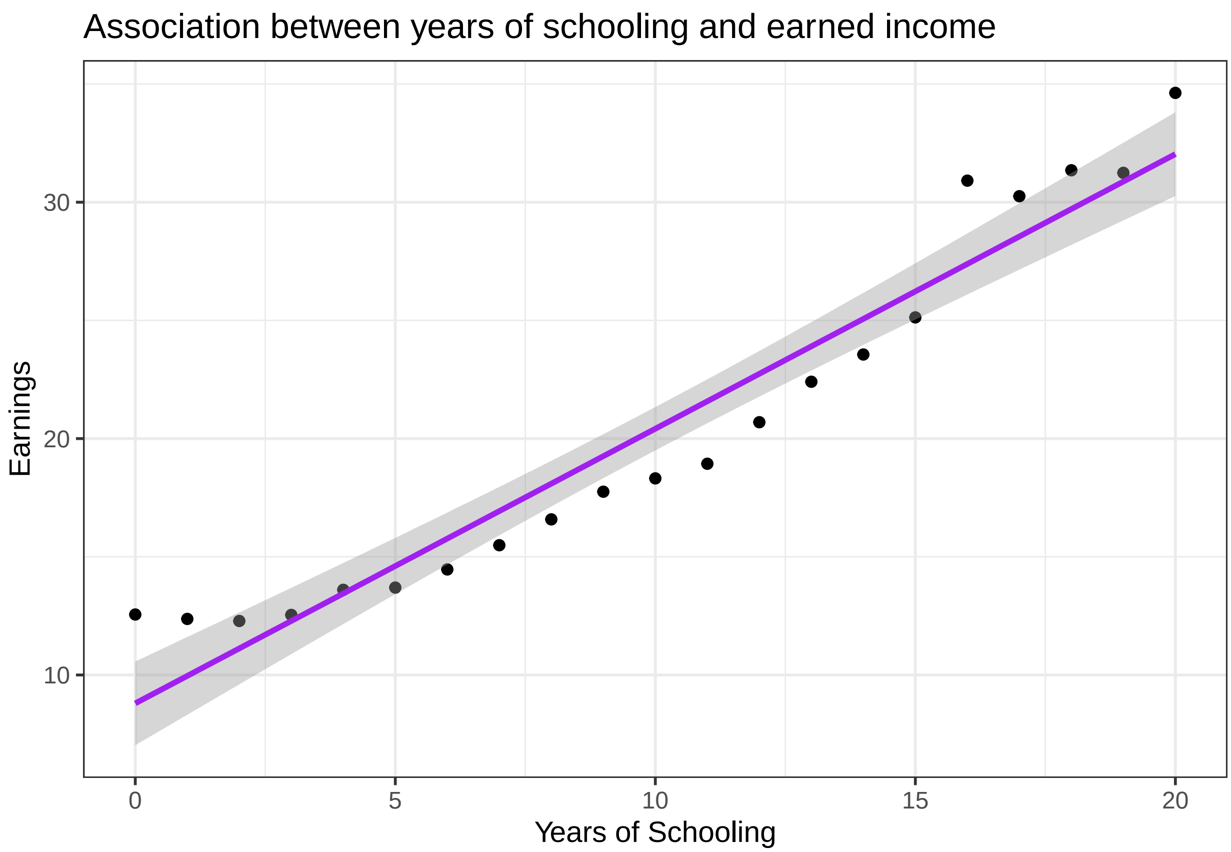
```r
# visualize a scatter plot of the data
schooling_data %>%
  ggplot(aes(x = schooling, y = earnings)) +
  geom_point() +
  labs(
    title = "Years of schooling versus earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```



At first blush, the data appear somewhat linear, so we visualize an OLS regression line on the plot.

```r
# visualize the regression line on a scatter plot
schooling_data %>%
  ggplot(aes(x = schooling, y = earnings)) +
  geom_point() +
  geom_smooth(color = "purple", method = lm) +
    labs(
    title = "Association between years of schooling and earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
# note: I do not add legends in this assignment.
# The graphics are fairly simple, which makes the time it would take to code
# a legend for each unnecessary.
```

The OLS regression appears to reasonably fit the data. It is noticeable, however, that over certain intervals the data are uniformly above or below the regression line. That means the line is systematically over- or underestimating the actual outcomes for stretches of the data. This make me question whether the relationship is linear.

b) Now run a fourth-order polynomial regression (i.e., including schooling, schooling squared, schooling to the third, and schooling to the fourth). Do those predictions meaningfully differ from the predictions coming from the linear regression?

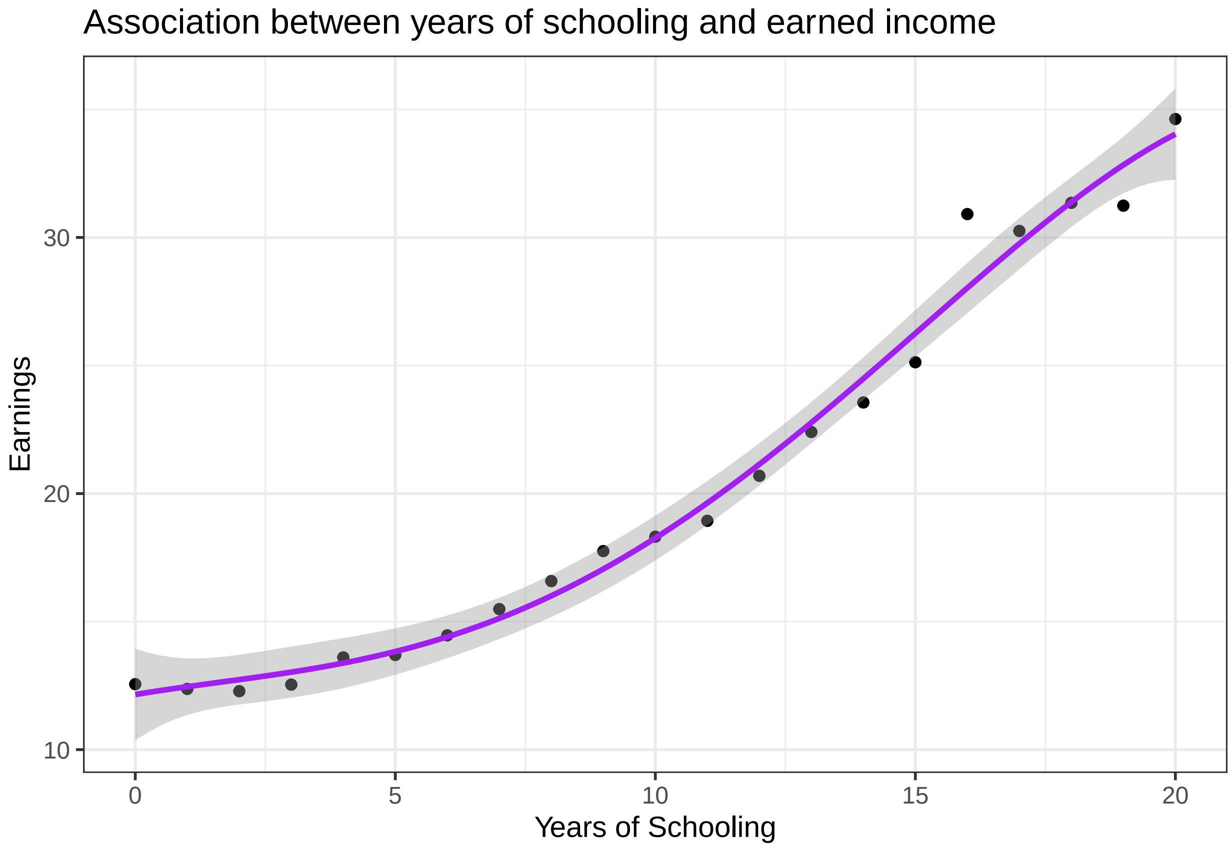
To further elucidate whether the variables are linearly associated, I run a fourth-order polynomial regression on earnings and years of schooling at first, second, third, and fourth powers. I show the regression parameters and a fit a regression line to a scatter plot below.

```
# fit a fourth-order regression and summarize the regression parameters
regression1 <- lm(data = schooling_data,
  earnings ~ poly(schooling, 4, raw=TRUE))
summary(regression1)
```

```
##
## Call:
## lm(formula = earnings ~ poly(schooling, 4, raw = TRUE), data = schooling_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58986 -0.45090 -0.02936  0.40589  2.88398
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      12.1513147  0.8425202  14.423 1.37e-10 ***
## poly(schooling, 4, raw = TRUE)1  0.3428344  0.6144678   0.558    0.585
## poly(schooling, 4, raw = TRUE)2 -0.0448806  0.1295554  -0.346    0.734
## poly(schooling, 4, raw = TRUE)3  0.0102209  0.0098493   1.038    0.315
## poly(schooling, 4, raw = TRUE)4 -0.0003049  0.0002442  -1.249    0.230
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.001 on 16 degrees of freedom
## Multiple R-squared:  0.9856, Adjusted R-squared:  0.982
## F-statistic: 274.4 on 4 and 16 DF,  p-value: 1.614e-14
```

```r
# visualize a fourth-order line of best fit to a scatter plot of the data
schooling_data %>%
  ggplot(aes(x = schooling, y = earnings)) +
  geom_point() +
  geom_smooth(color = "purple", method = lm,
              formula = y ~ poly(x, 4, raw=TRUE)) +
  labs(
    title = "Association between years of schooling and earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```



The predictions meaningfully differ from the linear regression. For for certain intervals of the data, a linear regression gave predictions that were systematically above or below the observed values. This is not the case for our fourth order polynomial regression.

In addition, there seems to be less error in the polynomial regression's predictions; the line (and therefore its predictions) appears uniformly closer to the observed values than the linear regression.

c) Now run different regressions for some different ranges of schooling. Do those lines look meaningfully different from the predictions you get from a single regression including all the data?

We now further test the linearity of the data by regressing different ranges of schooling. Comparing the linear regression against the scattered observations, there appear to be three distinct intervals (years 0 to 5, 6 to 15, and 16 to 20). The observations are systematically above, below, and above the predictions made by the line of best fit in each of these intervals.

In addition, the relationship between earnings and schooling appear to be roughly linear within each of those intervals. Earnings tend to increase very slowly, if at all, from years 0 to 4, then steadily from 5 to 16, and sporadically from 17 to 20. This makes sense because these years correspond to distinct phases of a person's education. Years 0 to 4 are elementary/primary school; 5 to 15 are intermediate, secondary, and college; and 16 to 20 are post-graduate. It is reasonable to think that associations with changes in income will be different across these three groups. (Post-graduates have very different employment situations than high school graduates, for instance.)

In light of this, I run a segmented linear regression, that is, I fit a separate model for data in each of these intervals. In addition, I show a table with relevant regression parameters and visualize all three lines of best fit on a scatter plot. I start by separating the data into these three groups.

```
#separate the data into three groups and append a group id
schooling_g1 <- schooling_data %>%
  filter(schooling < 6) %>%
  mutate(group = 1)

schooling_g2 <- schooling_data %>%
  filter(5 < schooling & schooling < 16) %>%
  mutate(group = 2)

schooling_g3 <- schooling_data %>%
  filter(15 < schooling) %>%
  mutate(group = 3)

# re-join them for later graphing
schooling_rejoined <- schooling_g1 %>%
  full_join(schooling_g2) %>%
  full_join(schooling_g3)
```

```
## Joining, by = c("schooling", "earnings", "group")
## Joining, by = c("schooling", "earnings", "group")
```

Defining three new data sets permits running a linear regression on each. Recombining them facilitates graphing on a common scatter plot. I perform these actions below.

```
# fit a separate linear regression for each group
# for group 1
regression2 <- lm(data = schooling_g1, earnings ~ schooling)
summary(regression2)
```

```
##
## Call:
## lm(formula = earnings ~ schooling, data = schooling_g1)
##
## Residuals:
##        1        2        3        4        5        6
##  0.40479 -0.05749 -0.42030 -0.44001  0.34694  0.16607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.15241    0.29958  40.564 2.21e-06 ***
## schooling    0.27567    0.09895   2.786   0.0495 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4139 on 4 degrees of freedom
## Multiple R-squared:  0.6599, Adjusted R-squared:  0.5749
## F-statistic: 7.762 on 1 and 4 DF,  p-value: 0.04951
```

```
# for group 2
regression3 <- lm(data = schooling_g2, earnings ~ schooling)
summary(regression3)
```

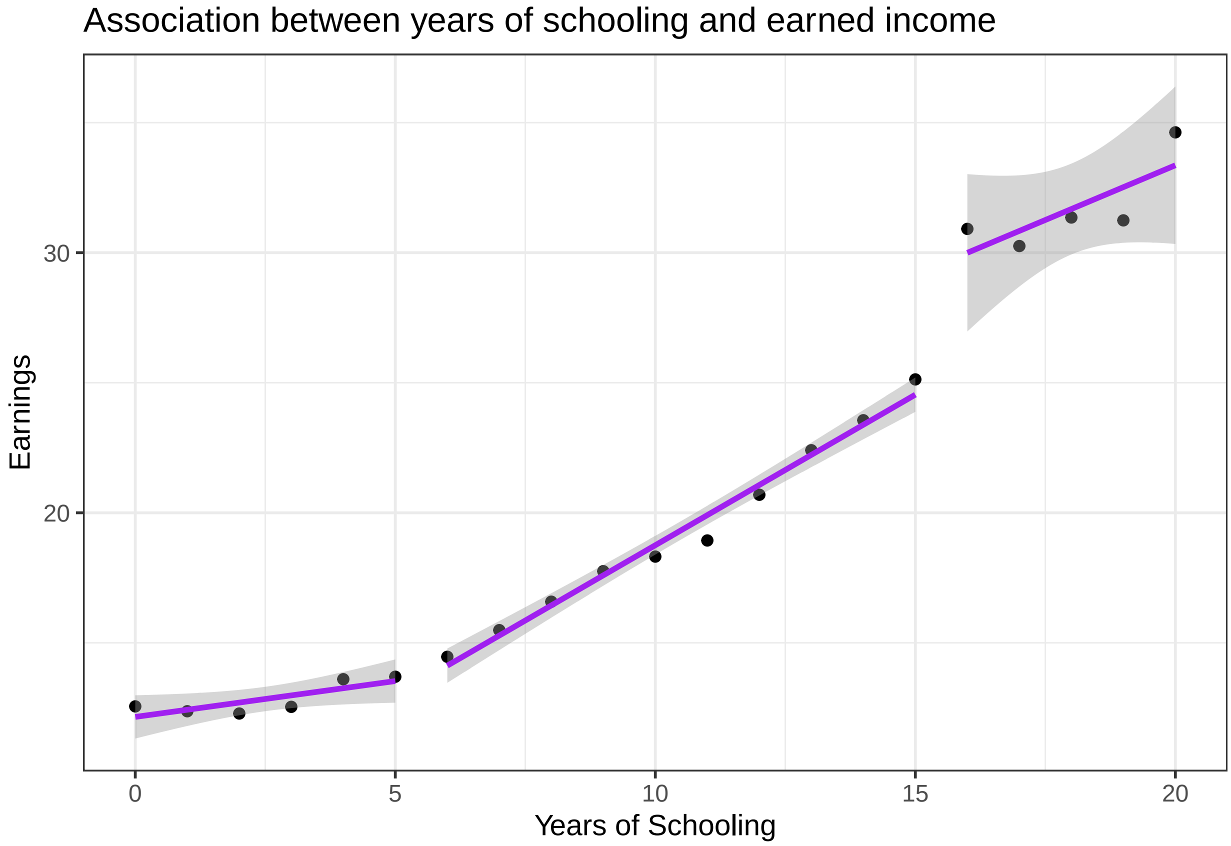
```
##
## Call:
## lm(formula = earnings ~ schooling, data = schooling_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9751 -0.2459  0.1655  0.2003  0.5860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.17716    0.58313   12.31 1.77e-06 ***
## schooling    1.15761    0.05357   21.61 2.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4866 on 8 degrees of freedom
## Multiple R-squared:  0.9832, Adjusted R-squared:  0.9811
## F-statistic:   467 on 1 and 8 DF,  p-value: 2.215e-08
```

```
# for group 3
regression4 <- lm(data = schooling_g3, earnings ~ schooling)
summary(regression4)
```

```
##
## Call:
## lm(formula = earnings ~ schooling, data = schooling_g3)
##
## Residuals:
##       1       2       3       4       5
##  0.9181 -0.5813 -0.3259 -1.2764  1.2656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.5377     7.0069   2.360   0.0994 .
## schooling     0.8411     0.3881   2.167   0.1188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.227 on 3 degrees of freedom
## Multiple R-squared:  0.6103, Adjusted R-squared:  0.4803
## F-statistic: 4.697 on 1 and 3 DF,  p-value: 0.1188
```

```
# plot segmented linear regression
schooling_rejoined %>%
  ggplot(aes(x = schooling, y = earnings)) +
  geom_point() +
  geom_smooth(aes(group = group), color = "purple", method = lm) +
  labs(
    title = "Association between years of schooling and earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



The lines do look meaningfully different from a regression using all of the data. First, the regression line is not continuous, increasing at different rates depending on the year of schooling. Second, the predictions appear, in general, to be more accurate. There is no consistent direction of error (positive or negative) as occurred in the first linear regression, and the observations appear to be much closer to the lines’ predictions, though the fourth order polynomial regression is similarly accurate in these respects. It is clear that segmenting the regression improves the accuracy of the predicted outcomes, though at the expense of some parsimony.

d) Does all this make you think the simple linear approach was reasonable or unreasonable?

It makes me think the simple linear approach was unreasonable. The simple linear regression was the most parsimonious since we only have to bear two regression parameters (beta and alpha) in mind. However, this parsimony comes at the expense of accuracy; the predictions made by the linear regression had greater error (which was the same direction over certain intervals) than the fourth order polynomial and segmented regressions.

It does not appear, therefore, that the relationship between years of schooling and earnings is approximately linear. While there is certainly a positive correlation, i.e. that earnings tends to increase with years of schooling on average, the magnitude of that change differs over time. Linearity implies an approximately constant rate of change. We observe inconstancy in the relationship between years of schooling and age, which indicates non-linearity, so a simple linear approach, though helpful, is inferior to higher order polynomial and segmented regression.

5.4) Similar to what we did with age and voter turnout, conduct some out-of-sample tests to evaluate your prediction strategy. Using only data from those with twelve years of schooling or less, see how well your different strategies from question 3 perform when predicting earnings for those with more than twelve years of schooling.

To validate my conclusion above, I now compare the efficacy of alternative strategies to the simple linear approach. I do this by conducting some out-of-sample tests, that is, I will fit one simple linear, one fourth order polynomial, and one segmented linear regression to only the observations with 12 or less years of schooling. I then test their predictions for observations with greater than 12 years of age. I show the relevant parameters for each regression via tables and then plot them on a common scatter plot.

Below I separate the data into two data frames (one for observations with less and one for those with more than 12 years of schooling) and then fit three regressions for the first group. I also recombine the two data sets to facilitate their graphing on a common scatter plot. I start by splitting the data into in sample and out of sample groups.

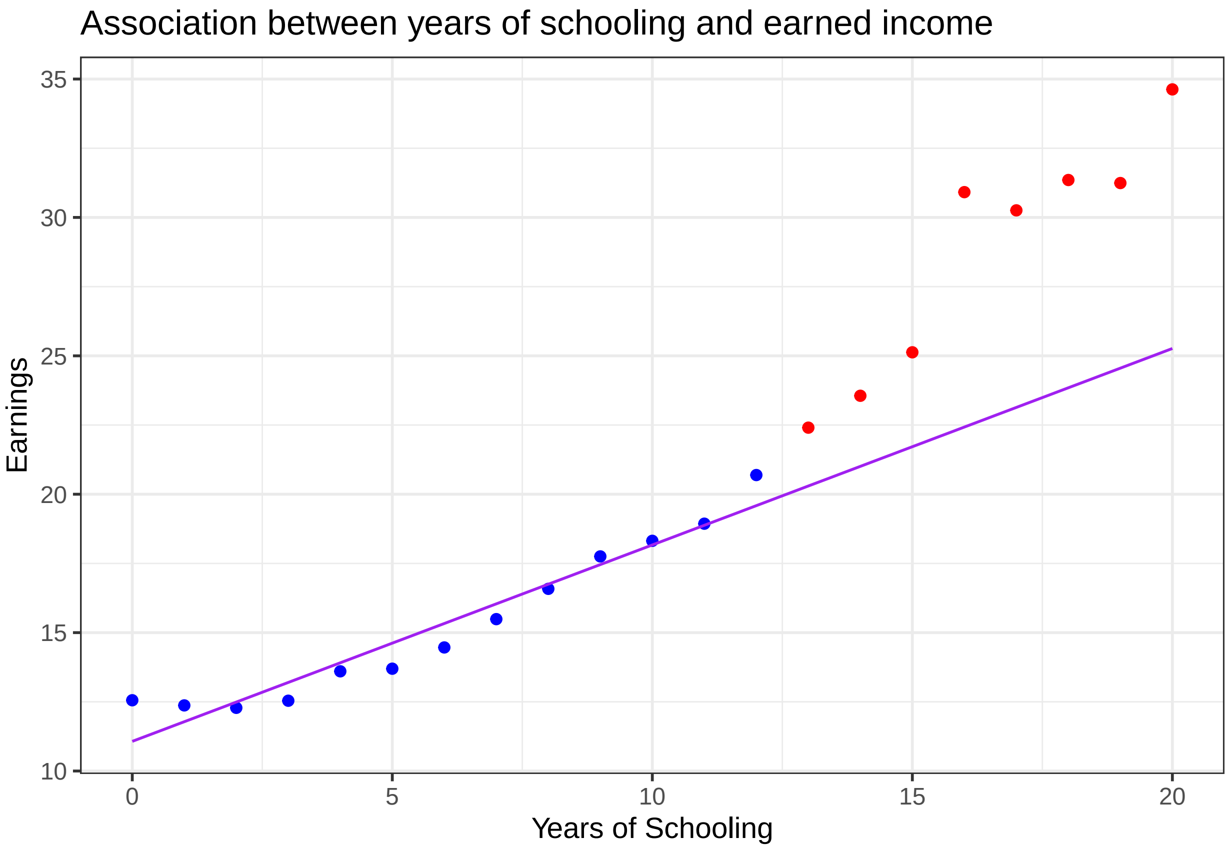
```
# create in and out of sample data
in_sample <- schooling_data %>%
  filter(schooling < 13)

out_of_sample <- schooling_data %>%
  filter(12 < schooling)
```

Now, I fit a simple linear model to the in sample data and test its predictions against the out of sample.

```
# fit a simple linear regression to in sample data
regression5 <- lm(data = in_sample, earnings ~ schooling)
summary(regression5)
```

```
##
## Call:
## lm(formula = earnings ~ schooling, data = in_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9228 -0.5511 -0.1664  0.2950  1.4859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.07127    0.39966   27.70 1.59e-11 ***
## schooling    0.70967    0.05652   12.56 7.29e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7625 on 11 degrees of freedom
## Multiple R-squared:  0.9348, Adjusted R-squared:  0.9288
## F-statistic: 157.7 on 1 and 11 DF,  p-value: 7.294e-08
```

```
# plot linear regression from in sample data showing out of sample data
ggplot() +
  geom_point(data = in_sample,
    aes(x = schooling, y = earnings), color = "blue") +
  geom_point(data = out_of_sample,
    aes(x = schooling, y = earnings), color = "red") +
  geom_function(fun = function(x) 0.70967 * x + 11.07127, color = "purple") +
   labs(
    title = "Association between years of schooling and earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```



It appears that a simple linear model fitted to the first twelve years of schooling is horribly predictive of out of sample data. While there is less error from years zero to 12 than there was along the simple linear regression of the entire data set, the observations are much higher than predicted when assuming linearity. This is further confirmation of my conclusion.

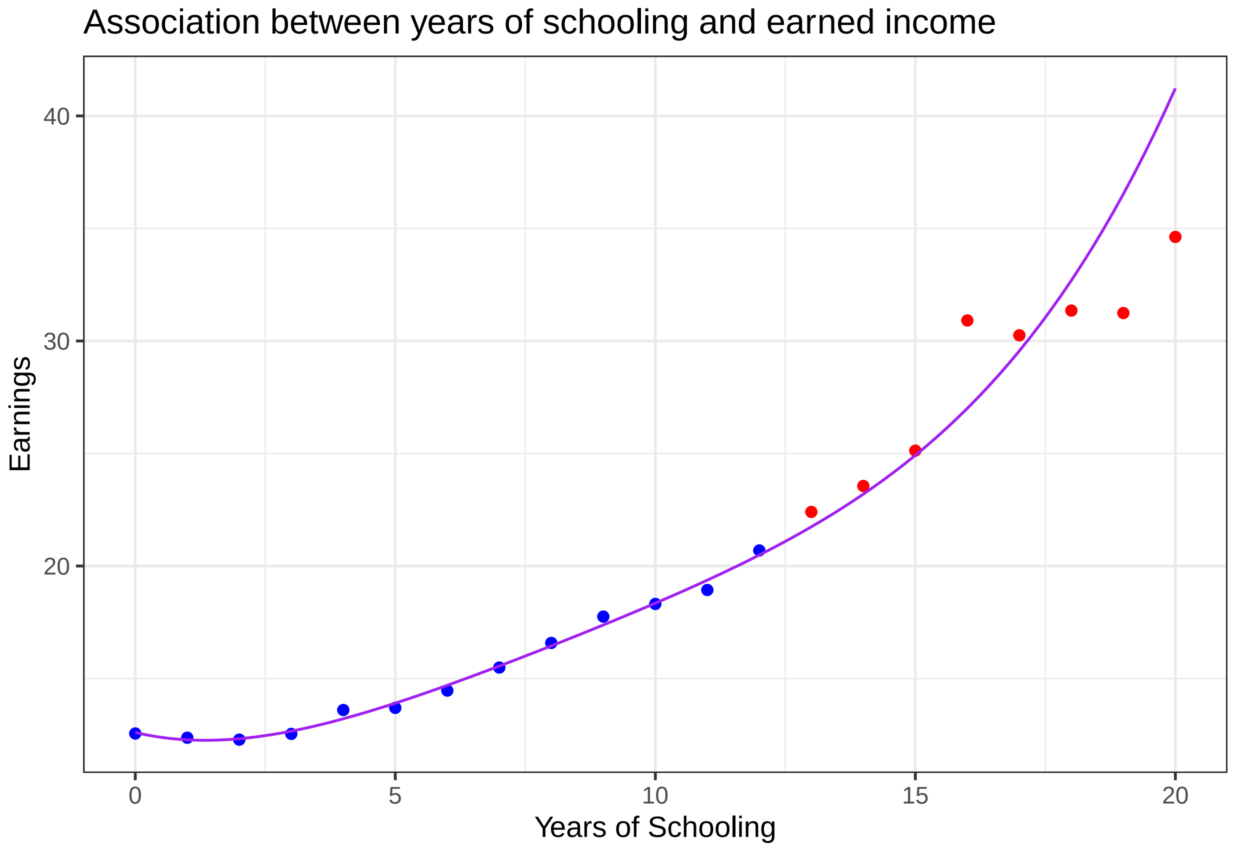
Below I fit a fourth order regression to the in sample data and test its predictions against the out of sample. In the visualizations below, in sample observations are plotted in blue and out of sample observations are plotted in red.

```
# fit fourth order regression to in sample data
regression6 <- lm(data = in_sample,
  earnings ~ poly(schooling, 4, raw=TRUE))
summary(regression6)
```

```
##
## Call:
## lm(formula = earnings ~ poly(schooling, 4, raw = TRUE), data = in_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43660 -0.12337 -0.04198  0.13032  0.39126
##
```

```
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     12.6042321  0.2711583  46.483 5.07e-11 ***
## poly(schooling, 4, raw = TRUE)1 -0.5514305  0.3414008  -1.615   0.1449
## poly(schooling, 4, raw = TRUE)2  0.2408554  0.1230226   1.958   0.0859 .
## poly(schooling, 4, raw = TRUE)3 -0.0185683  0.0157127  -1.182   0.2712
## poly(schooling, 4, raw = TRUE)4  0.0005741  0.0006492   0.884   0.4024
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2904 on 8 degrees of freedom
## Multiple R-squared:  0.9931, Adjusted R-squared:  0.9897
## F-statistic: 288.7 on 4 and 8 DF,  p-value: 1.114e-08
```

```
# plot fourth-order regression from in sample data showing out of sample data
ggplot() +
  geom_point(data = in_sample,
    aes(x = schooling, y = earnings), color = "blue") +
  geom_point(data = out_of_sample,
    aes(x = schooling, y = earnings), color = "red") +
  geom_function(fun = function(x) -0.5514305 * x +
                0.2408554 * x^2 +
                -0.0185683 * x^3 +
                0.0005741 * x^4 +
                12.6042321,
              color = "purple") +
  labs(
    title = "Association between years of schooling and earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```



The fourth order polynomial regression has much less error than the simple linear model. While the out of sample predictions are not amazingly accurate, they are much closer to the observed outcomes than the simple linear model. In addition, they are not systematically above or below observed outcomes as they were with the simple linear model. This suggests a polynomial regression might be an effective strategy.

Finally, I separate the in sample data into two groups (group from years 0 to 3 and group 2 from years 4 to 12) and fit a linear regression to each segment. I then test its predictions (that of the regression from group 2) against the out of sample data. It is all visualized on a common scatter plot. (I pick these intervals to regress for similar reasons as above; the rate of change appears to differ between years 3 and 4, though perhaps 4 and 5 would also serve as credible cut offs.)

```
#separate in_sample data into two groups and append group number
in_sample_g1 <- in_sample %>%
  filter(schooling < 4) %>%
  mutate(group = 1)

in_sample_g2 <- in_sample %>%
  filter(3 < schooling) %>%
  mutate(group = 2)

# recombine them for graphing
in_sample_segmented <- in_sample_g1 %>%
  full_join(in_sample_g2)
```

```
## Joining, by = c("schooling", "earnings", "group")
```

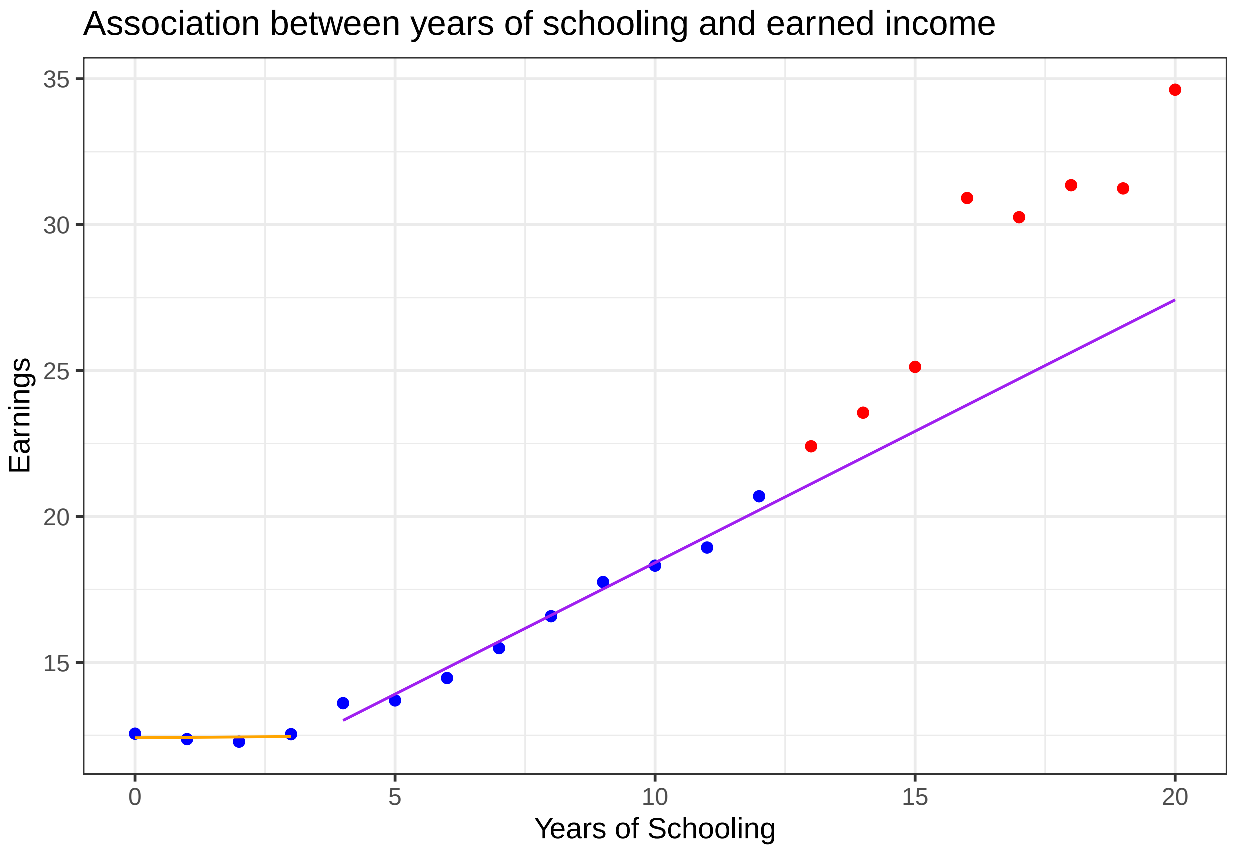
```r
# fit a segmented regression to in_sample data
# for group 1
regression7 <- lm(data = in_sample_g1, earnings ~ schooling)
summary(regression7)
```

```
##
## Call:
## lm(formula = earnings ~ schooling, data = in_sample_g1)
##
## Residuals:
##        1        2        3        4
##  0.09846 -0.07410 -0.14719  0.12283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.45874    0.13483  92.406 0.000117 ***
## schooling   -0.01405    0.07207  -0.195 0.863456
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1611 on 2 degrees of freedom
## Multiple R-squared:  0.01864,    Adjusted R-squared:  -0.472
## F-statistic: 0.038 on 1 and 2 DF,  p-value: 0.8635
```

```r
# for group 2
regression8 <- lm(data = in_sample_g2, earnings ~ schooling)
summary(regression8)
```

```
##
## Call:
## lm(formula = earnings ~ schooling, data = in_sample_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3811 -0.2258 -0.1003  0.2380  0.5907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.40808    0.41152   22.86 7.76e-08 ***
## schooling    0.90080    0.04895   18.40 3.47e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3792 on 7 degrees of freedom
## Multiple R-squared:  0.9797, Adjusted R-squared:  0.9769
## F-statistic: 338.6 on 1 and 7 DF,  p-value: 3.469e-07
```

```r
# plot segmented linear regression showing out of sample data
ggplot() +
  geom_point(data = in_sample,
    aes(x = schooling, y = earnings), color = "blue") +
  geom_point(data = out_of_sample,
    aes(x = schooling, y = earnings), color = "red") +
  geom_segment(aes(x = 0, y = 12.41659, xend = 3, yend = 12.45874),
               color = "orange") +
  geom_segment(aes(x = 4, y = 13.01128, xend = 20, yend = 27.42408),
               color = "purple") +
   labs(
    title = "Association between years of schooling and earned income",
    x = "Years of Schooling",
    y = "Earnings") +
  scale_fill_brewer(palette = "Set3") +
  theme_bw()
```



```
# note: the group 1 regression line is shown in orange; group 2's is purple.
```

The segmented regression is not a desirable strategy either. Conceptually, if segments are found to be useful, i.e. the rate of change is different for different intervals of x, then it makes little sense to assume linearity across the intervals of out of sample predictions. H

However, this assumption is required for prediction, and the relative inaccuracy of the model's predictiveness is the result. The segmented linear regression is less off the mark than the simple linear model, but its predictions are still systematically too low. The polynomial regression is better on both fronts and it thus the preferred strategy between these three alternatives.

5.5). Drop one observation, run a regression to try to predict the outcome for that missing observation, and see how far you were. Repeat this for each observation in the data set (you

should be able to do this with a loop) and average your errors. Try different strategies to see which one gives you the best out-of-sample predictions.

Though we can now prefer a strategy of polynomial regression over simple or segmented linear models, we can verify this preference even more rigorously. Below I loop through every observation, drop it (one-by-one) from the data, fit a model to the remaining observations, and then compare the mean error (i.e. difference between actual and predicted outcome) for that year of schooling. I do this for a simple linear and second, third, fourth, and fifth order polynomial regressions. I report and compare each of their standard errors to select a preferred model.

```r
# calculate average error using a linear and first through fifth order regressions

# initialize an empty character vector
errors1 <- c()

# initialize an empty character vector
errors2 <- c()

# initialize an empty character vector
errors3 <- c()

# initialize an empty character vector
errors4 <- c()

# initialize an empty character vector
errors5 <- c()

# for loop
for (i in c(1:21)){

  # split the data into training and testing
  training <- schooling_data[-i, ]
  testing <- schooling_data[i, ]

  # simple linear model
  linear = lm(earnings ~ schooling, data = training)
  temp <- (testing$earnings - predict(linear, testing))^2
  errors1 <- c(errors1, temp)

  # second order model
  second_order = lm(earnings ~ poly(schooling, 2, raw=TRUE), data = training)
  temp <- (testing$earnings - predict(second_order, testing))^2
  errors2 <- c(errors2, temp)

  # third order model
  third_order = lm(earnings ~ poly(schooling, 3, raw=TRUE), data = training)
  temp <- (testing$earnings - predict(third_order, testing))^2
  errors3 <- c(errors3, temp)

  # fourth order model
  fourth_order = lm(earnings ~ poly(schooling, 4, raw=TRUE), data = training)
  temp <- (testing$earnings - predict(fourth_order, testing))^2
  errors4 <- c(errors4, temp)

  # sith order model (did you catch that one?)
  fifth_order = lm(earnings ~ poly(schooling, 5, raw=TRUE), data = training)
  temp <- (testing$earnings - predict(fifth_order, testing))^2
  errors5 <- c(errors5, temp)

}

# She the mean squared error for each model.
mean(errors1)
```

```
## [1] 4.645274
```

```
mean(errors2)
```

```
## [1] 1.293558
```

```
mean(errors3)
```

```
## [1] 1.206018
```

```
mean(errors4)
```

```
## [1] 1.476035
```

```
mean(errors5)
```

```
## [1] 2.755554
```

From the results above, it appears that a third order polynomial regression minimizes the total squared error. Extra dimensionality (i.e. higher order regression) appears to capture more of the relationship than a simple linear model. However, the benefits of extra terms appears to wear off after the third order regression, probably due to over fitting. Though the fourth and fifth order functions are likely highly accurate between years zero and 12 of schooling, they are much less effective at predicting out of sample of observerations.

In general, this phenomenon (over fitting) occurs because the extra terms just by chance happen to correlate with the observed outcomes in our training sample, but they do not capture anything real about the relationship between schooling and earnings, reading every little noisy bump in the data as meaningful. They mistake noise for signal. As a result, they are less accurate in predicting out of sample data. The third order polynomial regression most effectively balances the gains from extra dimensions with the error introduced from overfitting.