**Question 1a**

A wine vintage is the year in which the grapes were harvested. A wine’s vintage can greatly affect the taste and quality, primarily because of the weather that affects the vines throughout the growing season.

**Question 1b**

The response variable for the three models described in this paper is Logarithm of price relative to 1961, called LNPRICE2.

**Question 1c**

The values of LNPRICE2 are missing(No.3,5 and 30 through 38) corresponding to:

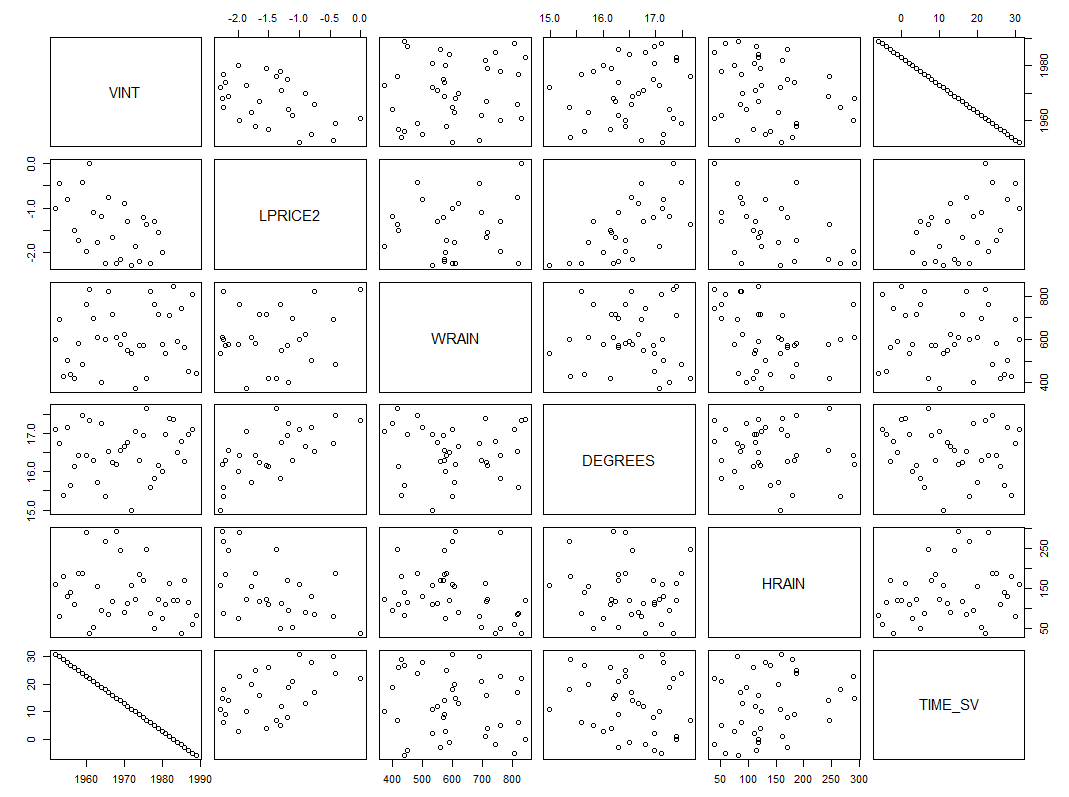
In 1954,1956,1981,1982,1983,1984,1985,1986,1987,1988,1989.

The reasons why they are missing:

Wines of 1954 and 1956 are now rarely sold.

Wines of 1981,1982,1983,1984,1985,1986,1987,1988 and1989 are not mature yet.

**Question 1d**

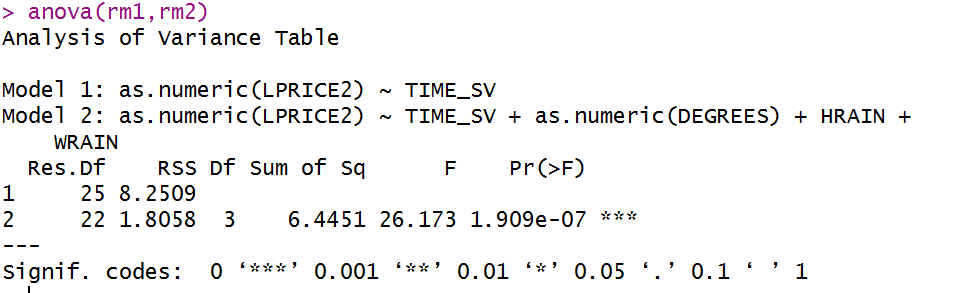


There is a positive linear relationship between VINT and LNPRICE2 and a negative linear relationship between TIME\_SV and LNPRICE2.The relation between VINT and TIME\_SV is complete negative relation.

We can also see some positive linear relationship between LNPRICE2 and WRAIN, lNPRICE2 and DEGREES and a slight negative linear relationship between LNPRICE2 and HRAIN.

**Question 1e**

The second model is better, which include variables:HRAIN,WRAIN,DEGREES and VINTAGE.Based on the output, model 2 has a significantly higher R-square than model 1.That means model 2 explains more variability in the LNPRICE2 than model1.

Hypothesis:

*H*0: β(WRAIN) = β(HRAIN) = β(DEGREES) = 0

*H*1: At least one of these slopes does not equal zero

Choose significance level : α = 0.05

Test statistic: F = 26.173

Which has an F distribution with 3 and 22 degrees of freedom

P-value = 1.909e-07 < α = 0.05

Conclusion: Reject null hypothesis; at least one of the slopes is significantly different from 0, therefore the second model is better than the first model.

I choose the same model as the author did.

**Question 1f**

The sample size of model is 27

**Question 1g**

LPRICE2 = -12.15 - 0.02385(1/years)\*TIME\_SV(years) + 0.6164(1/℃)\*DEGREES(℃) - 0.003861(1/ml)\*HRAIN(ml) + 0.00167(1/ml)\*WRAIN(ml)

Slope for TIME\_SV: if the age of wine increase 1 year, holding other exploratory variables DEGREES, WRAIN and HRAIN constant, the log of average vintage price relative to 1961 increases roughly 0.02385.

Slope for DEGREES: if the average temperature over the growing season increase 1℃, holding other exploratory variables TIME\_SV, WRAIN and HRAIN constant, the log of average vintage price relative to 1961 increases roughly 0.6164.

Slope for HRAIN: if the rainfall in September and August increase 1 ml, holding other exploratory variables TIME\_SV, DEGREES and WRAIN constant, the log of average vintage price relative to 1961 decreases roughly 0.003861.

Slope for WRAIN: if the rainfall in the months preceding the vintage increase 1 ml, holding other exploratory variables TIME\_SV, DEGREES and HRAIN constant, the log of average vintage price relative to 1961 increases roughly 0.00167.

Y-intercept: if the age of wine is 0 year, the average temperature over the growing season is 0℃, the rainfall both in the months preceding the vintage and in September and August is 0 ml, the log of average vintage price relative to 1961 would be -12.15.

The y-intercept does not have a practical interpretation. Since the minimum values of DEGREES, WRAIN and HRAIN are 14.98℃, 376.0 ml and 38.0 m, the y-intercept does not have a practical interpretation.

**Question 1h**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SSE | RMSE | PRESS | RMSE(jackknife) |
| Model1 | 8.2509 | 0.5745 | 9.395569 | 0.61304 |
| Model2 | 1.8058 | 0.2865 | 2.816975 | 0.35783 |

Based on the result above, model 2 has smaller SSE, RMSE, PRESS and RMSE(jackknife) than model 1.

I will not change my answer to part e.

**Question 1i**

We cannot use these models to predict the quality for wine produced in 2005.

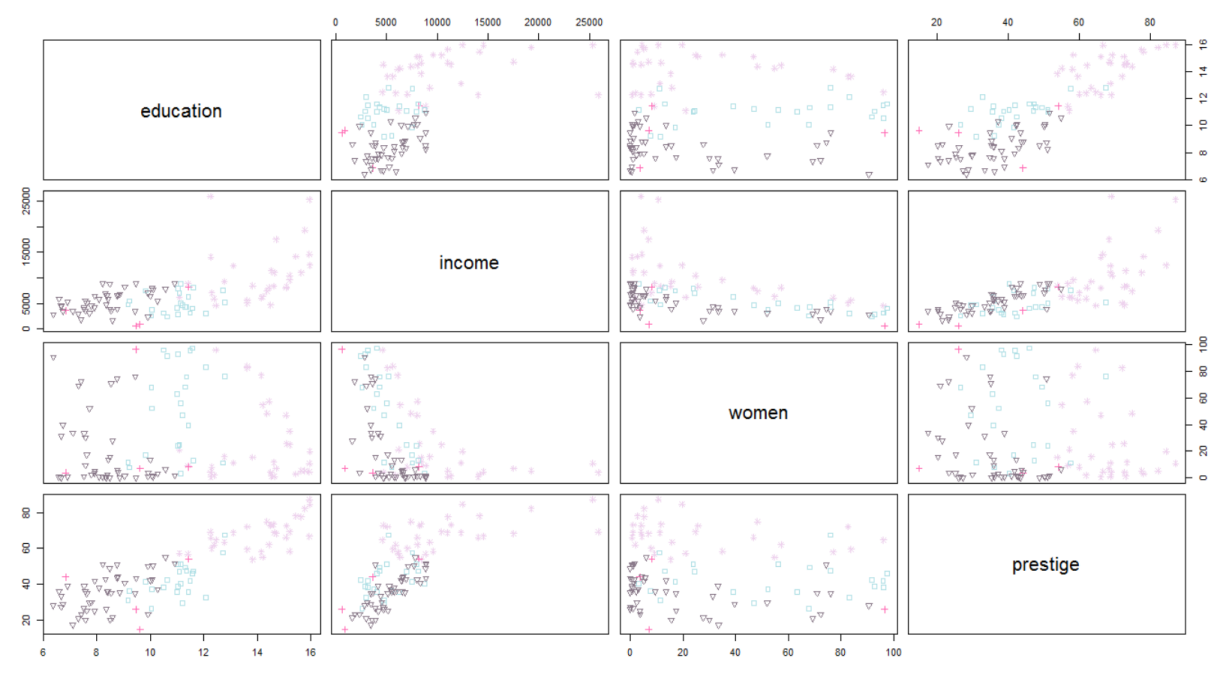
Because in the data, the VINT 2005 is not included.The prediction of the quality for wine produced in 2005 would be an extrapolation. Besides, we do not have any data for DEGREES, HRAIN and WRAIN in 2005.

**Question 2a**

The Pineo-Porter scale is a way for sociologists to measure occupational prestige.Occupational prestige refers to the public perception of an individual’s social standing based on their professional position, rather than any unique personal attributes the individual holds.It is a method to measure occupational prestige. Occupational prestige represents the respect that a specific occupation holds in a society. The ordinal classification schemes is improved by collapsing occupational titles into a limited number of categories. In 1977, Prineo, Porter and McRoberts introduced a widely used ordinal scale of 17 occupational categories with general educational development and skill properties of occupations. (Sources: Wikipedia:occupational prestige;A Socioeconomic Scale for Canada: Measuring Occupational Status from the Census)

I think it is a reliable measure.It significantly reduced the number of categories and reduced the burden of ranking among a large selection of occupations.It includes aspects such as income and education to support it.

**Question 2b**



Restricting our regression to only these variables make sense.

Because from the scatterplot above, we can see obvious linear relationship patterns between prestige and education and also prestige and income.Both education and income should be included in the model. The relationship between prestige and women is hard to describe, so the variable woman should not be included in the model.Type should be included in the model.Based on different types(marked with different colors in different shapes), we can make several segments for prestige.

**Question 2c**

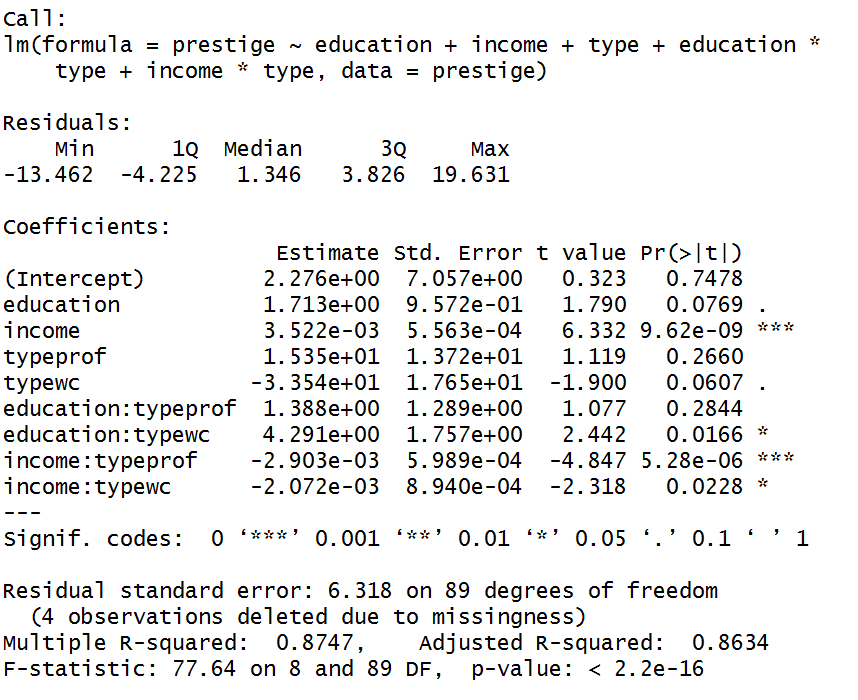
Athletes,newsboys, babysitters and farmers.

We should remove them from our data set.Since these four occupations are not the same type, it is not appropriate to group them together as a fourth professional category.

**Question 2d**

There seem to be an interaction between type and education and type and income.Based on the plot above(in question 2b), we can see different types(marked with different colors in different shapes), correspond to different education or income level.People who are in professional occupations have higher education and income than white collars do, and white collars have higher education and income than blue collar.

**Question 2e**



The adjusted R-squared(0.8634) means that 86.34% of variation of Pineo-Porter prestige score can be explained by the regression model.

Hypothesis test when α=0.05

Assume regression assumptions are satisfied;

Overall F-test significant;

Slope of education is not statistically significant;

Slope of income is statistically significant;

Slope of type when pro is not statistically significant;

Slope of type when wc is not statistically significant;

Slope of interaction term (education:typeprof) is not statistically significant;

Slope of interaction term (education:typewc) is statistically significant;

Slope of interaction term (income:typeprof) is statistically significant;

Slope of interaction term (income:typewc) is statistically significant;

Final model:

Type = bc

Prestige(points) = 2.28(points) + 1.7(points/years) \* education(years) + 0.00352(points/$) \* income($)

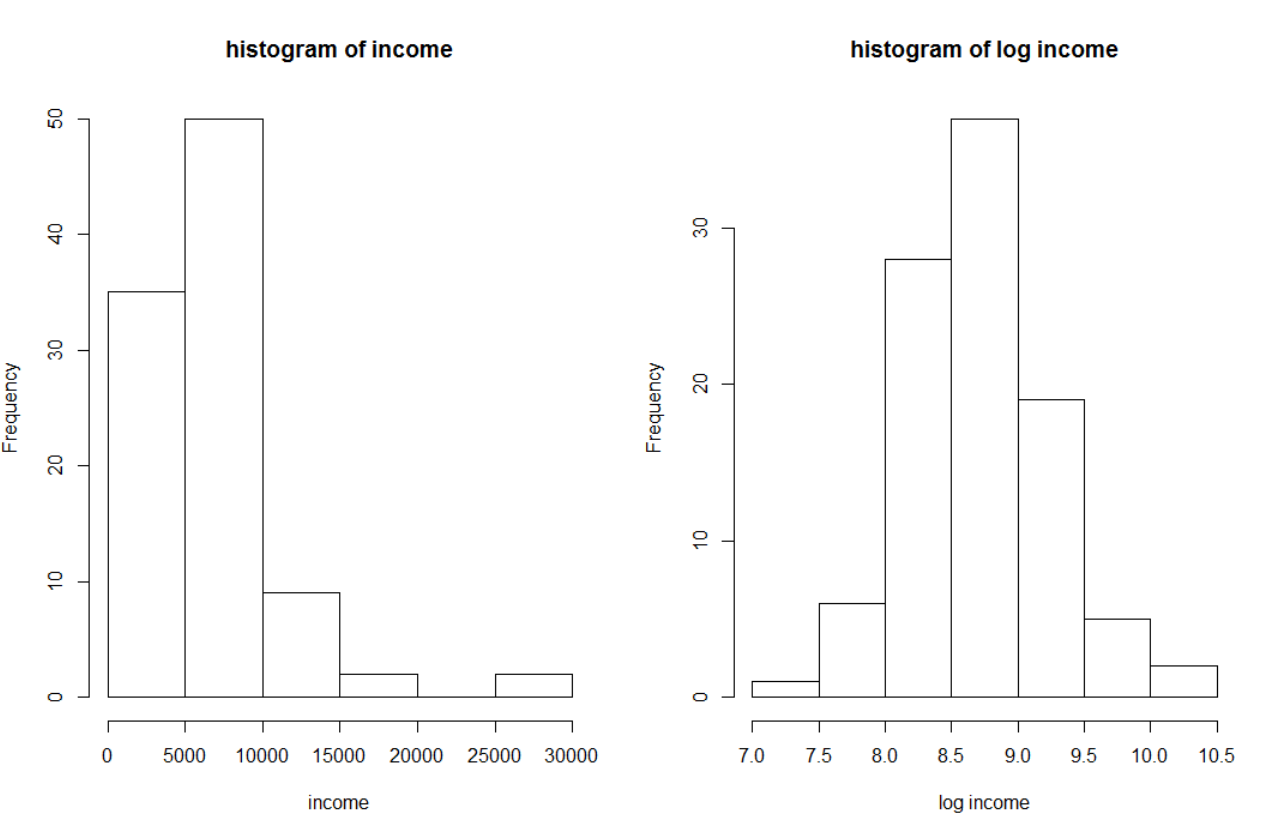
Type = prof

Prestige (points) = 3.8(points) + 3.1(points/years) \* education(years) + 0.00062(points/$) \* income($)

Type = wc

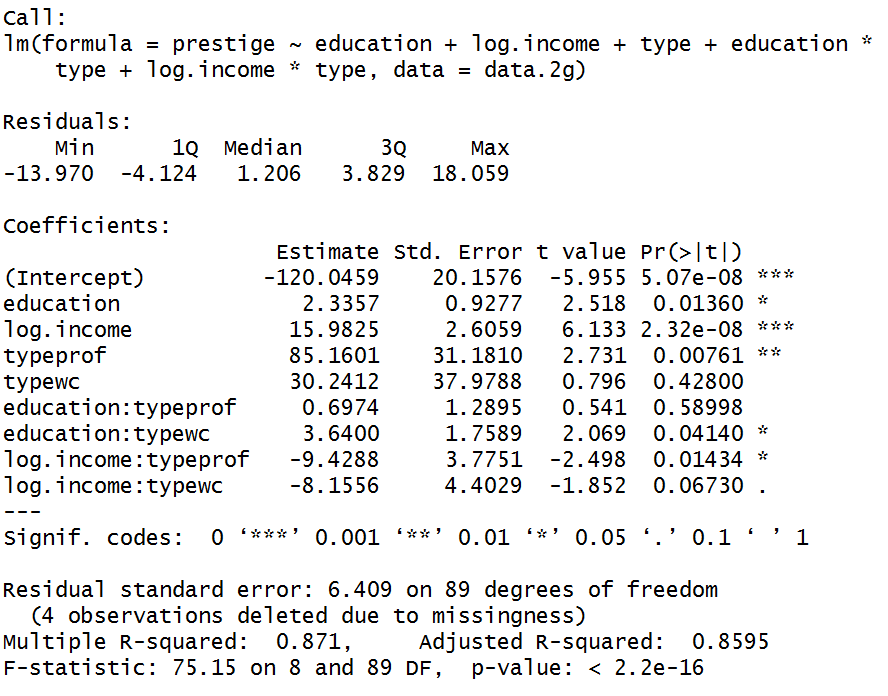
Prestige(points) = 1.94(points) + 6.0(points/years) \* education(years) + 0.00145(points/$) \* income($)

**Question 2f**



The distribution turned from right-skewed to symmetric.

**Question 2g**



The adjusted R-squared(0.8595) means 85.95% of variation of Pineo-Porter prestige score can be explained by the regression model.

Hypothesis test when α=0.05

Assume regression assumptions are satisfied;

Overall F-test significant;

Slope of education is statistically significant;

Slope of log.income is statistically significant;

Slope of type when pro is statistically significant;

Slope of type when wc is not statistically significant;

Slope of interaction term (education:typeprof) is not statistically significant;

Slope of interaction term (education:typewc) is statistically significant;

Slope of interaction term (log.income:typeprof) is statistically significant;

Slope of interaction term (log.income:typewc) is not statistically significant;

Final model:

Type = bc:

Prestige(points) = -120.0459(points) + 2.3357(points/years) \* education(years) + 15.9825(points/$) \* log.income($)

Type = prof:

Prestige (points) = -34.8858(points) + 3.0331(points/years) \* education(years) + 6.5537(points/$) \* log.income($)

Type = wc

Prestige(points) = -89.8047(points) + 5.98(points/years) \* education(years) + 7.8269(points/$) \* log.income($)

**Question 2h**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SSE | RMSE | PRESS | RMSE(jackknife) |
| Model(e) | 3552.9 | 6.318 | 4285.977 | 6.9395 |
| Model(g) | 3655.4 | 6.409 | 4399.257 | 7.0309 |

The model in (e) has higher adjusted-R-square(0.8634) than that(0.8595) in model in (g).

Also, we can compare the SEE, RMSE, PRESS and RMSE(jackknife).The model in (e) has smaller SSE, RMSE, PRESS and RMSE(jackknife) than model in (g). Thus, model in (e) is better than model in (g).

We cannot use partial F-test here because we cannot find one model in these two models to be the nested model of the other model.