Machine Learning Homework Assignment

Elliott Oates

November 16, 2023

Data Exploration

nr	year	black	exper	hisp	hours	married	occ1	occ2	occ3
Min. : 13	Min. :1980	Min. :0.0000	Min. : 0.000	Min. :0.000	Min. : 120	Min. :0.000	Min. :0.0000	Min. :0.00000	Min. :0.00000
1st Qu.: 2329	1st Qu.:1982	1st Qu.:0.0000	1st Qu.: 4.000	1st Qu.:0.000	1st Qu.:2040	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000
Median : 4569	Median :1984	Median :0.0000	Median : 6.000	Median :0.000	Median :2080	Median :0.000	Median :0.0000	Median :0.00000	Median :0.00000
Mean : 5262	Mean :1984	Mean :0.1156	Mean : 6.515	Mean :0.156	Mean :2191	Mean :0.439	Mean :0.1039	Mean :0.09151	Mean :0.05344
3rd Qu.: 8406	3rd Qu.:1985	3rd Qu.:0.0000	3rd Qu.: 9.000	3rd Qu.:0.000	3rd Qu.:2414	3rd Qu.:1.000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:0.00000
Max. :12548	Max. :1987	Max. :1.0000	Max. :18.000	Max. :1.000	Max. :4992	Max. :1.000	Max. :1.0000	Max. :1.00000	Max. :1.00000

occ4	occ5	occ6	occ7	occ8	occ9	educ	union	wage
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. :0.00000	Min. :0.0000	Min. : 3.00	Min. :0.000	Min. : 0.0279
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:11.00	1st Qu.:0.000	1st Qu.: 3.8602
Median :0.0000	Median :0.0000	Median :0.0000	Median :0.00000	Median :0.00000	Median :0.0000	Median :12.00	Median :0.000	Median : 5.3182
Mean :0.1115	Mean :0.2142	Mean :0.2021	Mean :0.09197	Mean :0.01468	Mean :0.1167	Mean :11.77	Mean :0.244	Mean : 5.9192
3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:12.00	3rd Qu.:0.000	3rd Qu.: 7.3235
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.00000	Max. :1.00000	Max. :1.0000	Max. :16.00	Max. :1.000	Max. :57.5043

Question 1 [5 Marks]

Compute the correlation matrix for all variables in the data

The Correlation Matrix for all variables can be computed as following:

	nr	year	black	exper	hisp	hours	married	occ1	occ2	occ3	occ4	occ5	occ6	occ7	occ8	occ9	educ	union	wage
nr	1.00	0.00	0.12	0.07	0.31	0.01	0.03	0.01	-0.03	0.00	-0.07	-0.02	0.01	0.02	0.02	0.05	-0.05	0.01	-0.03
year	0.00	1.00	0.00	0.81	0.00	0.22	0.28	0.05	0.10	0.03	-0.03	0.06	-0.06	-0.06	-0.05	-0.05	0.00	-0.01	0.26
black	0.12	0.00	1.00	0.04	-0.16	-0.02	-0.13	-0.05	-0.04	-0.04	0.02	-0.07	0.07	0.02	-0.03	0.09	-0.04	0.11	-0.07
exper	0.07	0.81	0.04	1.00	0.07	0.21	0.29	-0.07	0.02	-0.01	-0.06	0.12	0.01	-0.01	-0.00	-0.05	-0.34	0.01	0.15
hisp	0.31	0.00	-0.16	0.07	1.00	0.03	0.01	-0.05	-0.00	-0.05	-0.03	-0.02	0.04	0.02	0.04	0.05	-0.20	0.03	-0.03
hours	0.01	0.22	-0.02	0.21	0.03	1.00	0.20	-0.02	0.13	0.02	-0.09	0.03	0.04	-0.04	0.10	-0.12	0.02	-0.05	-0.03
married	0.03	0.28	-0.13	0.29	0.01	0.20	1.00	0.01	0.03	-0.03	-0.05	0.14	-0.01	-0.04	0.03	-0.10	0.02	0.04	0.17
occ1	0.01	0.05	-0.05	-0.07	-0.05	-0.02	0.01	1.00	-0.11	-0.08	-0.12	-0.18	-0.17	-0.11	-0.04	-0.12	0.27	-0.09	0.13
occ2	-0.03	0.10	-0.04	0.02	-0.00	0.13	0.03	-0.11	1.00	-0.08	-0.11	-0.17	-0.16	-0.10	-0.04	-0.12	0.15	-0.12	0.10
occ3	0.00	0.03	-0.04	-0.01	-0.05	0.02	-0.03	-0.08	-0.08	1.00	-0.08	-0.12	-0.12	-0.08	-0.03	-0.09	0.10	-0.10	0.10
occ4	-0.07	-0.03	0.02	-0.06	-0.03	-0.09	-0.05	-0.12	-0.11	-0.08	1.00	-0.18	-0.18	-0.11	-0.04	-0.13	0.05	0.01	-0.04
occ5	-0.02	0.06	-0.07	0.12	-0.02	0.03	0.14	-0.18	-0.17	-0.12	-0.18	1.00	-0.26	-0.17	-0.06	-0.19	-0.15	-0.01	0.03
occ6	0.01	-0.06	0.07	0.01	0.04	0.04	-0.01	-0.17	-0.16	-0.12	-0.18	-0.26	1.00	-0.16	-0.06	-0.18	-0.14	0.11	-0.04
occ7	0.02	-0.06	0.02	-0.01	0.02	-0.04	-0.04	-0.11	-0.10	-0.08	-0.11	-0.17	-0.16	1.00	-0.04	-0.12	-0.10	0.09	-0.07
occ8	0.02	-0.05	-0.03	-0.00	0.04	0.10	0.03	-0.04	-0.04	-0.03	-0.04	-0.06	-0.06	-0.04	1.00	-0.04	-0.05	-0.04	-0.08
occ9	0.05	-0.05	0.09	-0.05	0.05	-0.12	-0.10	-0.12	-0.12	-0.09	-0.13	-0.19	-0.18	-0.12	-0.04	1.00	-0.04	0.06	-0.13
educ	-0.05	0.00	-0.04	-0.34	-0.20	0.02	0.02	0.27	0.15	0.10	0.05	-0.15	-0.14	-0.10	-0.05	-0.04	1.00	-0.01	0.26
union	0.01	-0.01	0.11	0.01	0.03	-0.05	0.04	-0.09	-0.12	-0.10	0.01	-0.01	0.11	0.09	-0.04	0.06	-0.01	1.00	0.12
wage	-0.03	0.26	-0.07	0.15	-0.03	-0.03	0.17	0.13	0.10	0.10	-0.04	0.03	-0.04	-0.07	-0.08	-0.13	0.26	0.12	1.00

Which 4 variables are most correlated with wage?

One can visually inspect the above matrix to determine the 4 most correlated variables with wage, however it may be more efficient to use code. The four most correlated are shown in the table below (in order)

Variable	Description
educ	Number of School Years
year	Year
married	Marital Status (1 if married, 0 otherwise)
exper	Work Experience (in years)

Produce a pairwise scatterplot with wage and these 4 variables and Briefly interpret the scatterplot

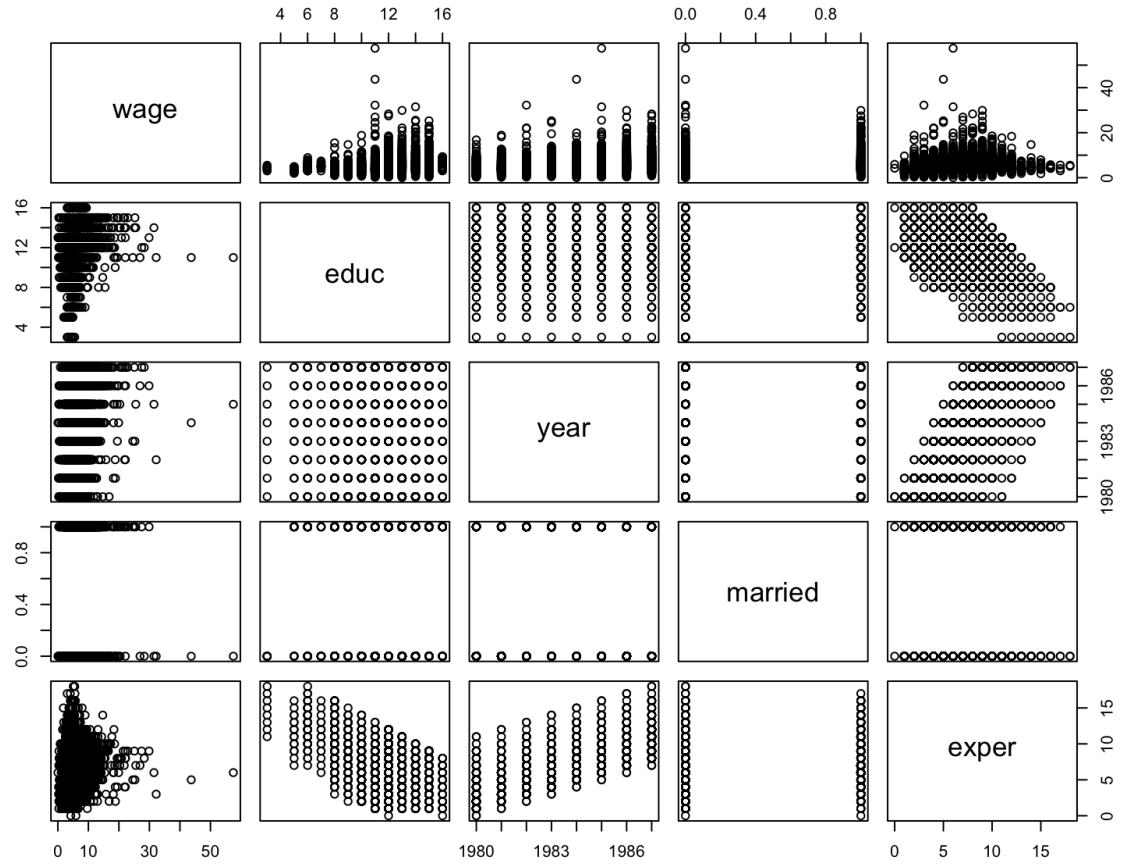


Figure 1: Pairwise Scatterplots with Wage and, Education, Year, Married and Experience

- From the top row, second from the left plot, we can see the scatterplot of Wage (Y-axis) and Educ (X-axis). Visually, we can't tell whether there the mean wage for given levels of interpretation increases, however, we can tell that the spread of wages increases as years of education increases. This has some economic intuition, the income of the highest earners at each level of education generally increases. There is a noticeable occurrence of 2 very high earners at the level of education of 12 years.
- Education and Experience, the second from the left, bottom row plot, suggests an expected linear relationship between the amount of time spent in each level of education as having a negative linear relationship with the level of work experience. This makes sense as spending a longer time in education prevents you from entering the workforce.
- From the top row, middle plot, we can visually see that the spread of income has increased as the year of the observation is increased.
- We can also see that the range of levels of experience in the dataset begins to increase very linearly with the year. This is because the same individuals are used each year and so their level of experience will just increase by 1 each observation.
- From the top row, second from the right plot, visually we can see that the marriage status does not have an incredible impact on wage; however, there are only observations of extreme income in the single population.

- From the top right plot, with wage on experience, we visually see something that begins to resemble a normal distribution. Wages gradually increase with experience, then after a certain point start to decrease again.

Question 2 [5 Marks]

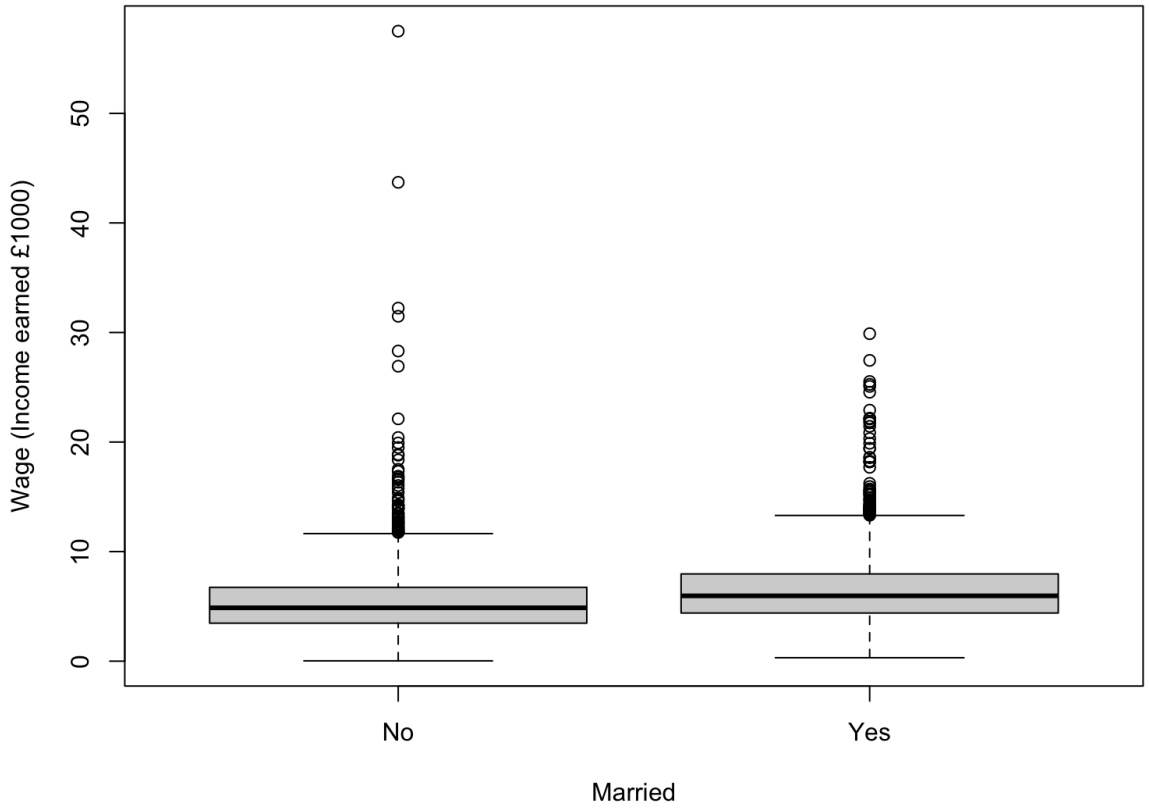


Figure 2: Boxplot with Married on X-axis and Wage on Y-axis

From figure 2 we can make the following observations:

- The mean wage of those that are married is higher than those that are single.
- The Interquartile range of those that are married is higher than those that are single.
- The minimum and maximum values (excluding outliers) are higher for those that are married than those that are single.
- Although the distribution of the two population groups, married and unmarried, is pretty similar, and they have a visually similar interquartile range, and in both groups, the mean is just slightly closer to the lower quartile value, they are different in that they are centered around different means. The mean married wage is higher.
- Both population subgroups have several observations where the wage is considered an outlier, or 1.5 times the interquartile range. However, there are only two observations of a wage above 35,000, and these both occur in the single population.

Question 3 [8 Marks]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-469.9863	74.4579	-6.31	0.0000
nr	-0.0000	0.0000	-1.39	0.1647
year	0.2369	0.0377	6.28	0.0000
black1	-0.4625	0.1434	-3.22	0.0013
exper	0.0979	0.0328	2.99	0.0028
hisp1	0.2127	0.1313	1.62	0.1053
hours	-0.0007	0.0001	-9.03	0.0000
married1	0.5949	0.0946	6.29	0.0000
occ11	1.7156	0.1919	8.94	0.0000
occ21	1.7663	0.1983	8.91	0.0000
occ31	2.2030	0.2306	9.55	0.0000
occ41	0.6290	0.1821	3.45	0.0006
occ51	1.2776	0.1612	7.93	0.0000
occ61	1.0101	0.1606	6.29	0.0000
occ71	0.5178	0.1914	2.71	0.0068
occ81	-0.1986	0.3839	-0.52	0.6049
educ	0.4686	0.0326	14.39	0.0000
union1	1.0824	0.1044	10.37	0.0000

Residual standard error: 2.85 on 4342 degrees of freedom
Multiple R-squared: 0.2109, Adjusted R-squared: 0.2078
F-statistic: 68.28 on 17 and 4342 DF, p-value: $< 2.2e-16$

3(a) [2 marks] Is there a relationship between wage and all the predictors?

Based on the model summary we can conclude that there is a statistically significant relationship between wage and the predictors. The models F-statistic is 62.28 and has an incredibly small p-value. This F-statistic represents a hypothesis test where the null hypothesis is that all coefficients equal to 0, against the alternative hypothesis that at least one of the coefficients does not equal 0.

If the null hypothesis is true, the F-statistic will be close to 1, and if the alternative hypothesis that there is overall significance, the F-statistic will be greater than 1. An F-statistic of 62.28 on 17 parameters and 4360 observations is sufficiently large to reject the null hypothesis and thus we can conclude there is overall statistical significance in the model.

3(b) [3 marks] How much is the residual standard error? Interpret it.

The residual standard error for the model is 2.85 on 4342 degrees of freedom. This is the average difference between the actual values for wage in the data set and the values for wage predicted by the model. In the context of the variable wages this means that the average difference between the actual and predicted income earnt in £1,000s was 2.85, so £2,850.

3(c) [3 marks] Interpret the coefficients on the married and hisp indicator variables.

The coefficient on married is 0.595 and is significant at all conventional significant levels. The coefficient is for a dummy (categorical variable). This means that given all other variables are fixed (the same), the difference in the predicted wage for someone who is single and someone who is married is 0.595. So with all other variables constant, someone who is married is will have a predicted income of £595 higher than someone who is single.

The coefficient on Hispanic is 0.213 and is not significant at any significant level. This means that given all other variables are fixed (the same), the difference in the predicted wage for someone who is Hispanic and someone who is not Hispanic is 0.213. So with all other variables constant, someone who is married

is will have a predicted income of £213 higher than someone who is single, however there is no statistical significance in this relationship in this model.

Question 4 [25 marks] Re-estimate the linear regression model including only significantly associated Variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-479.6720	74.0270	-6.48	0.0000
year	0.2418	0.0375	6.45	0.0000
black1	-0.5266	0.1386	-3.80	0.0001
exper	0.0940	0.0326	2.88	0.0040
hours	-0.0007	0.0001	-9.12	0.0000
married1	0.5859	0.0944	6.20	0.0000
occ11	1.7271	0.1869	9.24	0.0000
occ21	1.7949	0.1927	9.32	0.0000
occ31	2.2092	0.2258	9.78	0.0000
occ41	0.6551	0.1766	3.71	0.0002
occ51	1.2910	0.1538	8.39	0.0000
occ61	1.0321	0.1539	6.70	0.0000
occ71	0.5338	0.1861	2.87	0.0041
educ	0.4598	0.0321	14.33	0.0000
union1	1.0943	0.1041	10.51	0.0000

Residual standard error: 2.85 on 4345 degrees of freedom
Multiple R-squared: 0.2103, Adjusted R-squared: 0.2077
F-statistic: 82.63 on 14 and 4345 DF, p-value: $< 2.2e-16$

4(a) [8 marks] Produce diagnostic plots of the linear regression.

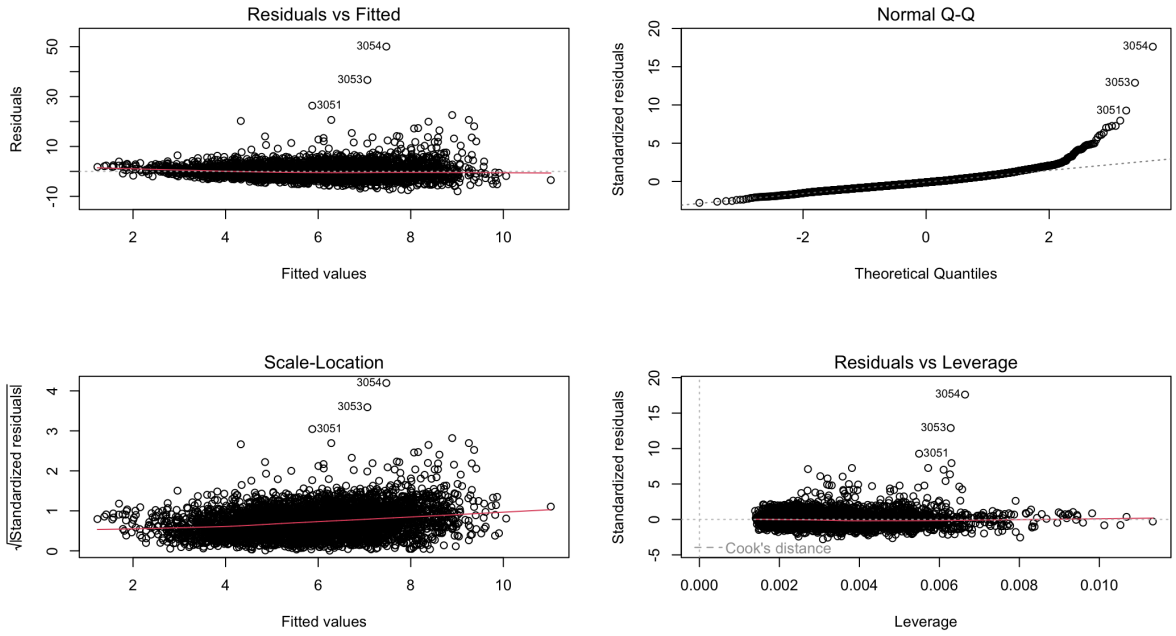


Figure 3: Linear Model with less predictors: Diagnostic Plots

The following interpretations can be made from figure 3.

- The Residuals vs Fitted values plot: This plot shows if residuals have non-linear patterns. It appears that in general there is no discernible non linear trends, the residuals for different fitted values levels roughly surround 0. However there is slight indication of non-constant variance the distribution of the residuals is concentrated around 0 for small fitted values, but spread out as the fitted values increase. This is an instance of "increasing variance". If the assumption of constant variance does not hold the confidence in our confidence intervals, prediction intervals, and the p values are compromised
- The Scale-Location Plot: This plot is a more sensitive approach to exploring the constant variance assumption (homoskedasticity). If you see significant trends in the red line on this plot, it tells you that the residuals (and hence errors) have non-constant variance. There is a clear upward trend and hence the residuals and errors have non-constant variance.
- The Normal Q-Q plot : This plot shows if residuals are normally distributed. The residuals are lined well on the straight dashed line which is good up until the theoretical quantile of 2. The upper tail is observed to be heavier and have larger values than what we would expect with the standard normal assumptions.
- The Residuals vs Leverage plot: This plot helps us to find influential cases (i.e., subjects) if there are any. Influential observations are ones where their omission from the data would significantly alter the regression line. In this plot there is no evidence of influential observations, the cooks distance curves are not even visible on the plot indicating that while there may be observations with high leverage or high residual, there are non with a problematic amount of both.

4(b) [6 marks] How many observations can be classified as outliers and high leverage points?

- 48 Observations can be considered as outliers with studentized residuals (residual/estimated std error) greater than 3.
- 3 Observations can be considered as high leverage points (where the leverage statistic is more than 3 times the average leverage statistic)

There is quite a large difference in the number of observations considered outliers and those considered as having high leverage. In multiple linear regression high leverage points can occur if each individual predictor's values are well within the expected range, but that is unusual in terms of the combination of the full set of predictors. Since there only 3 of these compared to 48 points where the actual wage is far from the predicted wage it suggests that there is no missing predictor or deficiency in the model. If there where a lot more high leverage points perhaps it could be that there is unexplored interaction effects but this is not the case.

4(c) [5 marks] Compute the variance inflation factor for all the predictor variables.

Variable	year	black	exper	hours	married	occ1	occ2
VIF	3.962029	1.054254	4.564458	1.106427	1.178985	1.745294	1.655993

Variable	occ3	occ4	occ5	occ6	occ7	educ	union
VIF	1.384048	1.657079	2.138047	2.050469	1.552468	1.684996	1.073109

Table 1: VIF values for the variables

Variance inflation factors are ratios of the variance of the estimated coefficient for a predictor when fitting it in full model to when it is fitted in a model on its own. Variance inflation factors are used to asses collinearity, which if present reduces the probability of correctly detecting a non-zero coefficient.

A variance inflation factor larger than 5 or 10 represents a problematic amount of collinearity between variables. In this model the largest is is 4.56 so we do not consider this model as having a problematic amount of collinearity.

4(d) [6 marks] The coefficient on the hours variable is negative. Is it expected?

The coefficient on the hours variable on wage is -0.00073100. This means that holding all other variables fixed a 1 unit increase in hours increase will result a -0.0073100 reduction in predicted wage. In the context of the data, a increase of 1 hour more worked in a year will result in a £7.3 reduction in predicted income earned.

This is not what one would initially expect, as it is generally the case that working more hours leads to more income, however that is the case for when you are on hourly wages. A possible explanation is that members of the labour union (and data-set) are on yearly salaries rather than hourly contracts and perhaps people earning higher salaries in managerial, executive positions will work less hours in a year than those in lower salary lower experience level jobs.

Question 5 [37 marks] Construct a new indicator variable, rich

The mean wage is computed as 5.919175, any observation with wage above this is given 1 for rich, any below is given 0.

5(a) [10 marks] Perform logistic regression on the training data to predict rich using educ, hours, exper, black, and married as the independent variables.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.4572	0.4642	-11.76	0.0000
educ	0.4723	0.0297	15.89	0.0000
hours	-0.0005	0.0001	-6.67	0.0000
exper	0.1434	0.0215	6.69	0.0000
black1	-0.6183	0.1305	-4.74	0.0000

Confusion Matrix with 0.4 Cutoff

	Actual	
Predicted	**0**	**1**
0	704	169
1	458	304

Test MSE = 0.3835

Confusion Matrix with 0.6 Cutoff

	Actual	
Predicted	**0**	**1**
0	1035	381
1	127	92

Test MSE = 0.3107

Confusion Matrix with 0.8 Cutoff

	Actual	
Predicted	**0**	**1**
0	1150	473
1	12	0

Test MSE = 0.2966

Figure 4: Confusion Matrix's for the Logistic Regression Model and Overall test error rates

The cut-off with the lowest overall test error rates is the 0.8 cut-off. Its test error rate is 0.2966361. If overall test MSE is the determinant of the most preferable cut-off then this is the best. However it is often important to consider the context and application of the model when choosing the best cut-off. While the overall test MSE is the lowest with a cut-off of 0.8, the 0.8 mis-classifies 100 percent of the observed rich observations and predicts them as not rich.

This may be problematic, if the intention of this model is to aid the labour union in better allocate their funds and resources among the union members that need it the most, then perhaps the model that mis-classifies every rich member as poor is not the best. In this case we could look at the cut-off that produces the lowest false negative error rate so as to make sure that there is the least amount of people that don't need union resources mis-classified as not rich. The cut-off with the lowest false negative error rate is 0.4.

5(b) [6 marks] Calculate the test classification error rates for the LDA method

Confusion Matrix with 0.4 Cutoff

	Actual	
Predicted	**0**	**1**
0	714	176
1	448	297

Test MSE = 0.3817

Confusion Matrix with 0.6 Cutoff

	Actual	
Predicted	**0**	**1**
0	1042	386
1	120	87

Test MSE = 0.3095

Confusion Matrix with 0.8 Cutoff

	Actual	
Predicted	**0**	**1**
0	1152	473
1	10	0

Test MSE = 0.2954

Figure 5: Confusion Matrix's for the LDA Method and Overall test error rates

5(c) [6 marks] Calculate the test classification error rates for the QDA method

Confusion Matrix with 0.4 Cutoff

	Actual	
Predicted	**0**	**1**
0	874	255
1	288	218

Test MSE = 0.3321

Confusion Matrix with 0.6 Cutoff

	Actual	
Predicted	**0**	**1**
0	1087	416
1	75	57

Test MSE = 0.3003

Confusion Matrix with 0.8 Cutoff

	Actual	
Predicted	**0**	**1**
0	1162	473
1	0	0

Test MSE = 0.2893

Figure 6: Confusion Matrix's for the QDA Method and Overall test error rates

5(d) [7 marks] Calculate the test classification error rate for the KNN classifier using the following values for K: 1, 10, 20, 50, 100 and 200.

Confusion Matrix with K = 1

	Actual	
Predicted	**0**	**1**
0	750	159
1	412	314

Test MSE = 0.3492

Confusion Matrix with K = 10

	Actual	
Predicted	**0**	**1**
0	701	158
1	461	315

Test MSE = 0.3786

Confusion Matrix with K = 20

	Actual	
Predicted	**0**	**1**
0	686	182
1	476	291

Test MSE = 0.4024

Confusion Matrix with K = 50

	Actual	
Predicted	**0**	**1**
0	605	172
1	557	301

Test MSE = 0.4459

Confusion Matrix with K = 100

	Actual	
Predicted	**0**	**1**
0	582	160
1	580	313

Test MSE = 0.4526

Confusion Matrix with K = 200

	Actual	
Predicted	**0**	**1**
0	574	160
1	588	313

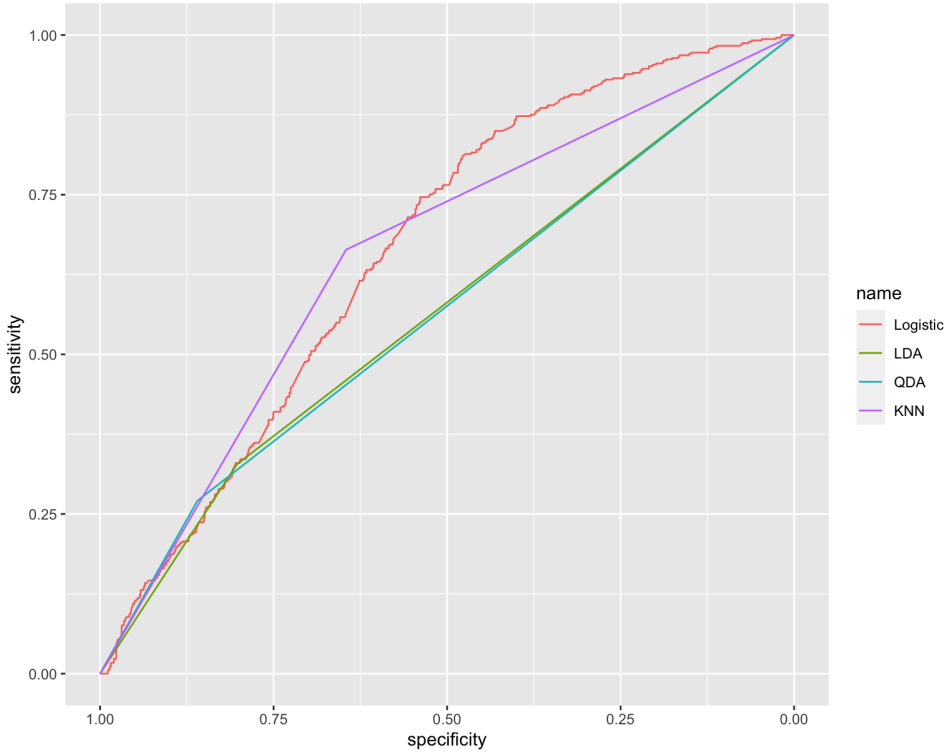
Test MSE = 0.4575

Figure 8: Confusion Matrices and Overall test error rates for KNN method

K = 1, is the preferred value for K. No improvements are made to the overall test error rate with additional neighbours used to determine the class.

5 (e) [8 marks] Which method appears to provide the best results on this data?

To determine the best method of classification for our data we can produce ROC curves and compute the area under each curve



ROC curves for different classification methods

Classification Method	AUC
Logistic Regression	0.6646
LDA	0.5653
QDA	0.5652
KNN (k=1)	0.6546

AUC values for different Classification Methods

ROC (false positive vs true positive) curves are a graphical tool that display the two types of errors for all posible thresholds. The average performance for a classification method under all possible thresholds is given as the area under the curve. A larger AUC indicates a better classification method.

The method with the Highest AUC is Logistic Regression followed closely by KNN with k = 1. LDA and QDA have similar and lower AUCs. LDA and QDA may be worse due to assumptions made on the parametric forms of the class distribution. Logistic regression may be outperforming LDA due to the assumption of the normal distribution with common covariance for classes not being met.

Question 6 [8 marks] Compute the standard errors for the LDA discriminant coefficients using the bootstrap method.

Using the complete data and the model specified in 5(b), the standard errors for the LDA coefficients are computed as follows.

Variable	Coefficient	Bias	Standard Error
educ	0.5999297139	-0.000535607607	0.0117167410
hours	-0.0003993535	-0.000001466487	0.0000715112
exper	0.2708835625	-0.000195206514	0.0124506682
black1	-0.6295563563	0.002243812637	0.1237358676

Standard Errors for LDA coefficients with bootstrap (1000 times)

Question 7 [12 marks] Calculate the LOOCV error and the k-fold CV error for the logistic regression model at a cutoff of 0.6 for k = 10.

LOOCV and K-fold CV estimates for the logistic model can be computed as the following

CV Approach	Raw CV Error Estimate	Adjusted CV Error Estimate
LOOCV	0.1894495	0.1895062
10-Fold CV	0.1910550	0.1902523

Table 2: Cross-Validation Results

The LOOCV is lower because as LOOCV is often optimistic in its error estimate since the model is tested on a dataset that is very similar to the one it was trained.

In k-fold cross validation, the observations are randomly split up into k groups of similar size. Each of the k folds is used as a validation set while the other k-1 sets are used to generate estimates of the test error and the k-fold cv estimated test error is the average of the estimates. LOOCV is essentially k-fold cross validation where k is the number of observations.

There is a difference in the values of the error estimates because there are different levels of bias and variance accompanied with different values of k.

- K Fold CV test error estimates have a lower variance than the LOOCV estimates because more observations are included in the test set.
- K fold cv test error estimates have a greater bias than the LOOCV estimates because fewer observations are included in the training set.

However k-fold cross validation is more accurate than LOOCV because although it increases bias marginally, it decreases variance by a greater magnitude and thus outperforms LOOCV in the bias-variance trade-off.

1 R Code

```
##
options(scipen = 10)
library(car)
library(MASS)
library(class)
library(pROC)https://www.overleaf.com/project/6555f17734d0d9bf484a0b8e
library(boot)
###Loading and Exploring Data###
data <- read.csv("labour.csv")
attach(data)

head(data)
summary(data)
str(data)


###[Question 1, 5 Marks]###
correlation_matrix <- cor(data) #Computing Correlation Matrix for all variables in data
top_correlated_vars <- names(sort((correlation_matrix["wage",]),decreasing=TRUE)[2:5]) #Finding which 4 variables are the most correlated
print(top_correlated_vars) # Education, Year, Marrige status and Experience are the variables most correlated, other than wage itself
pairs(data[c("wage",top_correlated_vars)]) #Producing Pairwise Scatter Plot, See PDF for interpretation


###[Question 2, 5 Marks]###
indicators <- c("black", "hisp", "married", "union", "occ1", "occ2", "occ3", "occ4", "occ5", "occ6", "occ7", "occ8", "occ9")
data[indicators] <- lapply(data[indicators], as.factor) #Convert Indicators to factor variables
str(data) #Check to see if they are now factors
# Assuming 'married' is a factor variable and 'wage' is numeric

attach(data)
plot(married, wage,xlab = "Married", ylab = "Wage (Income earned £1000)",names = c("No", "Yes"))
#Graph a box plot with the variable married on the x-axis and wage on the y-axis.
#See PDF for Interpretation
```

```
###[Question 3, 8 Marks]###
lm.fit <- lm(wage ~ . - occ9,data) #Estimate a linear regression model with wage as DV and all other variables as IDV except occ9
summary(lm.fit) # Produce regression output
#See PDF for Answers to 3(a),3(b),3(c)


###[Question 4, 25 marks]###
lm.fit1 <- lm(wage ~ . - nr - hisp - occ8 -occ9,data) #Exclude nr, hisp, occ8 as they where not deemed significantly associated and exclude occ9
summary(lm.fit1) #Produce regression output

###[4a]###
par(mfrow=c(2,2))
plot(lm.fit1) #See PDF for interpretation of plots

###[4b]###
index.outlier <- which(rstudent(lm.fit1) > 3 | rstudent(lm.fit1) < -3)
length(index.outlier) #48 observations considered outlier
highlev_cutoff <- 3*(length(coef(lm.fit1)))/nobs(lm.fit1) #Highlev cuttof = 3*(p+1)/n, length(coef(lm.fit1))=p+1
index.highlev <- which(hatvalues(lm.fit1) > highlev_cutoff) #Identify observations with leverage values > Cuttoff
length(index.highlev)#3 observation considered high leverage points
#See PDF for inference on difference between High Leverage points and

###[4c]###
vif(lm.fit1) #Use vif function from car package, to create variance inflation factors for all predictor variables
#See PDF for inference of VIFs

###[4d]###
coef(lm.fit1)["hours"]  #Check to see if coefficient is negative
#See PDF for explanation.



###[Question 5, 37 marks]###
data$rich <- as.factor(ifelse(wage > mean(wage), 1, 0)) #Create Variable based on ifelse condition, and convert it to factor
test <- data[year <= 1982, ]
train<- data[year > 1982, ]
```

```
###[5a]###
logit.fit <- glm(rich ~ educ + hours + exper + black,train, family = 'binomial')
summary(logit.fit) #Produce Regression output info

logit.pred <- predict(logit.fit, test, type = 'response') #predict probabilities for Rich on test data

logit.pred.40 <- rep(0, dim(test)[1]) #Create variable full of 0s, length = length test data
logit.pred.40[logit.pred > 0.4] <- 1 #Assign values of 1, if probability >0.4
table(logit.pred.40,test$rich) #Create confusion matrix with predicted in rows, actuall in collumns
mean(logit.pred.40 != test$rich) #0.3834862 is test error rate for 0.4 cuttoff. Equivilant to (458+169)/1635

logit.pred.60 <- rep(0, dim(test)[1]) #Create variable full of 0s, length = length test data
logit.pred.60[logit.pred > 0.6] <- 1 #Assign values of 1, if probability >0.6
table(logit.pred.60,test$rich) #Create confusion matrix with predicted in rows, actuall in collumns
mean(logit.pred.60 != test$rich) #0.3107034 is test error rate for 0.8 cuttoff. Equivilant to (127+381)/1635

logit.pred.80 <- rep(0, dim(test)[1]) #Create variable full of 0s, length = length test data
logit.pred.80[logit.pred > 0.8] <- 1 #Assign values of 1, if probability > 0.8
table(logit.pred.80,test$rich) #Create confusion matrix with predicted in rows, actuall in collumns
mean(logit.pred.80 != test$rich) #0.2966361 test error rate for 0.8 cuttoff. Equivilant to (12+473)/1635

#Preffered Cuttoff is 0.8, it has the lowest overall test error.

table(logit.pred.40,test$rich)[1, 2] / sum(table(logit.pred.40,test$rich)[,2 ])
table(logit.pred.60,test$rich)[1, 2] / sum(table(logit.pred.60,test$rich)[,2 ])
table(logit.pred.80,test$rich)[1, 2] / sum(table(logit.pred.80,test$rich)[,2 ])
#0.4 has the lowest false negative error rate. See PDF for explanation why this may be the most preffered.

###[5b]###
lda.fit <- lda(rich ~ educ + hours + exper + black, train) #Fit LDA model on training data

lda.pred <- predict(lda.fit,test) #Predict Rich on test data with lda model

lda.pred.40 <- rep(0,dim(test)[1]) #Create variable full of 0s, length = length test data
lda.pred.40[lda.pred$posterior[,2] > 0.4] <- 1 #Assign value of 1 if posterior probability >0.4
table(lda.pred.40,test$rich) #Create confusion matrix
```

```
mean(lda.pred.40 != test$rich) #Calculate overall test error rate = 0.3816514

lda.pred.60 <- rep(0,dim(test)[1]) #Create variable full of 0s, length = length test data
lda.pred.60[lda.pred$posterior[,2] > 0.6] <- 1 #Assign value of 1 if posterior probability >0.6
table(lda.pred.60,test$rich) #Create confusion matrix
mean(lda.pred.60 != test$rich) #Calculate overall test error rate = 0.3094801

lda.pred.80 <- rep(0,dim(test)[1]) #Create variable full of 0s, length = length test data
lda.pred.80[lda.pred$posterior[,2] > 0.8] <- 1 #Assign value of 1 if posterior probability >0.6
table(lda.pred.80,test$rich) #Create confusion matrix
mean(lda.pred.80 != test$rich) #Calculate overall test error rate = 0.2954128
#0.8 cuttoff has lowest overall test error rate for lda model


###[5c]###
qda.fit <- qda(rich ~ educ + hours + exper + black, train) #Fit QDA model on training data

qda.pred <- predict(qda.fit,test) #Predict Rich on test data with qda model

qda.pred.40 <- rep(0,dim(test)[1]) #Create variable full of 0s, length = length test data
qda.pred.40[qda.pred$posterior[,2] > 0.4] <- 1 #Assign value of 1 if posterior probability >0.4
table(qda.pred.40,test$rich) #Create confusion matrix
mean(qda.pred.40 != test$rich) #Calculate overall test error rate = 0.3321101

qda.pred.60 <- rep(0,dim(test)[1]) #Create variable full of 0s, length = length test data
qda.pred.60[qda.pred$posterior[,2] > 0.6] <- 1 #Assign value of 1 if posterior probability >0.4
table(qda.pred.60,test$rich) #Create confusion matrix
mean(qda.pred.60 != test$rich) #Calculate overall test error rate = 0.3003058

qda.pred.80 <- rep(0,dim(test)[1]) #Create variable full of 0s, length = length test data
qda.pred.80[qda.pred$posterior[,2] > 0.8] <- 1 #Assign value of 1 if posterior probability >0.4
table(qda.pred.80,test$rich) #Create confusion matrix
mean(qda.pred.80 != test$rich) #Calculate overall test error rate = 0.2892966
#0.8 cuttoff has lowest overall test error rate for qda model

###[5d]###
set.seed(10)
```

```
knn.pred.1 <- knn(as.matrix(train), as.matrix(test), train$rich, k = 1) #Fit KNN model with k = 1
table(knn.pred.1, test$rich) #Confusion matrix
mean(knn.pred.1!= test$rich) #Calculate overall test error rate = 0.3492355

knn.pred.10 <- knn(as.matrix(train), as.matrix(test), train$rich, k = 10) #Fit KNN model with k = 10
table(knn.pred.10, test$rich) #Confusion matrix
mean(knn.pred.10!= test$rich) #Calculate overall test error rate = 0.3785933

knn.pred.20 <- knn(as.matrix(train), as.matrix(test), train$rich, k = 20) #Fit KNN model with k = 20
table(knn.pred.20, test$rich) #Confusion matrix
mean(knn.pred.20!= test$rich) #Calculate overall test error rate = 0.4024465

knn.pred.50 <- knn(as.matrix(train), as.matrix(test), train$rich, k = 50) #Fit KNN model with k = 20
table(knn.pred.50, test$rich) #Confusion matrix
mean(knn.pred.50!= test$rich) #Calculate overall test error rate = 0.4458716

knn.pred.100 <- knn(as.matrix(train), as.matrix(test), train$rich, k = 100) #Fit KNN model with k = 100
table(knn.pred.100, test$rich) #Confusion matrix
mean(knn.pred.100!= test$rich) #Calculate overall test error rate = 0.4525994

knn.pred.200 <- knn(as.matrix(train), as.matrix(test), train$rich, k = 100) #Fit KNN model with k = 200
table(knn.pred.200, test$rich) #Confusion matrix
mean(knn.pred.200!= test$rich) #Calculate overall test error rate = 0.4574924

#k=1 is prefered value, as it minimizes overall error rate

###[5e]###
#See PDF for explanation as to why this AUC of ROC curves can determine the best method.
logiROC<-roc(test$rich,logit.pred)
ldaROC<-roc(test$rich,as.numeric(lda.pred$class))
qdaROC <- roc(test$rich,as.numeric(qda.pred$class))
knnROC<-roc(test$rich,as.numeric(knn.pred.1))

ggroc(list(Logistic=logiROC,LDA = ldaROC,QDA = qdaROC,KNN = knnROC))
auc(logiROC)
auc(ldaROC)
```

```
auc(qdaROC)
auc(knnROC)

###[Question 6, 8 marks]###
boot.fn <- function(data,index){
  coef(lda(rich ~ educ + hours + exper + black, data,subset = index))
}

boot.fn(data,1:dim(data)[1]) #Test function on full data-set as specified in quesiton
boot(data,boot.fn,1000) # Use the boot function to get 1000 bootstrap estimates for the coefficients and standard errors

###[Question 6, 8 marks]###
cost <- function(rich, pi = 0) mean(abs(rich-pi) > 0.6) #Since Response is binary variable, create cost function with 0.6 cuttoff

glm.fit <- glm(rich ~ educ + hours + exper + black, data, family = 'binomial') #Fit Logistic model on full dataset as per question

LOOCV.err <- cv.glm(data, glm.fit, cost)$delta #calculate Leave one out cross validation model, this is esentially k fold with n folds
LOOCV.err

kfoldCV.err<- cv.glm(data, glm.fit, cost, K = 10)$delta #calculate 10 fold cross validation error
kfoldCV.err
#See PDF for full comparison and discussion of LOOCV and Kfold CV estimates of test error


...
```