SDGB 7844 HW 2: Vintage Wine

Darshan Patel

2/7/2019

Question 1: Read the posted article "Bordeaux wine vintage quality and weather," by Ashenfelter, Ashmore, and LaLorne (CHANCE, 1995). Three regression models are considered in this article. Answer the following questions.

(a) What is a wine "vintage"?

Answer: A wine "vintage" is a way to describe aged wine; it tells the year/place in which the wine was produced. According to the paper, "bad" vintages are young and usually overpriced whereas "good" vintages may be underpriced and old.

(b) What is the response variable for the three models described in this paper?

Answer: The response variable for the three models described in this paper is the (log or regular) price of vintages.

Now download the data in "wine.dat". This is some of the data the authors used to fit their models. The columns are vintage (`VINT`), log of average vintage price relative to 1961 (`LPRICE2`), rainfall in the months preceding the vintage in mL (`WRAIN`), average temperature over the growing season in °C (`DEGREES`), rainfall in September and August in mL (`HRAIN`) and age of wine in years (`TIME_SV`).

NOTE: the average temperature in September is not available in this data set so the third regression model from the paper cannot be fit.

```
# Import tidyverse
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------- tidyverse 1.2
## v ggplot2 3.1.0.9000     v purrr   0.2.5
## v tibble  2.0.1          v dplyr   0.7.8
## v tidyr   0.8.2          v stringr 1.3.0
## v readr   1.1.1          v forcats 0.3.0
## Warning: package 'tibble' was built under R version 3.4.4
## Warning: package 'tidyr' was built under R version 3.4.4
## Warning: package 'purrr' was built under R version 3.4.4
## Warning: package 'dplyr' was built under R version 3.4.4
## -- Conflicts -------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Read in data and note the delimiter and NA values
df <- read_delim("wine.dat", delim = ',', na = '.')
```

```
## Parsed with column specification:
## cols(
##   VINT = col_integer(),
##   LPRICE2 = col_double(),
##   WRAIN = col_integer(),
##   DEGREES = col_double(),
```

```
##   HRAIN = col_integer(),
##   TIME_SV = col_integer()
## )
```

(c) Which values of LPRICE2 are missing and, according to the article, why have they been omitted?

Answer:

```
# Find data that has missing LPRICE2
df[is.na(df$LPRICE2),]
```

```
## # A tibble: 11 x 6
##     VINT LPRICE2 WRAIN DEGREES HRAIN TIME_SV
##    <int>   <dbl> <int>   <dbl> <int>   <int>
##  1  1954      NA   430    15.4   180      29
##  2  1956      NA   440    15.6   140      27
##  3  1981      NA   535    17.0   111       2
##  4  1982      NA   712    17.4   162       1
##  5  1983      NA   845    17.4   119       0
##  6  1984      NA   591    16.5   119      -1
##  7  1985      NA   744    16.8    38      -2
##  8  1986      NA   563    16.3   171      -3
##  9  1987      NA   452    17.0   115      -4
## 10  1988      NA   808    17.1    59      -5
## 11  1989      NA   443    NA      82      -6
```

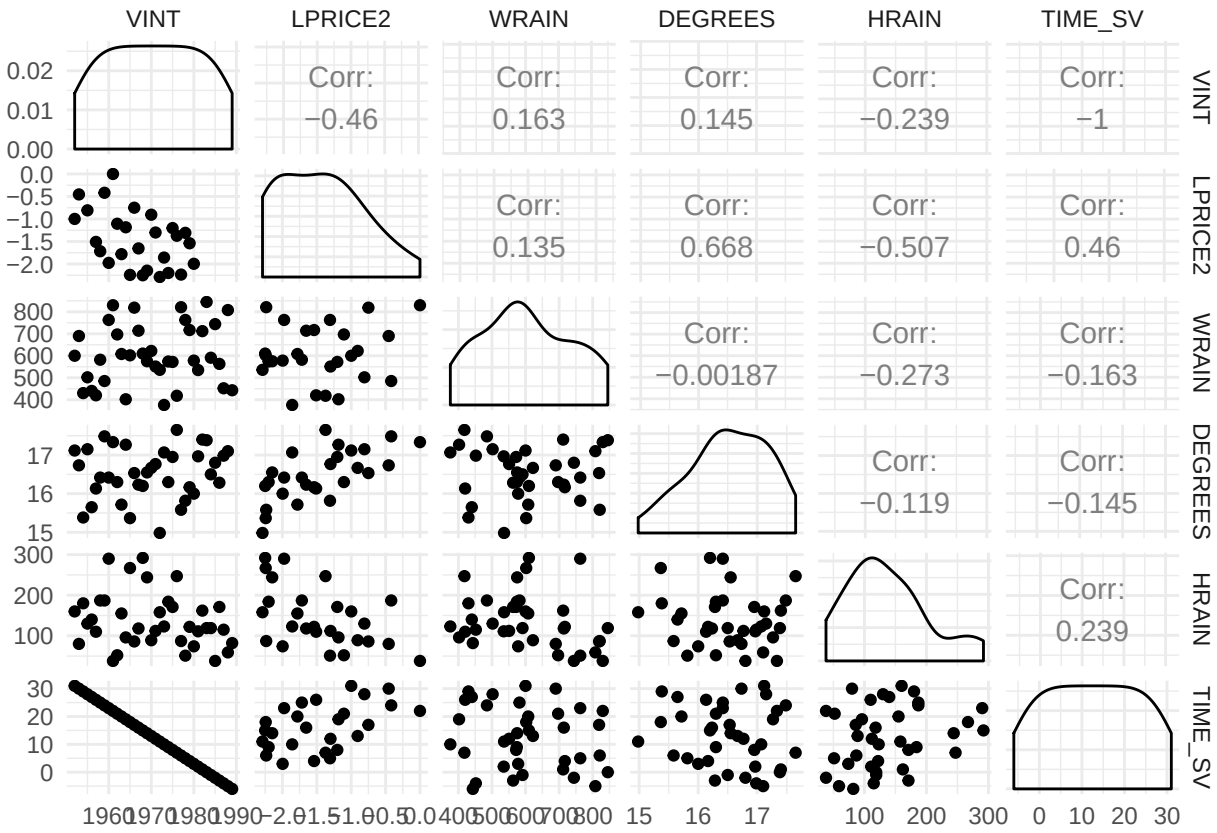
```
# Save count of missing data rows
nulls <- nrow(df[is.na(df$LPRICE2),])
```

Ten of the 11 observations that have missing values for LPRICE2 is shown above. These values have been omitted due to one of two reasons. The two vintages from the 1950s were omitted because they were no longer being sold at the time of the publication of the article. The other nine vintages, from 1981 to 1989, were omitted because the weather conditions that created these vintages were abnormal compared to the typical growing season for vintages. The increased temperature at the time created wine that was excellent yet young. It did not make sense to add vintages that were created using abnormal conditions.

(d) Make a scatterplot matrix of the variables (explanatory and response) included in the models? Discuss findings.

Answer:

```
# Import GGally
library(GGally)
# Create scatterplot matrix of all variables
ggpairs(df) + theme_minimal()
```



`VINT` and `TIME_SV` are linearly correlated, which makes sense because `TIME_SV` only measures the age of the wine (in years) at the time of the study, and `VINT` gives the year the wine was made. `LPRICE2` appears to have a linear correlation with `VINT`, `DEGREES` and `TIME_SV`. No other variables appear to look correlated.

(e) Fit the two regression models from the paper. Which is the best regression model? Justify the answer and include relevant output (let $\alpha = 0.05$). Did you choose the same model as the authors? Answer:

```
# Model 1 - Log price vs vintage
model_vint_price <- lm(LPRICE2 ~ VINT, df)
summary(model_vint_price)
```

```
##
## Call:
## lm(formula = LPRICE2 ~ VINT, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8545 -0.4788 -0.0718  0.4562  1.2457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.23162   26.87182   2.539   0.0177 *
## VINT        -0.03543    0.01366  -2.593   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5745 on 25 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.1804
```

```
## F-statistic: 6.725 on 1 and 25 DF,  p-value: 0.01567
```

At an α level of 0.05, it appears that the coefficient estimate of β_{VINT} is not statistically significant because its p-value is 0.0157 which is greater than the significance level.

```
# Model 2 - Log price vs weather vars
model_weather_price <- lm(LPRICE2 ~ WRAIN + DEGREES + HRAIN, df)
summary(model_weather_price)
```

```
##
## Call:
## lm(formula = LPRICE2 ~ WRAIN + DEGREES + HRAIN, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62819 -0.17924  0.02273  0.21987  0.62861
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.344e+01  1.969e+00  -6.827 5.82e-07 ***
## WRAIN        1.282e-03  5.765e-04   2.224  0.03628 *
## DEGREES      7.123e-01  1.088e-01   6.549 1.11e-06 ***
## HRAIN       -3.624e-03  9.646e-04  -3.757  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3436 on 23 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.7407, Adjusted R-squared:  0.7069
## F-statistic:  21.9 on 3 and 23 DF,  p-value: 6.246e-07
```

It is found that at the α level of 0.05, the coefficient estimates for β_{WRAIN}, β_{DEGREES} and β_{HRAIN} are statistically significant because their p-values are less than the significance level. This signifies that the null hypotheses, namely, $\beta_{\text{WRAIN}} = 0$, $\beta_{\text{DEGREES}} = 0$ and $\beta_{\text{HRAIN}} = 0$ can be rejected.

The better regression model is the second model, with `WRAIN`, `DEGREES` and `HRAIN`. This is because the model has an adjusted R^2 statistic of 0.7069, meaning that 70.69% of the variation in the log price of vintages were explained by the three variables. This contrasts with the first model that has a R^2 statistic of 0.212, meaning that only 21.2% of the variation in the log price of vintage was explained by `VINT`. In addition, the RSE in the second model is lower than the RSE in the first model. This is the opposite choice from the authors' choice. The authors' choice of the best model was the one that used the year of the vintages.

(f) What is the sample size for the models?

Answer:

```
# Print sample size
sample_size <- nrow(df) - nulls
sample_size
```

```
## [1] 27
```

The sample size of the models are 27. 11 observations (from the original 38) were removed from the size of the dataset due to missingness.

(g) Write out the regression equation of the model in part (e). Include the units of measurement. Interpret the partial slopes and the y-intercept. Does the y-intercept have a practical interpretation?

Answer:

For the first model,

```
# Print coefficients of first model
summary(model_vint_price)$coef[,1]
```

```
## (Intercept)        VINT
## 68.23161768 -0.03542956
```

$$Y = 68.231 - 0.0354X$$

where Y is the log price of the vintage in dollars and X is the year the vintage was made in, in year. According to this model, an increase in 1 in the vintage (or year a wine is made) is associated with a decrease of 0.03543 in the log price of the vintage. Furthermore, the log price of vintage is 68.213 in the year 0. This y-intercept has no practical interpretation here.

For the second model,

```
# Print coefficients of second model
summary(model_weather_price)$coef[,1]
```

```
##   (Intercept)         WRAIN       DEGREES         HRAIN
## -13.444330639   0.001282014   0.712317651  -0.003624186
```

$$Y = -13.444 + 0.001X_1 + 0.712X_2 - 0.003X_3$$

where Y is the log of average price of the vintage (in dollars), X_1 is the rainfall in the months preceeding the vintage (in mL), X_2 is the average temperature over the growing season (in °C) and X_3 is the rainfall in September and August (in mL). According to this model, an increase of 1 mL in the average rainfall in the months preceding the vintage, while keeping average temperature and rainfall in September and August constant, is associated with a 0.001 increase in the log of average price of the vintage in dollars. Likewise, an increase of 1° C is associated with an increase of 0.712 in the log average price of the vintage in dollars, while keeping the other two rainfall variables constant. In addition, an increase of 1 mL in the average rainfall in September and August is associated with a a decrease of 0.003 in the log average price of the vintage, while keeping the other two weather variables constant. Now, when the amount of rainfall in the months preceding the vintage is 0 mL, and the amount of rainfall in September and August is 0 mL, AND the average temperature over the growing season is 0°C, then the log average price of vintages is −13.444. This has no practical interpretation here.

(h) Make a table with the following statistics for both models: SSE, RMSE, PRESS, and $\text{RMSE}_{\text{jackknife}}$. Compare the relevant statistics. Based on this information, would you change your answer to part (e)? Justify your answers.

Answer:

```
# Import DAAG
library(DAAG)
```

```
## Loading required package: lattice
```

```
# Model 1 Calculations
sse1 <- summary(model_vint_price)$sigma^2 * (sample_size - 1)
rmse1 <- sqrt(sse1 / (sample_size - (1 + 1)))
press1 <- press(model_vint_price)
rmsejk1 <- sqrt(press1 / (sample_size - (1 + 1)))
# Model 2 Calculations
sse2 <- summary(model_weather_price)$sigma^2 * (sample_size - 1)
rmse2 <- sqrt(sse2 / (sample_size - (3 + 1)))
press2 <- press(model_weather_price)
```

```
rmsejk2 <- sqrt(press2 / (sample_size - (3 + 1)))
# Store and print statistics in a nice dataframe
stats_df <- data.frame("Model 1" = c(sse1, rmse1, press1, rmsejk1),
                       "Model 2" = c(sse2, rmse2, press2, rmsejk2))
rownames(stats_df) <- c("SSE", "RMSE", "PRESS", "RMSE_JACKKNIFE")
stats_df
```

```
##                   Model.1   Model.2
## SSE            8.5809185 3.0687767
## RMSE           0.5858641 0.3652740
## PRESS          9.3955689 3.9173019
## RMSE_JACKKNIFE 0.6130438 0.4126954
```

According to this table, the SSE, RMSE, PRESS and $\text{RMSE}_{\text{jackknife}}$ statistics from the second model are lower than the oens from the first model. This means that the second model made fewer errors than the first model. Hence the answer given to part (e) remains the same, i.e., the model using `WRAIN`, `DEGREES` and `HRAIN` to predict `LPRICE2` is better than the one using only `VINT`.

(i) Could we use these regression models to predict quality for wines produced in 2005? Justify your answer.

Answer: It is not advisable to use these regression models to predict quality for wines produced in 2005. This is because 2005 falls way beyond the scope of this study, which ends in

```
tail(df$VINT, 1)
```

```
## [1] 1989
```

Predicting the quality of wines produced 16 years in the future from when this dataset has data on is considered extrapolating. There is no clear information whether the relationships found above will hold 16 years into the future.

Question 2: Model the prestige level of occupations using variables such as education and income levels. This data was collected in 1971 by Statistics Canada (the Canadian equivalent of the US Census Bureau or the National Bureau of Statistics of China). The data is in the file "prestige.dat" and the variables are described below:

```
# Data description table
prestige_col_df <- data.frame("variable" = c("prestige (y)", "education", "income",
                                             "women", "census", "type"),
                              "description" = c("Pineo-Porter prestige score for occupation,
                                                from a social survey conducted in the mid-1960s",
                                                "average education of occupational incumbents,
                                                years, in 1971",
                                                "average income of incumbents, dollars, in 1971",
                                                "percentage of incumbents who are women",
                                                "Canadian Census occupational code",
                                                "type of occupation: bc = blue collar,
                                                prof = professional / managerial / technical,
                                                wc = white collar"))
# Import knitr
library(knitr)
```

```r
# Print data description table
kable(prestige_col_df, caption = "Column Description")
```

Table 1: Column Description

variable	description
prestige (y)	Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s
education	average education of occupational incumbents, years, in 1971
income	average income of incumbents, dollars, in 1971
women	percentage of incumbents who are women
census	Canadian Census occupational code
type	type of occupation: bc = blue collar, prof = professional / managerial / technical, wc = white collar

(a) Do some internet research and write a short paragraph in your own words about how the Pineo-Porter prestige score is computed. Include the reference(s) you used. Do you think this score is a reliable measure? Justify your answer.

Answer: In a social study done by Pineo and Porter in Canda during the 1960s, participants were asked to rank occupations in how they viewed it in terms of respect. At the same time, the US-based National Opinion Research Center also conducted a similar study where people were asked to rank occupations. The results from these two surveys were compared and used to compute the Pineo-Porter prestige score by matching the occupations to census data and then regressing prestige ranking on education and income using a subset of occupations. Then the prestige score was found by forming predictions using the regression equation for all the occupations.

Source: http://homes.chass.utoronto.ca/~boydmon/research_papers/SES_scales/Recasting_Rethinking.pdf.

I believe this is a reliable measure because education and income does play a big role in how jobs are looked at by people. People tend to view high-paying jobs, ones that require higher education, as better than low-paying jobs that only require the minimalist amount of education.

(b) Create a scatterplot matrix of all the **quantitative* variables. Use a different symbol for each profession type: no type (`pch=3`), "bc" (`pch=6`), "prof" (`pch=8`) and "wc" (`pch=0`) when making the plot. For the remainder of this question, use the explanatory variables: income, education, and type. Does restricting our regression to only these variables make sense given the explanatory analysis? Justify your answer.

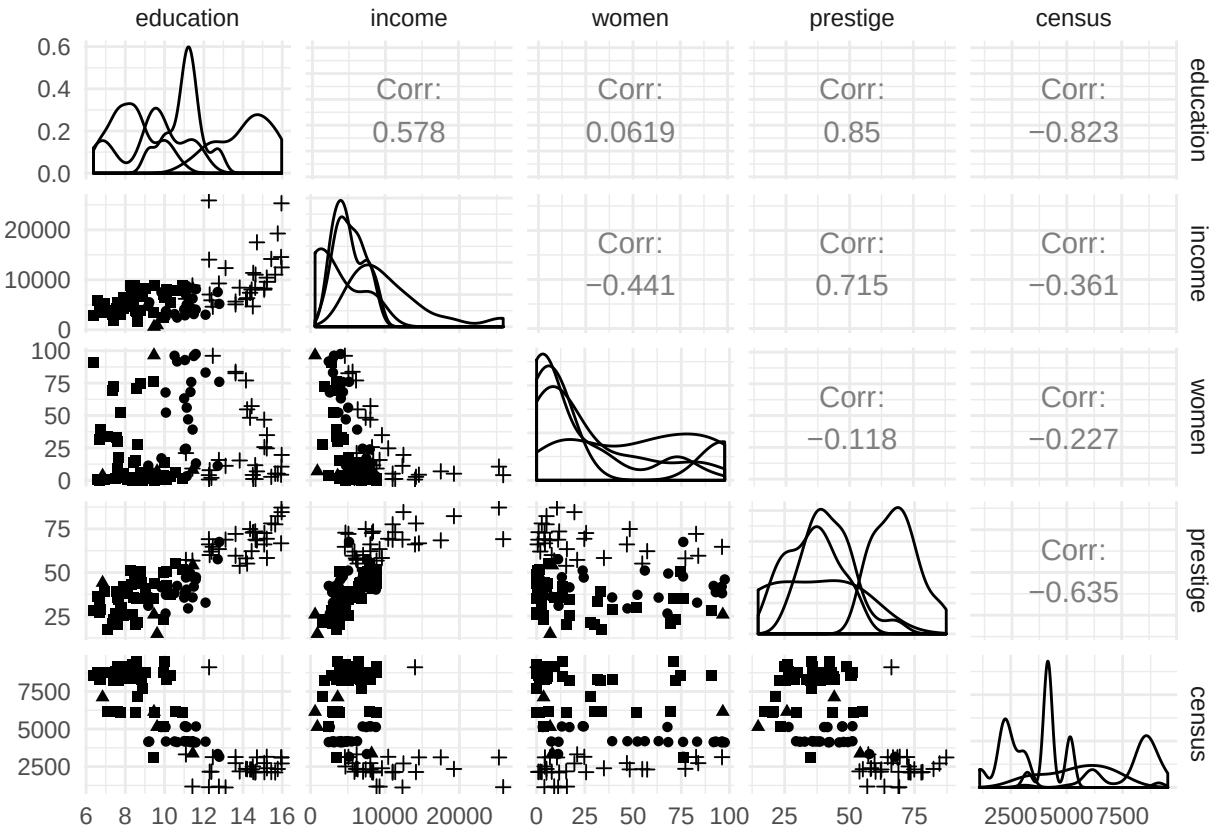
Answer:

```r
# Read in data
df2 <- read_delim("prestige.dat", delim = ',')

# Code the profession type using pch
pch_indicator <- rep(3, nrow(df2))
pch_indicator[df2$type == "bc"] <- 6
pch_indicator[df2$type == "prof"] <- 8
pch_indicator[df2$type == "wc"] <- 0

# Create the scatterplot matrix of all quantitative
# varaibles, making use of profession type symbols
ggpairs(df2[sapply(df2, class) != "character"],
```

```
      mapping = aes(pch = as.factor(pch_indicator))) +
theme_minimal()
```



Given the plot above, it makes sense to regress on only income, education and type. There is no correlation between prestige and percentage of incumbents who were women. The magnitude of the correlation between prestige and percentage of incumbents who were women is also not high enough to warrant it useful for regression. It can also be seen in that plot that there is a linear relationship between prestige and census but it is very scattered.

(c) Which professions are missing "type"? Since the other variables for these observations are available, we could group them together as a fourth professional category to include them in the analysis. Is this advisable or should we remove them from our data set? Justify your answer.

Answer:

```
# Print rows with missing profession type
df2[is.na(df2$type),]
```

```
## # A tibble: 4 x 7
##   occupation.group education income women prestige census type
##   <chr>                <dbl>  <int> <dbl>    <dbl>  <int> <chr>
## 1 athletes              11.4    8206  8.13     54.1   3373 <NA>
## 2 newsboys               9.62    918  7        14.8   5143 <NA>
## 3 babysitters            9.46    611 96.5      25.9   6147 <NA>
## 4 farmers                6.84   3643  3.6      44.1   7112 <NA>
```

The four professions that have missing "type" are: athletes, newsboys, babysitters and farmers. It would not make sense to group them together as a fourth professional category because there are only 4 observations as such and it can introduce some bias in the analysis. It is advisable to remove them from our data set.

(d) Visually, does there seem to be an interaction between type and education and/or type and income? Justify your answer.

Answer:

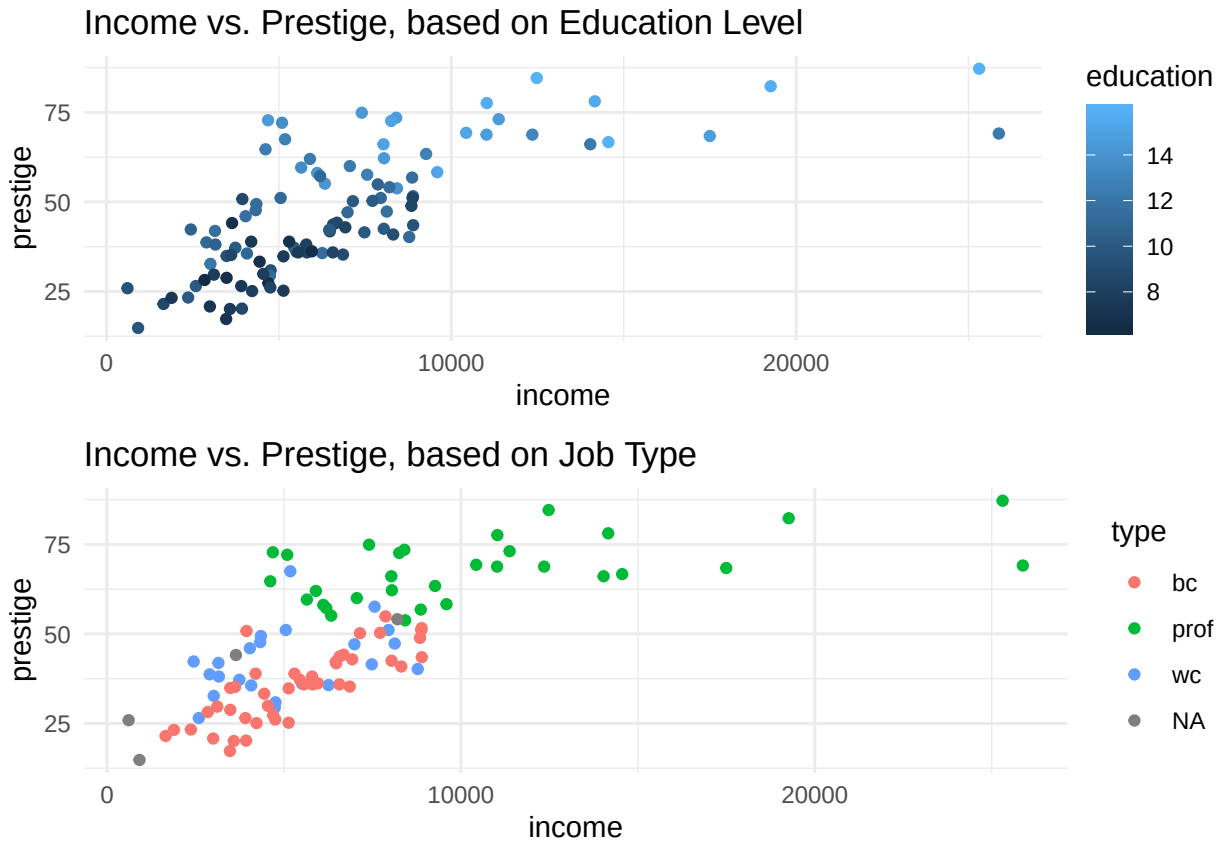
```
# Create plots of the interactions
g1 <- ggplot(df2, aes(x = income, y = prestige, color = education)) + geom_point() +
  ggtitle("Income vs. Prestige, based on Education Level") +
  theme_minimal()
g2 <- ggplot(df2, aes(x = income, y = prestige, color = type)) + geom_point() +
  ggtitle("Income vs. Prestige, based on Job Type") +
  theme_minimal()

# Import gridExtra
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# Print the graphs aesthetically
grid.arrange(g1, g2, nrow = 2)
```



There apears to be an interaction between type and education, as well as type and income. It can be seen that as education level goes up, people have higher income and higher prestige. This contrasts with the people who have low education level; their income is in the lower range, as well as their prestige. On the other plot, it can be seen that `prof` have higher prestige, yet have scattering incomes. Other than `prof`, `bc`

and `wc` appear to be in the same income range as well as prestige group. This interaction, although apparent somewhat, is weak since `bc` and `wc` does not appear to be distinguishable in their income and prestige. There is a visible interaction between type and education.

(e) Fit a model to predict prestige using: income, education, type and any interaction terms based on your answer to part (d). Evaluate the model and include relevant output. Use your answer to part (c) to determine which observations to use in your analysis.

Answer:

```
# Create model using income, education, type to predict prestige
# and print output
model_prestige <- lm(data = df2, prestige ~ income + education + type + type*education)
summary(model_prestige)
```

```
##
## Call:
## lm(formula = prestige ~ income + education + type + type * education,
##     data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1168  -4.1751   0.4384   5.1625  15.2362
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.331e+00  7.783e+00  -0.299    0.765
## income               1.052e-03  2.201e-04   4.782 6.66e-06 ***
## education            3.852e+00  9.406e-01   4.096 9.12e-05 ***
## typeprof             2.209e+01  1.520e+01   1.454    0.149
## typewc              -2.822e+01  1.959e+01  -1.440    0.153
## education:typeprof -1.227e+00  1.304e+00  -0.941    0.349
## education:typewc     2.270e+00  1.872e+00   1.213    0.228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.036 on 91 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8411, Adjusted R-squared:  0.8306
## F-statistic: 80.27 on 6 and 91 DF,  p-value: < 2.2e-16
```

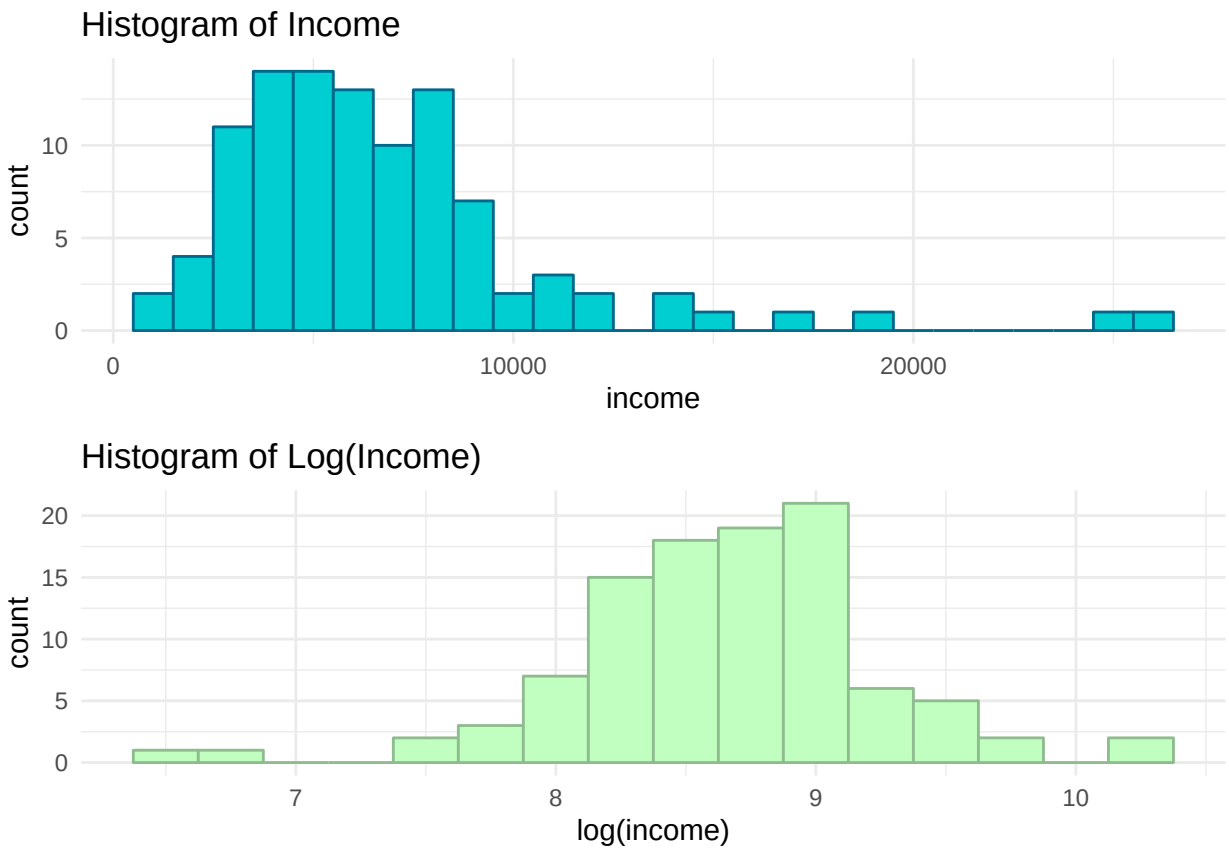
This model has an adjusted R^2 value of 0.8306, meaning that 83.06% of the variance in prestige was explained by the income, education, type and interaction term (education and type). Furthermore, the coefficient estimates for β_{income} and $\beta_{\text{education}}$ are statistically significant since their p-values are less than $\alpha = 0.05$. The observations used in this analysis were the ones that had an actual `type` of profession and not an empty value in its place. This means that the regression was not run for the following occupations: athletes, newsboys, babysitters and farmers.

(f) Create a histogram of income and a second histogram of log(income) (i.e., natural logarithm). How does the distribution change?

Answer:

```
# Create histograms using income and log(income)
g3 <- ggplot(df2, aes(x = income)) + geom_histogram(binwidth = 1000,
                                                    color = "deepskyblue4",
                                                    fill = "darkturquoise") +
  ggtitle("Histogram of Income") + theme_minimal()
```

```
g4 <- ggplot(df2, aes(x = log(income))) + geom_histogram(binwidth = 0.25,
                                                         color = "darkseagreen",
                                                         fill = "darkseagreen1") +
  ggtitle("Histogram of Log(Income)") + theme_minimal()
# Print histograms
grid.arrange(g3, g4, nrow = 2)
```



By applying the natural logarithm to income, the distribution of income goes from skewed right to skewed left.

(g) Fit the model in (e) but this time use log(income) (i.e., natural logarithm) instead of income. Evaluate this model and provide the relevant output.

Answer:

```
model_prestige_logincome = lm(data = df2, prestige ~ log(income) + education + type + type*education)
summary(model_prestige_logincome)
```

```
##
## Call:
## lm(formula = prestige ~ log(income) + education + type + type *
##     education, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.962  -3.797   1.041   4.092  16.503
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -81.6672    14.3681  -5.684 1.57e-07 ***
## log(income)         10.6833     1.7108   6.245 1.33e-08 ***
## education            3.1407     0.9004   3.488 0.000751 ***
## typeprof            15.6176    14.2168   1.099 0.274871
## typewc             -30.4466    18.3465  -1.660 0.100451
## education:typeprof  -0.5801     1.2211  -0.475 0.635887
## education:typewc     2.6675     1.7551   1.520 0.132018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.585 on 91 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8608, Adjusted R-squared:  0.8516
## F-statistic: 93.79 on 6 and 91 DF,  p-value: < 2.2e-16
```

This model has an adjusted R^2 value of 0.8516, meaning that 85.16% of the variance in prestige was explained by the log income, education, type and interaction term (education and type). In addition, the RSE of the model is 6.585, which is lower than the the RSE of the previous model to predict `prestige`. Furthermore, the coefficient estimates for $\beta_{\text{log(income)}}$ and $\beta_{\text{education}}$ are statistically significant since their p-values are less than $\alpha = 0.05$.

(h) Is the model in (e) or (g) better? Justify your answer. Why can't a partial F-test be used here?

Answer: The model in (g) is *slightly* better than the one in (e). This is because the adjusted R^2 value goes slightly up when log of income is used rather than income, and the RSE goes down. A partial F-test cannot be used here because the features in one model are not a subset of the features in the model. If the income variable was removed from either models (not both), then a partial F-test can be done.