Name: Kitu Komya
Student ID: 404-491-375

Please read the following and provide your signature to indicate that you agree to and understand these terms. If you do not understand, please ask us to clarify.

I agree that I will not seek any help from any human, robot, artificial intelligence, or internet resource. I understand that the only resources I can consult the TA (Ryan), the professor, the online textbook, and any resources provided on the CCLE (excluding Piazza.)

Signature: Kitu Komya

This final exam is based on data provided by the Wall Street Journal (WSJ) to assess the problem of high student loans. I believe, but am not certain, that the data set includes only non-profit educational institutions. (This is important, because some researchers have found that for-profit institutions have a much greater loan default rate.) If you have any questions about the context of the data (what the variables mean, what a student loan is, what it means to default) please ask Prof. Gould. The data come from 2012. For the most part, you will be working with a random sample of the WSJ's data set.

Your answers should be typed directly into this file. Keep your answers short and precise. Long, rambling answers will lose points. You can copy-and-paste any graphs directly into this file.

You must also provide a .R file that contains documented code. Be sure to indicate which question the code belongs to.

1) The model shown below explains the median payment amount for student loans as a function of the graduation rate, the default rate, the net price of the institution, the change in the net price over the last year, the percent of students who transfered out of the institution, and the undergraduate enrollment.

```
> summary(m1)

Call:
lm(formula = MEDIAN_PAYMENT ~ GRAD_RATE + DEFAULT_RATE + NETPRICE +
    NETPRICE_CHANGE + TRANSFEROUT_RATE + UG_ENROLLMENT, data = wsj)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-266.620 -30.620  -1.413  30.705 313.934

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.855e+01  3.756e+00  23.575   <2e-16 ***
GRAD_RATE         5.405e-01  5.022e-02  10.764   <2e-16 ***
DEFAULT_RATE     -2.127e+00  1.045e-01 -20.357   <2e-16 ***
NETPRICE          5.237e-03  1.381e-04  37.913   <2e-16 ***
NETPRICE_CHANGE  -5.150e-02  2.948e-02  -1.747   0.0808 .
TRANSFEROUT_RATE -1.744e-01  7.293e-02  -2.392   0.0168 *
UG_ENROLLMENT    -8.189e-06  1.105e-04  -0.074   0.9409
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.35 on 3387 degrees of freedom
  (596 observations deleted due to missingness)
Multiple R-squared:  0.5613,  Adjusted R-squared:  0.5606
F-statistic: 722.4 on 6 and 3387 DF,  p-value: < 2.2e-16---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.49 on 3482 degrees of freedom
  (503 observations deleted due to missingness)
Multiple R-squared:  0.5398,  Adjusted R-squared:  0.539
```

a) (4) Write an interpretation of the slope for DEFAULT_RATE. Assume the model is valid.

Among individuals in which all of the other variables are held constant, those with a default rate of 1% more on average have a median payment of $2.13 less. Since the p-value is significant at a significance level of 0.05, this interpretation is significant.

b) (4) A junior analyst suggests fitting a new model that excludes both UG_ENROLLMENT and NETPRICE_CHANGE. Your boss asks you if this is a good idea. What answer would you give your boss?

This is a bad idea because removing variables may result in a change in the p-values of other variables which would in turn remove significant variables. We must use theory to determine which variables belong in the model, and in this case, it makes sense to keep these variables by using financial intuition. These variables likely have significant interaction with the other variables leading us to conclude that by themselves the insignificant variables may not seem useful, but as a whole group they are.

c) (4) The same junior analyst suggests that if you had entered UG_ENROLLMENT into the model first, then it might have had a smaller p-value in the summary table given above. Your boss wants your opinion on this. What would you tell him?

First, I would convince my boss to consider firing this incompetent junior analyst. The summary function is independent of order, but in ANOVA it does matter. The summary table does not change test statistics because its function is to compute the significance of a variable given that the other variables are already in the model.

d) (4) The ANOVA table below is based on the model whose output is below. Some of the entries have been accidentally erased and replaced with "xxxxxx". Replace the x's with the correct values. The total sums of squares is 19572874.

```
> anova(m1)
Analysis of Variance Table

Response: MEDIAN_PAYMENT
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
GRAD_RATE	1	5901908	5901908	2328.2618	< 2e-16 ***
DEFAULT_RATE	1	706725	706725	278.7982	< 2e-16 ***
NETPRICE	1	4356516	4356516	1718.6153	< 2e-16 ***
NETPRICE_CHANGE	1	**7442**	**7442**	**2.9357**	0.08672 .
TRANSFEROUT_RATE	1	14567	14567	5.7465	0.01658 *
UG_ENROLLMENT	1	14	14	0.0055	**0.940886**
Residuals	3387	8585702	2535		

2) Consider the model
DEFAULT_RATE = B_0 + B_1*NETPRICE

At a significance level of 0.05, NETPRICE is not significantly associated with the default rate, which could lead us to conclude that the default rate does not depend on the price of the institution-- a surprising result, since you'd expect more expensive institutes to require their students to borrow more, and so to therefore have a higher default rate. However, that annoying junior analyst has a theory. Perhaps graduates of different types of institutions have different earning potentials. And so graduates with bachelor degrees earn more and so can better pay off their loans. Since their education might also cost more, we need to control for the type of institution to see if the price is associated with the default rate.

a) (2) Give the R-code for a model that explains the default rate as a function of the net price, controlling for the institution group (INSTITUTION_GROUP).

lm(DEFAULT_RATE~NETPRICE+INSTITUTION_GROUP, data = wsj)

b) (4) Fit the model. (You don't need to report the output). On average, how does the mean default rate of bachelor-degree institutions differ from the mean default rate at associate-degree granting institutions?

Given that all of the other variables are held constant, on average the mean default rate of bachelor-degree institutions is 9.875% less than the rate at associate-degree granting institutions.

c) (4) State the value of the intercept and interpret, assuming the model is valid.

The intercept is 16.65. The intercept is the mean y value when all predictors are set to 0. Here that means finding the average default rate with no NETPRICE. When the INSTITUTION_GROUP variable is set to 0, we must be referring to associate-degree granting institutions, the only INSTITUTION_GROUP not in the summary. So, we conclude that the predicted mean value of default rate at associate-degree granting institutions is 16.65%.

d) (2) Again, assuming a valid model, report a 95% confidence interval for the mean default rate of bachelor-degree granting institutions with a net price of 15000.

[8.448469%, 9.564715%]

e)(4) A politician says that bachelor degrees cost too much. Since everyone knows, the politician says, that the default-rate on student loans is higher for schools with higher cost, bachelor-degree institutions should lower their costs. How would you respond to the politician, using the results from this model?

The politician's claim is actually wrong. When fitting the model DEFAULT_RATE = B_0 + B_1*NETPRICE, we learned that at a significance level of 0.05, NETPRICE is not significantly associated with the default rate, which lead us to conclude that the default rate does not depend only on the price of the institution. Instead we learned that a default rate depends on multiple variables, such as the INSTITUTION_GROUP. When INSTITUTION_GROUP is added to the model, then NETPRICE does become significant. Moreover, in general, since correlation does not imply causation, we cannot expect that lowering the net price will cause a default rate to decrease. We have merely observed an association.

3) The junior analyst fit the following model:
DEFAULT_RATE~MEDIAN_BORROWING+MEDIAN_PAYMENT+NETPRICE+GRAD_RATE.

a) (2) Report the variance inflation factors:

MEDIAN_BORROWING: 672268500
MEDIAN_PAYMENT: 672269100
NETPRICE: 1.789688

GRAD_RATE: **1.482188**

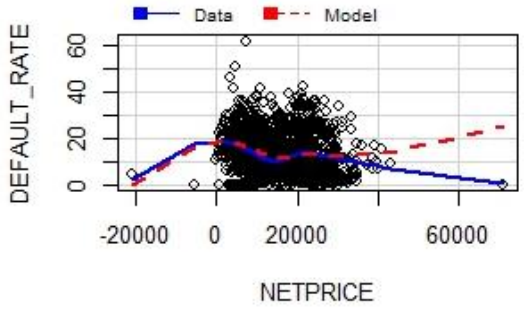
b) (4) Based on the variance inflation factors, what advice would you give to the junior analyst for improving the model? Explain to your boss why your suggestion is an improvement (without looking at diagnostic plots).

The rule of thumb to ensure that collinearity is not a severe problem is for all variance inflation factors to be less than 5. Thus, we have a problem with MEDIAN_BORROWING and MEDIAN_PAYMENT which are clearly beyond the cut-off since these two variables are correlated which will lead bias in our conclusions since the p values will be inflated. With this in mind, we could justify omitting one of these variables from our model since its information is redundant and fully contained in the other variable. In our case, let's omit MEDIAN_PAYMENT and keep MEDIAN_BORROWING in our model.

4) Improve the junior analysts model in Question 3 by fitting a new model following the suggestion you made in (3).

a) (2) Give the R code for this model:
lm(DEFAULT_RATE~MEDIAN_BORROWING+NETPRICE+GRAD_RATE, data = wsj)

b) (4) Now, provide the marginal model plot for NETPRICE (copy and paste into this document). What improvements to the model does this plot suggest?



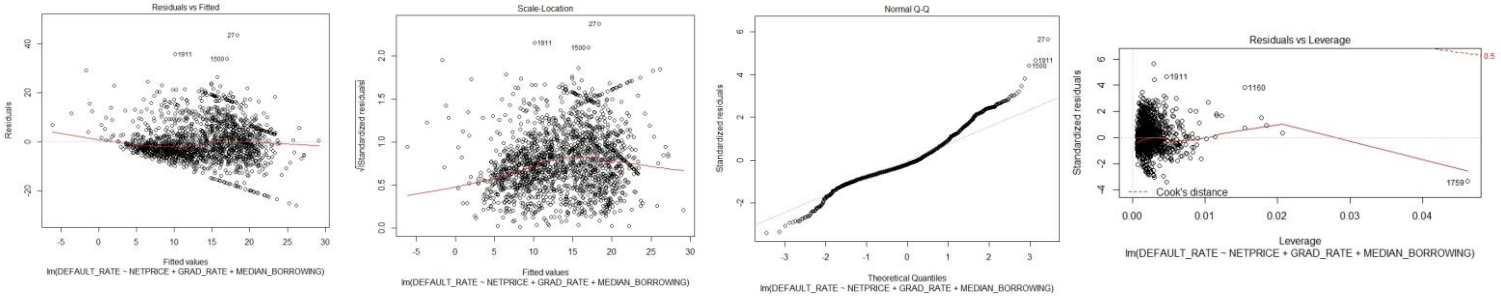
The marginal model plot demonstrates that the nonparametric estimates given as a solid blue curve is markedly different from the smooth fitted values shown as dashed red lines. The influential point is skewing the data so no model will work until that has been removed. But in the meantime, if we try a square root transformation on the response variable, our model will be bettered.

c) (2) Give the R-code to fit an improved model based on your answer to (b).

lm(sqrt(DEFAULT_RATE)~MEDIAN_BORROWING+NETPRICE+GRAD_RATE, data = wsj)

5) Fit a model that explains the DEFAULT_RATE as a function of NETPRICE, GRAD_RATE and MEDIAN_BORROWING.

a) (8) Comment on the validity of the model. Include any and all necessary plots. Be sure to comment on every plot that you provide. (In other words, don't provide a plot unless you think it is helpful for checking the validity of the model.)



From left to right, we have the residual plot, the scale-location plot, the normal QQ plot, and the residuals vs leverage plot. From the residual plot we can see that the residuals are, for the most part, randomly scattered with no obvious pattern. The red smoother line is decently flat. This meets our linearity assumption as well as our constant variance assumption.

The scale-location plot further affirms that a linear model is appropriate and it also follows the constant variance assumption as it too has data points that are scattered randomly without any obvious trend. The red smoother, although not perfectly flat, is still in decent shape.

The normal QQ plot, however, is a line that seems to deviate substantially from the linear line at its extrema. This leads us to question the normality of the errors.

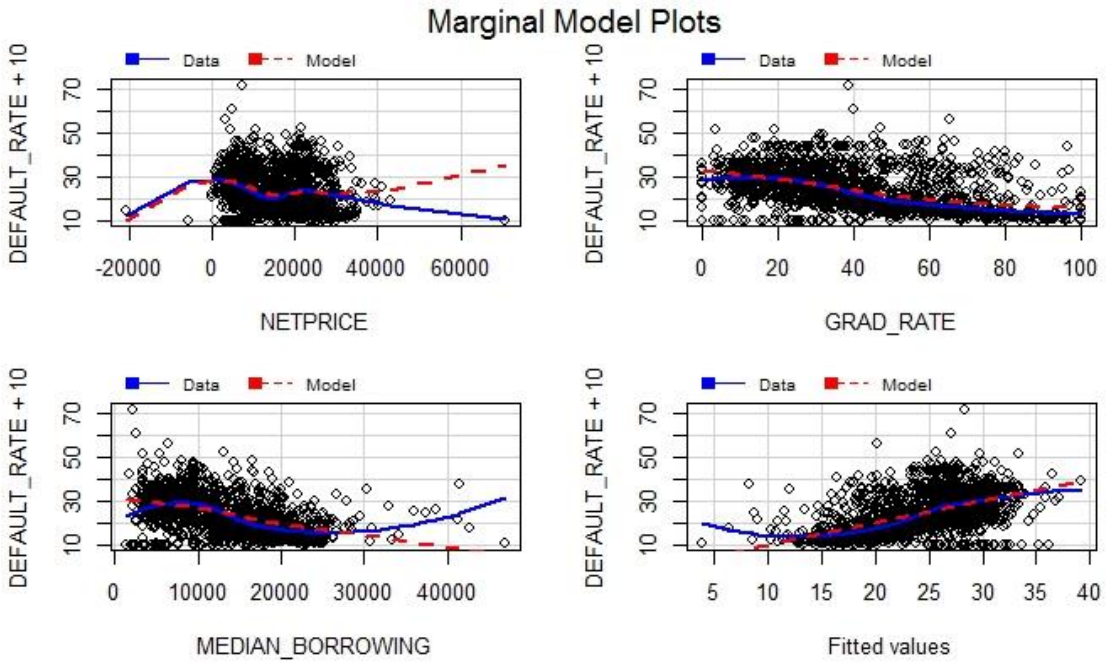
In the residuals vs leverage plot, we see that there is only one high leverage point, but since it falls within the standardized residual interval of [-2, 2], it is a good high leverage point.

Looking at the variation inflation factors, we see:

NETPRICE = 1.78248
GRAD_RATE = 1.479761
MEDIAN_BORROWING = 1.939163

Since none of these values are above 5, we can be assured that collinearity is not a problem in our data.

In the summary statistics, we see that the adjusted r square value is 0.311 which is a very weak value. This lends the possibility of finding a better model.

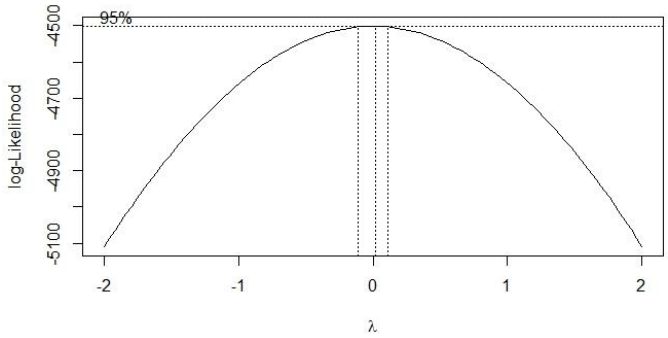


The marginal model plot demonstrates that the nonparametric estimates in the net price given as a solid blue curve is still markedly different from the smooth fitted values shown as dashed red lines. The influential point is still skewing the data so no model will work until that has been removed. We have a similar situation in the median borrowing marginal model plot. However, in the graduation rate and the fitted values, the marginal model plot demonstrates that the nonparametric estimates given as a solid blue curve is indistinguishable from the smooth fitted values shown as dashed red lines, leading us to believe that this is a somewhat okay model.

Although this model is valid, it could be better since its primary weakness lies in its possible violation of the normality of the errors.

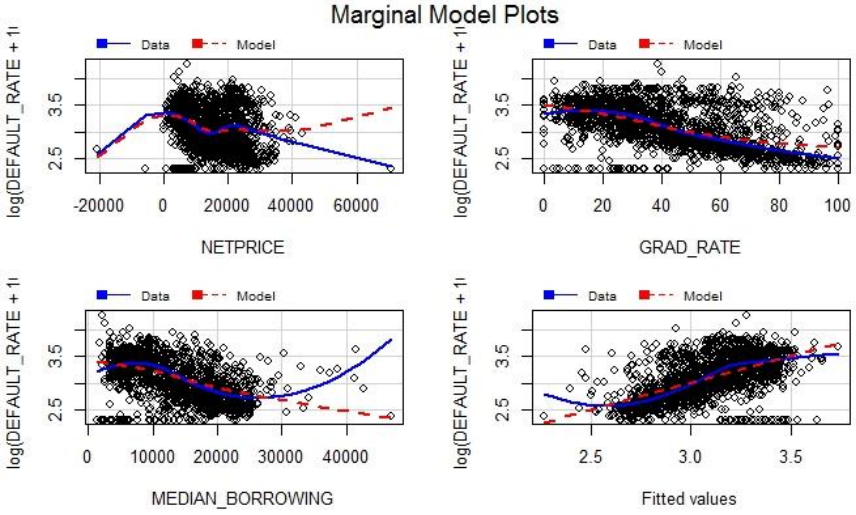
b) (4) Suggest a transformation for the **response variable only**. Explain why you think this will be a useful transformation. (Your explanation should not be based on actually fitting the transformed model). Provide any needed graphs or summary statistics.

We shall use the Box-Cox approach to transform the response variable. This will be a useful transformation because as seen in 5a, a linear model is not the ideal model, and with Box-Cox we can fix the weaknesses we saw and thus fit a better model.

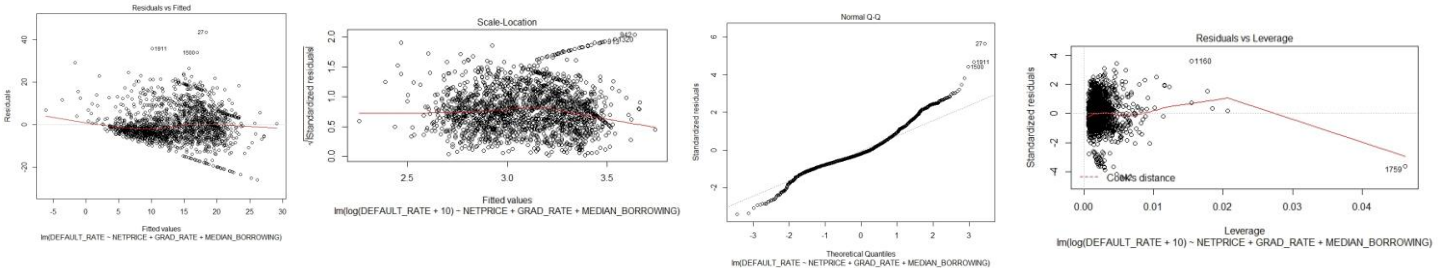


The plot above shows a 95% confidence interval between the dotted vertical lines. We observe that the interval falls in negative and positive lambda values, making us wary. Using the summary statistics, we learn that the estimate for lambda is 0.0013. Moreover, at a lambda = 0, the high p value suggests that we fail the reject the null hypothesis so it's best to do a log transform, which is consistent with the estimate for lambda. And at a lambda = 1, the low p value suggests that we reject the null hypothesis and thus a transformation is necessary. Therefore, we conclude a log transform on the response variable is needed.

c) (8) Follow your suggestion and transform the response variable. Does it improve the validity of the model? Explain, including any necessary plots and summary statistics.



The marginal model plot demonstrates that the nonparametric estimates in the net price given as a solid blue curve is still markedly different from the smooth fitted values shown as dashed red lines. The influential point is still skewing the data so no model will work until that has been removed. We have a similar situation in the median borrowing marginal model plot. However, in the graduation rate and the fitted values, the marginal model plot demonstrates that the nonparametric estimates given as a solid blue curve is indistinguishable from the smooth fitted values shown as dashed red lines, leading us to believe that this is a somewhat better model.



From left to right, we have the residual plot, the scale-location plot, the normal QQ plot, and the residuals vs leverage plot. From the residual plot we can see that the residuals are very much randomly scattered with no obvious pattern. The red smoother line is decently flat. This meets our linearity assumption as well as our constant variance assumption.

The scale-location plot further affirms that a linear model is appropriate and it also follows the constant variance assumption as it too has data points that are scattered randomly without any obvious trend. The red smoother, is very flat as well.

The normal QQ plot is a line that deviates substantially less from the linear line at its extrema in comparison to the non-transformed model. This is affirmation that this model is much better than the non-transformed model since we can meet the normality of the errors assumption.

In the residuals vs leverage plot, we see that there is only one high leverage point, but since it falls within the standardized residual interval of [-2, 2], it is a good high leverage point.

The adjusted r square value is 0.3425 which means that about 34.25% of the variation in the default rate variable is explained by our model. This value is an improvement from the adjusted r square value of 0.311 from the non-transformed model.

Looking at the variation inflation factors, we see:

NETPRICE = 1.78248

GRAD_RATE = 1.479761
MEDIAN_BORROWING = 1.939163

Since none of these values are above 5, we can be assured that collinearity is not a problem in our data. Thus, our model is appropriate for our data, and is in fact, better than the non-transformed model, as seen by the adjusted r square value.