Homework 13

Alizeh Azhar

11/24/2021

Problem 1 (20 points)

Before beginning the problems, please load the `prestige` dataset into R using `prestige_df <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/carData/Prestige.csv")`- alternatively, you can download and install the `carData` package and load the dataset using `data(Prestige)`. You can read more about this dataset at https://vincentarelbundock.github.io/Rdatasets/doc/carData/Prestige.html.

(a) How many variables are there in this dataset? How many are numerical, and how many are categorical? (1 pt)

```
#From the description, the first column contains the row names
prestige_df <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/carData/Prestige.csv",
                        row.names = 1)
str(prestige_df)
```

```
## 'data.frame':    102 obs. of  6 variables:
##  $ education: num  13.1 12.3 12.8 11.4 14.6 ...
##  $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
##  $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
##  $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
##  $ type     : chr  "prof" "prof" "prof" "prof" ...
```

The dataset has 6 variables. Only one variable is categorical, and 5 of them are numeric.

(b) Fit a model predicting prestige from education and type. (1 pt)

```
prestige_df$type= factor(prestige_df$type, levels = c("bc","prof","wc"))
prestige_model <- lm(prestige~education+type, data = prestige_df)
summary(prestige_model)
```

```
##
## Call:
## lm(formula = prestige ~ education + type, data = prestige_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.410  -5.508   1.360   5.694  17.171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.6982     5.7361  -0.470   0.6392
## education     4.5728     0.6716   6.809 9.16e-10 ***
## typeprof      6.1424     4.2590   1.442   0.1526
## typewc       -5.4585     2.6907  -2.029   0.0453 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.814 on 94 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7975, Adjusted R-squared:  0.791
## F-statistic: 123.4 on 3 and 94 DF,  p-value: < 2.2e-16
```

(c) State the model's adjusted R^2. (1 pt)

```
summary(prestige_model)$adj.r.squared
```

```
## [1] 0.7910444
```

The model's adjusted R^2 is 0.791

(d) Predict the prestige of a white collar position an average of 12 years education. (1 pt)

```
-2.6982+4.5728*12-5.4585
```

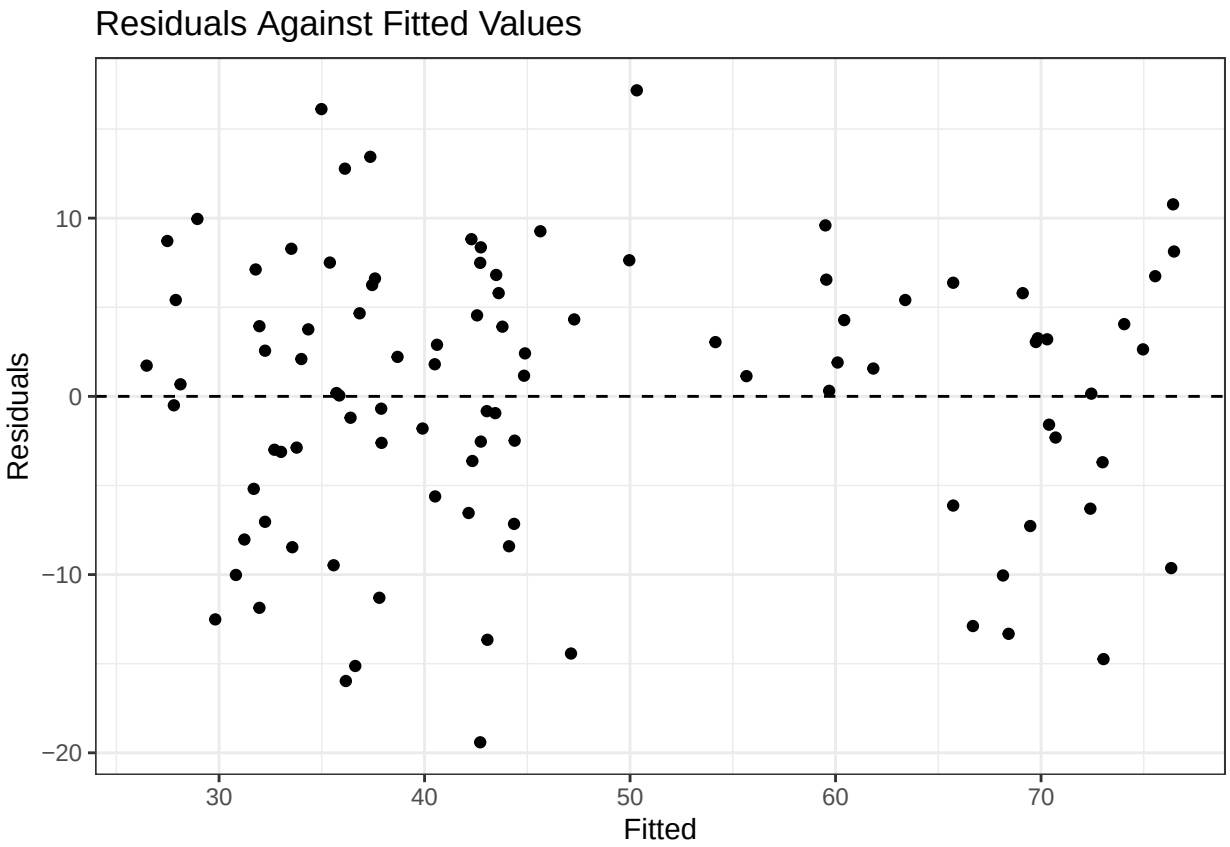
```
## [1] 46.7169
```

(e) Run a hypothesis test for the variable education at $\alpha = 0.05$. Be sure to include the appropriate hypotheses, report the test statistic and p-value, and clearly interpret your decision in the context of the problem. (4 pts)
H_0 : The coeffcient of educaton is equal to zero (not significant) vs H_1 : The coeffcient of educaton is not equal to zero (significant)
From the above results, the education coefficient (t=6.809, p= 0.000) has a p- value less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the coeffcient of educaton is not equal to zero. i.e., Education is a significant predictor.

(f) Interpret the point estimate for the coefficient on education. Is increased education expected to lead to an increase or decrease in perceived prestige? How do you know? (2 pts)
The point estimate of the coefficient of education is 4.5728. An unit increase in education will lead to an increase in percieved prestige by 4.5728. The increase is significant (p=0.000), hence education is a significant predictor. Moreover, There is a positive relationship between education and percieved prestige.

(g) Plot the residuals for this model against the fitted values. Is this model appropriate for the data?Explain. (3 pts)

```
library(ggplot2)
ggplot(data = prestige_model, aes(x=.fitted, y= .resid))+
  geom_point()+
  geom_hline(yintercept = 0, linetype= "dashed")+
  xlab("Fitted")+
  ylab("Residuals")+
  ggtitle("Residuals Against Fitted Values")+
  theme_bw()
```



This model is appropriate for the data because the negative and positive residuals are almost equal. There are no patterns

(h) Consider the variable type. What is the reference level? Interpret the point estimate for `typeprof`. (2pts)
The reference level in this model is 'bc'. The point estimate for 'typeprof' is 6.1424. The expected number of Professional, Managerial, and Technical('prof') is 6.1424 more than the Blue Collar('bc')

(i) Fit models using education alone and type alone (2 pts). Out of the three models you have picked, which one describes the data best? (1 pt) Why? (2 pts)

```
education_model <- lm(prestige~education, data = prestige_df)
summary(education_model)$r.squared
```

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

```
type_model <- lm(prestige~type, data = prestige_df)
summary(type_model)$r.squared
```

```
## [1] 0.6976287
```

The first model which has both 'education' and 'type' as independent variable describes the data best. It has the largest value of R^2 compared to the models having education alone and type alone.