Linear Models Group Project

Robyn Chilton, Mark Godfrey, Harry Pennock, James Strong, Lingyu Tan

November 2019

In this project we are trying to find a model that best fits the data set given in the *Project Beauty.csv* file. It is a different data set to that which is used in the *Beauty in the classroom* paper and so different results are potentially to be expected. Variable names used in our analysis correspond to the data as follows: Course Evaluation is *eval*, Tenured is *tenu*, Minority is *min*, Age is *age*, Female is *fem*, Students is *stud*, Lower is *low*, and Beauty is *pulch*.

1 Data Analysis

The first thing we do is to plot the response variable, *eval*, and all continuous explanatory variables, *pulch*, *stud*, and *age*, against each other to check for any potential relationships between them.

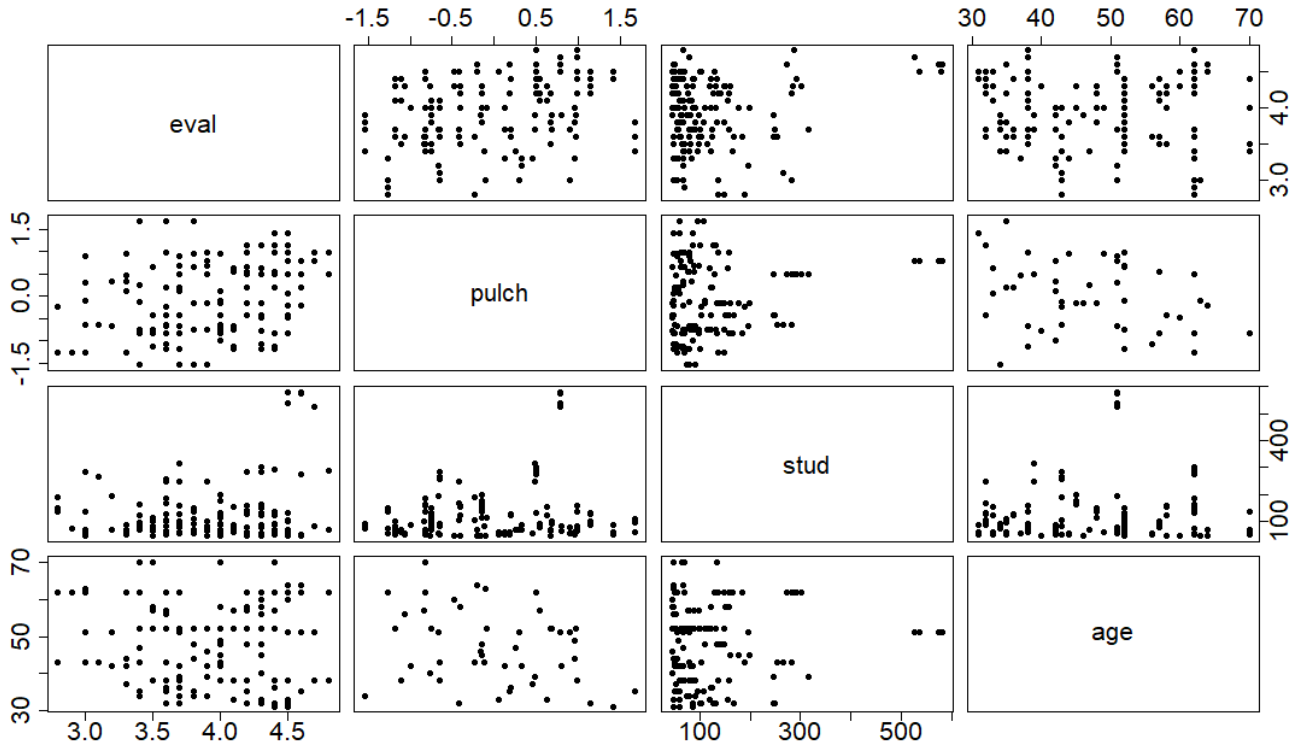

Figure 1: A plot of the response variable, Course Evaluation, *eval*, and all continuous explanatory variables, Beauty, Students, and Age, (*pulch*, *stud*, and *age* respectively), against each other.

From Figure 1, observing the scatter graph of *pulch* against *eval*, one may see that there appears to be a medium-strength, positive, linear relationship. Viewing *stud* against *eval*, a weak, positive relationship may be recognised. Upon studying *age* against *eval*, there doesn't appear to be a distinct relationship between the two.

In the interest of minimising multi-collinearity we will look at the relationships between the explanatory variables. The relationship in the *age* against *pulch* graph appears to be a medium-strength, negative,

linear relationship. There don't appear to be any distinct relationships between *stud* and *pulch* or *stud* and *age*. Hence, upon observing the scatter plots alone, it seems that correlation between pairs of explanatory variables is not going to be a problem, but to help determine the presence and magnitude of linear relationships between the variables, we shall check the correlation values. Correlation between pairs of variables (explanatory and response) can be seen in Table 1 below:

Correlations	eval	pulch	stud	age
eval	1.000	0.314	0.172	-0.126
pulch	0.314	1.000	0.148	-0.321
stud	0.172	0.148	1.000	0.116
age	-0.126	-0.321	0.116	1.000

Table 1: Table of correlation between pairs of variables (explanatory and response).

The correlation values back up our comments made upon inspection of the scatter graphs. The correlation value between *pulch* and *eval* is 0.314 which suggests a moderate, positive linear relationship between beauty and course evaluation, which would agree with the findings in the paper. There is also a weak positive relationship between *stud* and *eval* with a correlation value of 0.172 which may imply that as the number of students in a class increases, so does their average score given on course evaluation. The correlation value between *age* and *pulch* is -0.321 which suggests that as age increases, beauty decreases.

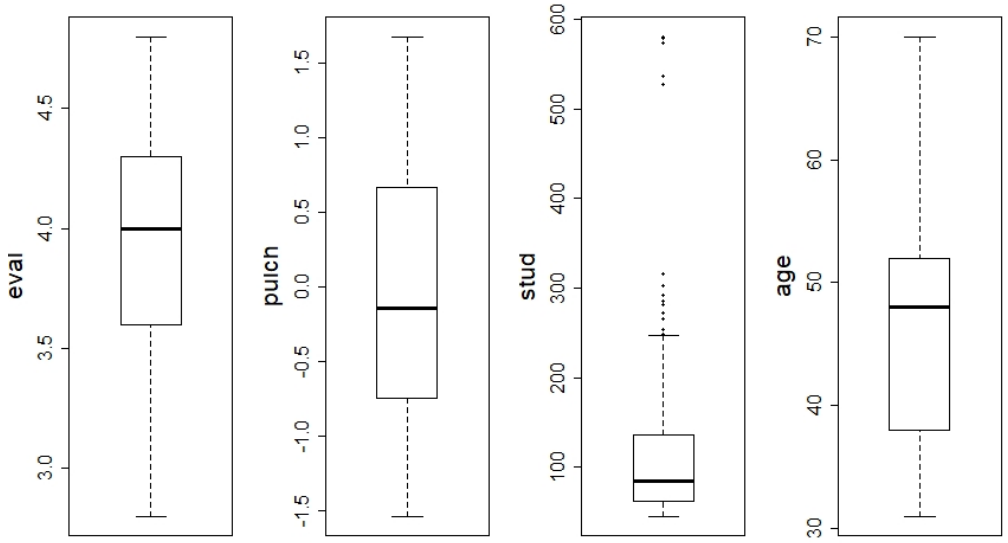

Figure 2: Box plots of the distributions of the response variable, Course Evaluation, *eval*, and all continuous explanatory variables, Beauty, Students, and Age, (*pulch*, *stud*, and *age* respectively).

From Figure 2, the distributions of evaluation scores, beauty scores, and age are fairly symmetric but the distribution of number of students is quite asymmetric so we will do a transformation of *stud* to try to attain a better, more symmetric distribution before continuing. We decided to use a natural log transformation (this gives a more symmetric distribution than a square root transformation) which gives the following distribution in Figure 3 (here *lnstud* denotes the Natural Log of Students variable).

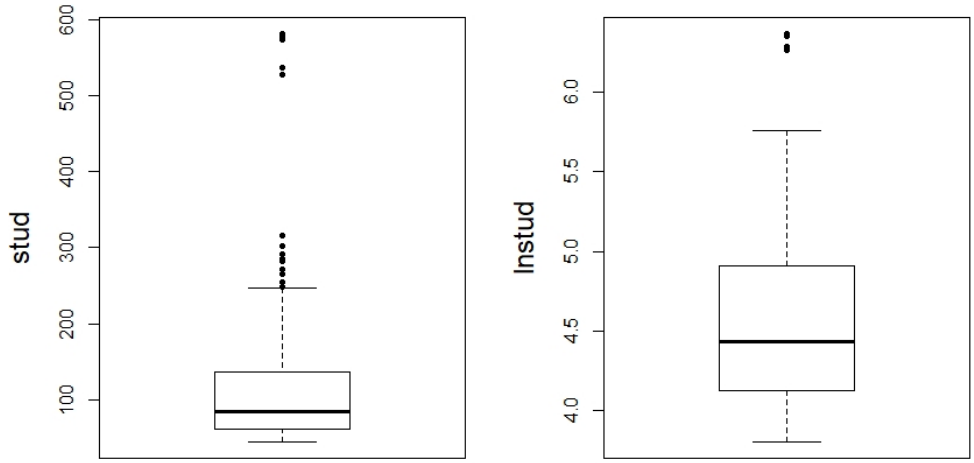

Figure 3: Box plots of Students, and the Natural Log of Students, (*stud*, and *lnstud* respectively).

lnstud is evidently much more symmetric than *stud* and so we will replace *stud* with *lnstud* henceforth whilst searching for a model.

Now we will produce our first model, *m1*, with the response variable as *eval*, and the explanatory variables as *tenu*, *min*, *age*, *fem*, *lnstud*, *low*, and *pulch* (in that order):

$$eval = \beta_0 + \beta_1 \times tenu + \beta_2 \times min + \beta_3 \times age + \beta_4 \times fem + \beta_5 \times lnstud + \beta_6 \times low + \beta_7 \times pulch + \varepsilon$$

We will now check the variance inflation factors (VIF) of these variables in *m1* to determine whether multi-collinearity will be a problem that may warrant the removal of variables. Table 2 shows these values for each of the explanatory variables:

tenu	min	age	fem	lnstud	low	pulch
1.532427	1.250070	1.942251	1.519571	1.379267	1.409053	1.360987

Table 2: Table of variance inflation factors of the variables in the model, *m1*.

One sees that none of the VIF values are particularly high and none are close to or above 4 which means that none of the variables are strongly correlated with each other. As a result of this, no variables need to be removed from a multi-collinearity standpoint.

Upon executing *summary(m1)* in R, we attain the following:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.996081	0.336659	11.870	< 2e-16
tenu	-0.183825	0.087486	-2.101	0.0374
min	-0.648082	0.132572	-4.889	2.69e-06
age	-0.003897	0.004403	-0.885	0.3776
fem	-0.070226	0.083555	-0.840	0.4020
lnstud	0.089690	0.064279	1.395	0.1651
low	-0.193540	0.080679	-2.399	0.0177
pulch	0.204918	0.047309	4.332	2.77e-05

Table 3: Summary of *m1* from R.

From Table 3 we can see that *min* and *pulch* are both highly significant with p-values of 2.69e-06 and 2.77e-05 respectively and *tenu* and *low* are significant at the 5% level with p-values of 0.0374 and 0.0177 respectively. The R^2 value for this model is 0.2853 which suggests that around 28% of the variation in the response variable is explained by variation in the explanatory variables. Also, the adjusted R^2 value is 0.2504 which is quite good.

We also check the residuals of this model. In plotting the fitted values and *age*, *lnstud*, and *pulch* explanatory variables against the studentised residuals we see that most points lie within [-2,2] and there is an even spread of points with no signs of changing variance. There are only 4 studentised residuals below -2, with some close to -3, and 1 above 2. The number of points outside isn't worrying as 5 out of 151 in total is within the 5% threshold but the proximity of some of these points to -3 is somewhat concerning, we will investigate this later on using Leverages and Cook's distances. There don't seem to be any systematic changes to the variance in any of the plots, and the spread of residuals is approximately uniform in all plots (except maybe the Natural Log of Students plot). In the Q-Q plot for this model the points stay close to the line with the expected ripples close to it - the only points that may be of concern are those in the tails. The p-value of the Anderson-Darling Normality Test for this particular model is 0.4327 which is high. From all this one may say that the constant variance and normality assumptions are validated.

Following on from this we will consider whether any points are influential and affect the model's coefficients unduly. Figures 4 and 5 are plots of the leverages and Cook's distances of all points based on this initial model, *m1*:

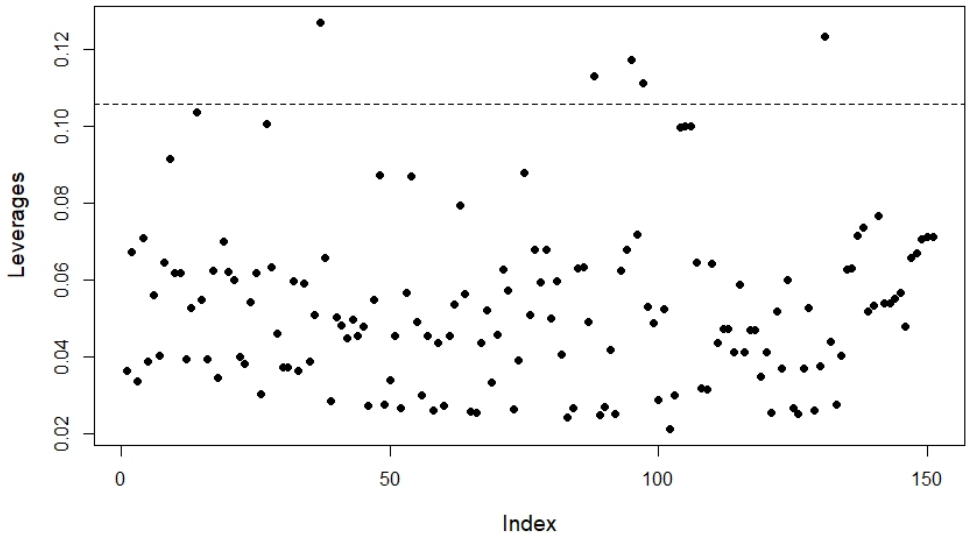

Figure 4: Scatter plot of the leverages of all points of *m1*.

The cut-off line for leverages is given by $\frac{2(p+1)}{n}$, in this case p (the number of explanatory variables in the model) is 7 and n (the sample size) is 151, placing our cut-off line at $\frac{16}{151} = 0.1059$ (3 d.p.). We can see that there are 5 points above the cut-off line. The fact that there are so many points with similarly high leverages suggests that leverages are not too much of a concern. It is still worthwhile checking Cook's distances to see if any of these points are particularly influential.

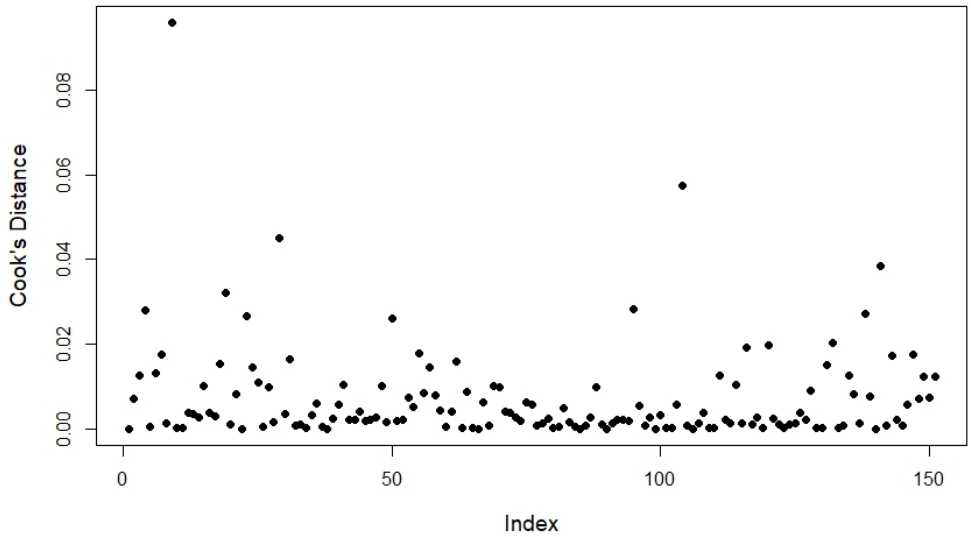

Figure 5: Scatter plot of the Cook's distance's of all points of *m1*.

We can see from Figure 5 that the most influential point is point 9 with Cook's distance of 0.096, followed by point 104 with Cook's Distance 0.057, these points could be having an overbearing influence on the coefficients of our model, so it will be worth carrying out sensitivity analysis with these points before we choose our final model. The rest of the points all have relatively low Cook's distances.

Now we will commence the process of removing variables from the model. We will start from a modified version of *m1* which has the variables in order from most to least significant. Upon doing this, the new model will be:

$$eval = \beta_0 + \beta_1 \times min + \beta_2 \times pulch + \beta_3 \times low + \beta_4 \times tenu + \beta_5 \times lnstud + \beta_6 \times age + \beta_7 \times fem + \varepsilon$$

2 Model Building

2.1 Backwards Elimination

This model has the same p-values for all variables from *summary* as they are just in a different order and so based on the p-values in the last table, we remove *fem* as it has the highest p-value at 0.4020. Upon doing *summary* again with variables in the same order but without *fem*, we see that *age* now has the highest p-value at 0.4918 and so we remove it. Doing *summary* again, now without *age*, we see that *lnstud* now has the largest p-value and so we remove it. If we do *summary* one more time, without *lnstud*, we see that all variables are now significant. The relevant information for all this is found in Tables 4, 5,and 6.

m2:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.860066	0.294912	13.089	< 2e-16
min	-0.682027	0.126141	-5.407	2.60e-07
pulch	0.206366	0.047299	4.369	2.37e-05
low	-0.187749	0.080303	-2.338	0.0208
tenu	-0.174794	0.086735	-2.015	0.0457
lnstud	0.102038	0.062514	1.632	0.1048
age	-0.002925	0.004244	-0.689	0.4918

Table 4: Summary of *m2*, (the model arrived to after removing Female as an explanatory variable from *m1*), from R.

m3:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.77506	0.26739	14.118	< 2e-16
min	-0.68005	0.12588	-5.402	2.63e-07
pulch	0.22030	0.04260	5.171	7.60e-07
low	-0.16962	0.07574	-2.240	0.02644
tenu	-0.20556	0.07423	-2.769	0.00635
lnstud	0.09346	0.06115	1.528	0.12861

Table 5: Summary of *m3*, (the model arrived to after removing Age as an explanatory variable from *m2*), from R.

m4:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.16965	0.06989	59.663	< 2e-16
min	-0.68624	0.12639	-5.430	2.3e-07
pulch	0.23013	0.04231	5.439	2.2e-07
low	-0.12045	0.06888	-1.749	0.0824
tenu	-0.18478	0.07331	-2.521	0.0128

Table 6: Summary of *m4*, (the model arrived to after removing Natural Log of Students as an explanatory variable from *m3*), from R.

Therefore, if we strictly follow the backward elimination until all the variables are significant, then our model would be:

$$eval = \beta_0 + \beta_1 \times min + \beta_2 \times pulch + \beta_3 \times low + \beta_4 \times tenu + \varepsilon$$

We notice, however, that after removing *lnstud*, there has been a slight decline in adjusted R^2 from 0.2546 to 0.2478 , suggesting that removing it may be unwise. We then check this using the Akaike Information Criterion and Mallow's C_p.

2.2 Akaike Information Criterion

The Akaike Information Criterion (AIC) is another way of testing which variables may be removed from a model. The *steps* command in R performs these calculations. Upon doing this using *direction = "both"* we come upon the model:

Start: $AIC = -262.44$
$eval \sim min + pulch + low + tenu + lnstud + age + fem$

Step: $AIC = -263.7$
$eval \sim min + pulch + low + tenu + lnstud + age$

Step: $AIC = -265.2$
$eval \sim min + pulch + low + tenu + lnstud$

From the AIC output we can see that removing Female and Age has lowered AIC, removing any additional variables increases AIC and so this procedure comes to a different model when compared to the model from backwards elimination. Using AIC *direction = "forward"* and *direction = "backward"* also result in the same model being suggested.

$$eval = \beta_0 + \beta_1 \times min + \beta_2 \times pulch + \beta_3 \times low + \beta_4 \times tenu + \beta_5 \times lnstud + \varepsilon$$

2.3 Mallow's C_p

p	8	7	6	5
C_p	8	6.7	5.2	5.5

Table 7: Rows p and C_p corresponding to the total number of variables in a model including the intercept and Mallow's C_p (to 1 d.p.) respectively.

The values in Figure 7 for C_p were calculated by hand with, $n = 151$, $m = 7$, and $R^2_{m,adj} = 0.2504$ (obtained from *summary*). The model with $p = 8$ is the ordered *m1* model, $p = 7$ is *m2*, $p = 6$ is *m3*, and $p = 5$ is *m4*. We see that C_p is less than p until $p = 5$ at which point we have $C_p = 5.5 > 5$. This tells us that the adjusted R^2 value decreases from *m3* to *m4* ($p = 6$ to $p = 5$) and so we should take note that less of the variation in the response variable is explained relative to the number of variables the model contains in *m4* than in *m3*. If we follow what Mallow's C_p is saying then our model going forward would be:

$$eval = \beta_0 + \beta_1 \times min + \beta_2 \times pulch + \beta_3 \times low + \beta_4 \times tenu + \beta_5 \times lnstud + \varepsilon$$

2.4 Model Determination

From these three methods of removing variables, we decide to follow the model produced by both Mallow's C_p and AIC (forward, backward and in both directions), i.e. include Natural Log of Students as a variable: this model produces a larger value of adjusted R^2 and only manual backwards elimination invokes the removal of Natural Log of Students.

Having, once again, looked at the leverages and Cook's distances, this time of all points in *m3* we concluded that no points affected the model's coefficients unduly.

2.5 Interaction Analysis

Two-way interaction	p-value
min:pulch	0.555
min:low	0.777
min:tenu	0.624
min:lnstud	0.632
pulch:low	0.023
pulch:tenu	0.096
pulch:lnstud	$2.10e-07$
low:tenu	0.926
low:lnstud	0.294
tenu:lnstud	0.567

Three-way interaction	p-value
min:pulch:low	0.665
min:pulch:tenu	0.974
min:pulch:lnstud	0.553
min:low:tenu	0.818
pulch:low:tenu	0.033
pulch:low:lnstud	0.970
pulch:tenu:lnstud	0.357
low:tenu:lnstud	0.145

Table 8: Tables containing the p-values (to 3d.p.) for two-way and three-way interactions when added by themselves to the end of the model *m3* as measured from the *anova* command.

Table 8 contains the p-values for interactions when added by themselves to the end of the model called *m3* as measured from the *anova* command. The only significant interactions by this method are *pulch*:*low*, *pulch*:*lnstud* and *pulch*:*low*:*tenu*.

Response: eval	Df	Sum Sq	Mean Sq	F value	Pr(>F)
min	1	2.3685	2.3685	16.9055	6.664e-05
pulch	1	5.1231	5.1231	36.5678	1.271e-08
low	1	0.3956	0.3956	2.8240	0.095095
tenu	1	1.0649	1.0649	7.6013	0.006611
lnstud	1	0.3880	0.3880	2.7692	0.098333
pulch:lnstud	1	4.1218	4.1218	29.4207	2.489e-07
pulch:low	1	0.0117	0.0117	0.0836	0.772848
pulch:tenu	1	0.0312	0.0312	0.2227	0.637741
low:tenu	1	0.0057	0.0057	0.0404	0.841075
pulch:low:tenu	1	0.2992	0.2992	2.1358	0.146137
Residuals	140	19.6140	0.1401		

Table 9: Table containing the p-values for *m3* and all the singly significant interactions from most to least significant added to the end measured from the *anova* command in R.

From Table 9 we see that the only remaining significant interaction is *pulch*:*lnstud* and so we retain that interaction and discard the rest to obtain the model *m5*:

$$eval = \beta_0 + \beta_1 \times min + \beta_2 \times pulch + \beta_3 \times low + \beta_4 \times tenu + \beta_5 \times lnstud$$
$$+\beta_6 \times pulch : lnstud + \varepsilon$$

If we use the *summary* command with *m5* then R outputs Table 10 below:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.133220	0.252960	16.339	< 2e-16
min	-0.610576	0.115704	-5.277	4.73e-07
pulch	-1.597863	0.335695	-4.760	4.67e-06
low	-0.173358	0.069193	-2.505	0.01334
tenu	-0.178055	0.068003	-2.618	0.00978
lnstud	0.006625	0.058091	0.114	0.90936
pulch:lnstud	0.405670	0.074395	5.453	2.10e-07

Table 10: Summary output for *m5* from R.

The *m5 summary* command output seen in Table 10 above shows the p-values for all of the included variables. All p-values are significant except for that of *lnstud* which is 0.90936. The R^2 of the model is 0.4028 which means that around 40% of the variation in the response variable is explained by variation in the explanatory variables. The adjusted R^2 of the model is 0.3779 which is quite high and takes into account the number of variables included in the model. The p-value of the interaction term, *pulch*:*lnstud* is highly significant at 2.10e-07.

2.6 Model Adequacy

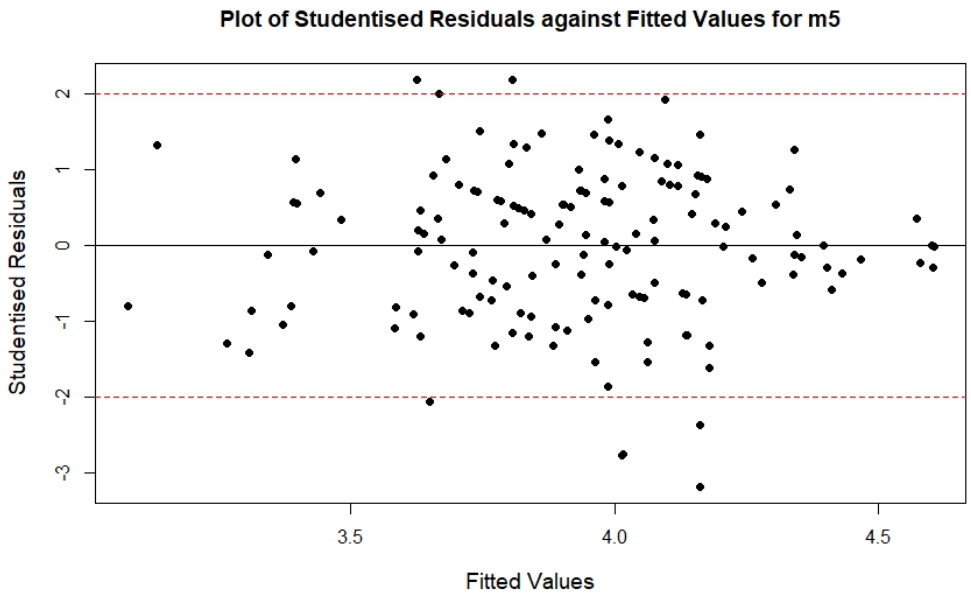

Figure 6: Scatter plot of Studentised Residuals against Fitted Values of *m5*.

From Figure 6, there is an approximately uniform spread of points across the plot and there are no signs of systematically changing variance. Seven points find themselves outside of the [-2,2] region with three of them above 2 and the other 4 below -2, 3 of these close to -3. 7 out of 151 is within the 5% threshold so this is not concerning. Influential points were checked for the first model and we found that points 9 and 104 had the highest Cook's distances. We shall check for influential points again for our final model.

The Q-Q plot below in Figure 7 shows the points lie close to the line the preponderance of the time with fairly large deviations appearing exclusively at the tails. There appear to be shallow ripples about the line

but this is expected and they are shallow for the most part. The Anderson-Darling statistic is A = 0.44094 with a fairly large p-value of 0.2866. From all of this, we determine that the normality assumption is valid.

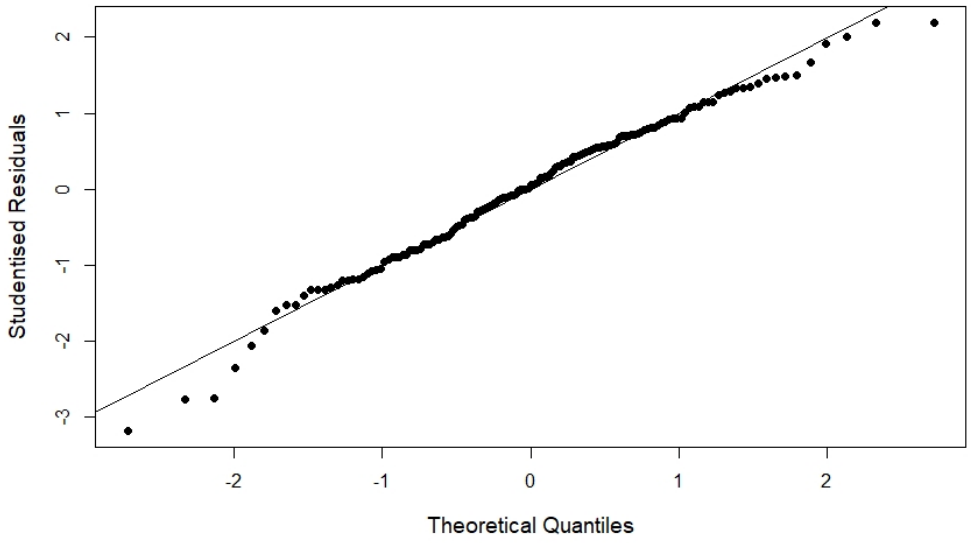

Figure 7: Normal Q-Q Plot of *m5*.

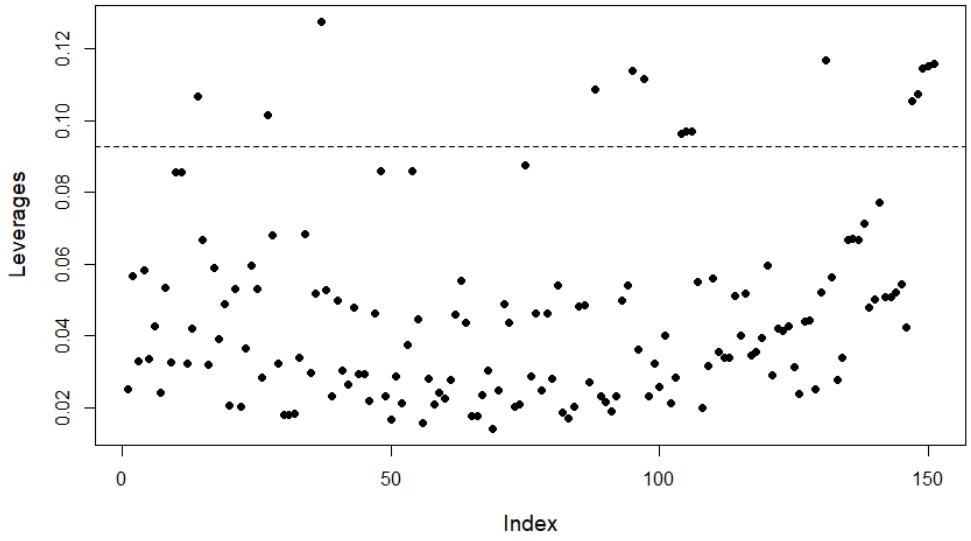

Figure 8: Leverages for *m5*.

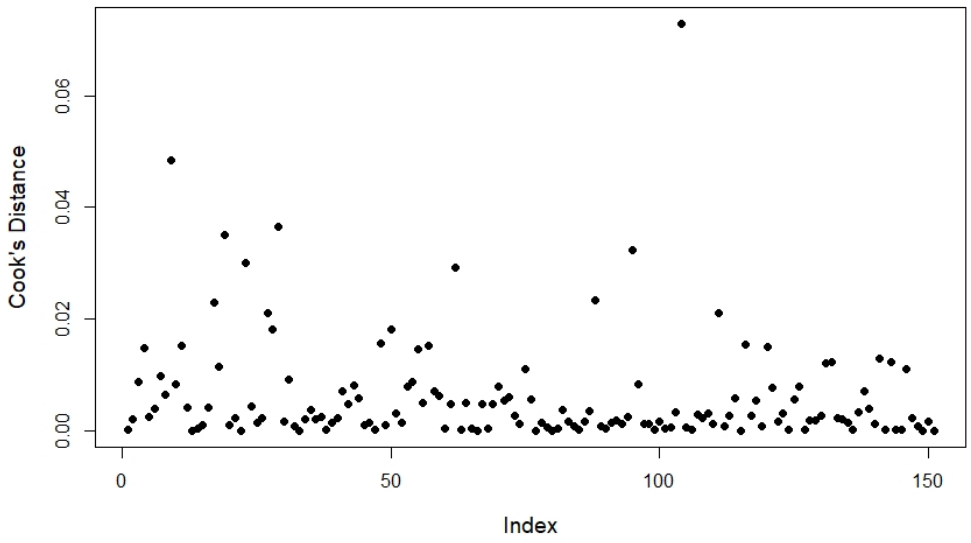

Figure 9: Cook's distances for *m5*.

We can see from Figure 9 that the most influential point is point 104 with Cook's distance of 0.073. This point was highlighted when we originally investigated the Cook's distances (for *m1*) and had a Cook's distance of 0.057, so this point has become more influential.

Carrying out sensitivity analysis to investigate the importance of this point, we remove it from the data set and refit the model. We found that removing the point did not significantly change the values of any of the coefficients, and so in the interest of retaining the full data set we decided not to remove the point. None of the other points really stand out in the Cook's distances plot.

Having plotted the leverages in Figure 8, we have no concerns about their values since none of the points stand out as being greatly higher than others. Moreover, point 104, the point with the largest Cook's distance is not above the cut-off line (at 14/151), the case for not removing it is thus strengthened.

3 Conclusion

The final model based on this analysis is *m5*:

$$\widehat{eval} = 4.133 - 0.611 \times min - 1.598 \times pulch - 0.173 \times low - 0.178 \times tenu + 0.007 \times lnstud$$

$$+0.406 \times pulch : lnstud,$$

where the coefficients are rounded to (3 d.p.).

Adjusted R^2 has a value of 0.3779 in this model which is quite good. R^2 has a value of 0.4028 in this model suggesting that around 40% of variation in Course Evaluation is explained by variation in Minority, Beauty, Lower, Tenured and Natural Log of Students and the interaction between Beauty and Natural Log of Students. This seems like a fairly high percentage of variation to be explained by those variables since you would assume that actual lecturing technique would be of greater significance, i.e. whether they use slides or write the notes by hand, how difficult the module is, or whether the lectures are recorded etc. Hence, these percentages could be unrealistically high.

The coefficient for Beauty is negative which disagrees with the paper; this at first glance seems strange however it could be explained by students being envious of good-looking lecturers. This could also be

partially explained by the positive coefficient for the interaction term between Beauty and Natural Log of Students. The coefficient for Natural Log of Students is relatively small and quite close to 0 mainly due to the 2-way interaction between Beauty and Natural Log of Students.

All of the other coefficients have the same sign as the model in the paper. The coefficient of Minority being negative could be explained by a language barrier between the students and the lecturer. The negative coefficient of lower can possibly be explained by foundation year courses being simpler so students could be less focussed on the content and more focussed on the lecturers' teaching style. The coefficient of Tenured being negative could potentially be explained since tenured lecturers don't have to try as hard in order to keep their job.

The interaction term suggests that the slope of Beauty changes as the slope of Natural Log of Students changes. Additionally, having included an intersection related to beauty, the coefficient of beauty, -1.598, cannot simply be explained as the unique effect of beauty on evaluation. In fact, the coefficient of beauty becomes $0.406 \times lnstud - 1.598$, which means the larger the number of students is, the more influential the beauty will be, and vice-versa. It may be that, in those larger classes, students are more likely to be distracted and cannot focus on the lecturer. So when the lecturer is beautiful, the students are more willing to focus on what he/she is talking and thus score better.

4 Appendix

Each member of the group has modelled the data in R to ensure that we all agreed on one model. Each member of the group added to the Latex document and we all edited it together.

Approved: _____
Robyn Chilton

Approved: _____
Mark Godfrey

Approved: _____
Harry Pennock

Approved: _____
James Strong

Approved: _____
Lingyu Tan

Everyone should get equal marks for this project, i.e. everyone did about the same level of work.