XIMAN LIU

DATA ANALYSIS AND REGRESSION

Assignment-5 | Total Points: 35 pts for DSC 423 and 25 pts for DSC 323
Due Date: 5/3/2021 by 11:59 pm

Note:

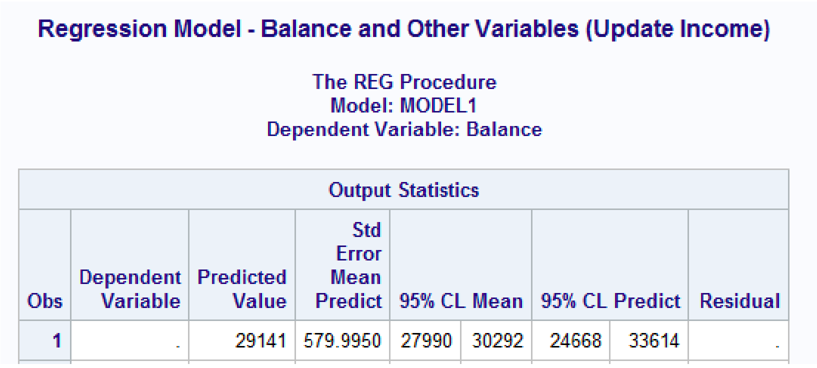
- All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
- No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (8-days) to submit your assignments.
- Submitting work that is not yours is grounds for an automatic 'F' for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
- SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won't be able to help you with DePaul software related issues.

***Note:* For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.**

Problem 1 [5 pts] – to be answered by everyone

You will continue the prediction, confidence interval and prediction interval for the **banking** dataset that was analyzed in Assignment 4. Since you would have altered the dataset to exclude outliers/influential points and/or multicollinearity, use the dataset and the code that was used to generate your final model. Note: Make sure you rerun the whole banking code from assignment 4, before you do this last part.

a) Use the fitted regression model from Assignment 4 to predict the average bank balance for a specific zip code area where there is a plan to open a new branch. Census data in that area show the following values: median age is 34 years, median education is 13 years, median income is $89,000, median home value is $160,000, median wealth is 140,000. Using SAS, compute the predicted average bank balance, 95% confidence interval and prediction interval for your estimate. Make sure to use SAS coding to determine the values. Include all relevant outputs. Discuss your findings.



Regression Model - Balance and Other Variables (Update Income)

The REG Procedure
Model: MODEL1
Dependent Variable: Balance

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	29141	579.9950	27990	30292	24668	33614	.

The final model removed income and six observations which were outliers and influential points.

With the values of census data in that area, the predicted average bank balance is $29,141. The predicted average bank balance is within 95% of confidence interval and between $27,990 and $30,292.

b) Copy and paste your FULL SAS code into the word document along with your answers.

```
TITLE "Analysis - Bankingfull";

PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment5\Bankingfull.txt"
out=Balance replace;
getnames=yes;
delimiter='09'x;
RUN;

DATA Balance_new2;
set Balance;
drop Income;
RUN;

DATA Balance_new3;
set Balance_new3;
if _n_=58 then delete;
RUN;

DATA Balance_pred;
input Balance Age Education HomeVal Wealth;
datalines;
. 34 13 160000 140000
;

DATA Balance_new3;
set Balance_pred Balance;

PROC PRINT;
RUN;

PROC REG;
title "Regression Model - Balance and Other Variables (Update Income)";
model Balance= Age Education HomeVal Wealth /p clm cli alpha=0.05;
RUN;
```

PROBLEM 2 [20 pts] – to be answered by everyone

This problem asks you to build a model for the college dataset (college.csv) that contains the following variables:

School	*School name*
Private	*public/private indicator. YES if university is private, NO if university is public.*
Accept.pct	*percentage of applicants accepted*
Elite10	*Elite schools with majority of students from the top 10% of their high school class (0- Not Elite, 1-Elite)*

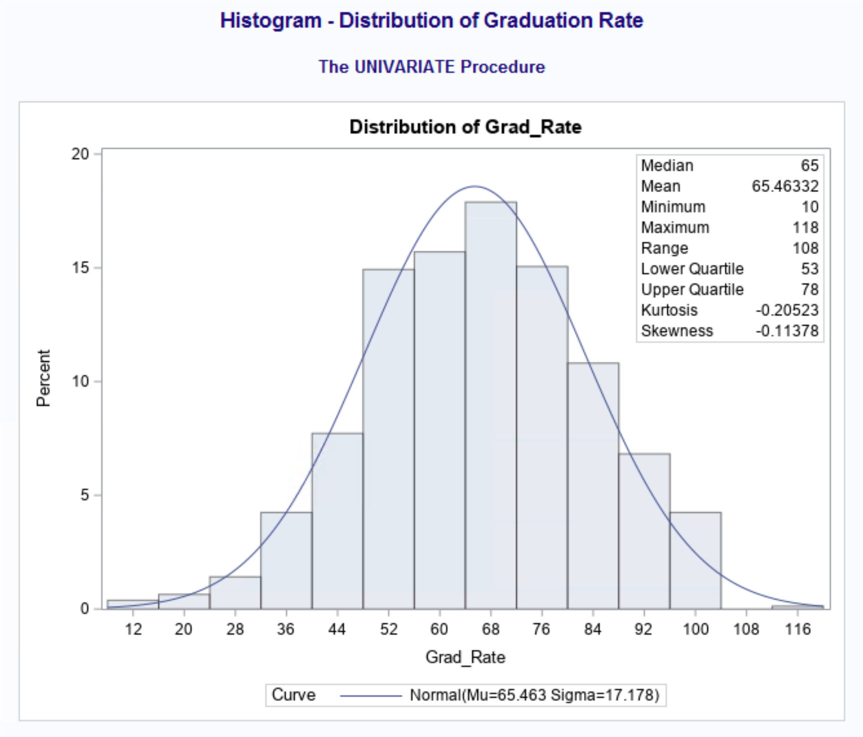
F.Undergrad	*number of full-time undergraduate students*
P.Undergrad	*number of part-time undergraduate students*
Outstate	*Out-of-state tuition*
Room.Board	*room and board costs*
Books	*estimated book costs*
Personal	*Estimated personal spending*
PhD	*Percent of faculty with PhD*
Terminal	*Faculty with terminal degrees (terminal degree is a university degree that is either highest on the academic track or highest on the professional track in a given field of study)*
S.F.Ratio	*Student/faculty ratio*
perc.alumni	*Percent of alumni who donate*
Expend	*Instructional expenditure per student*
Grad.Rate	*Graduation rate in 4 years*

Apply regression analysis techniques to analyze the relationship among the observed variables and build a model to predict Graduation Rates (Grad.Rate). **Note: Depending on how you import you data (INFILE or IMPORT) the SAS may relabel the column names. Make sure to use the variable names that appear when you use a proc print.**

Note: Before you start, open the college.csv file, and examine the data.

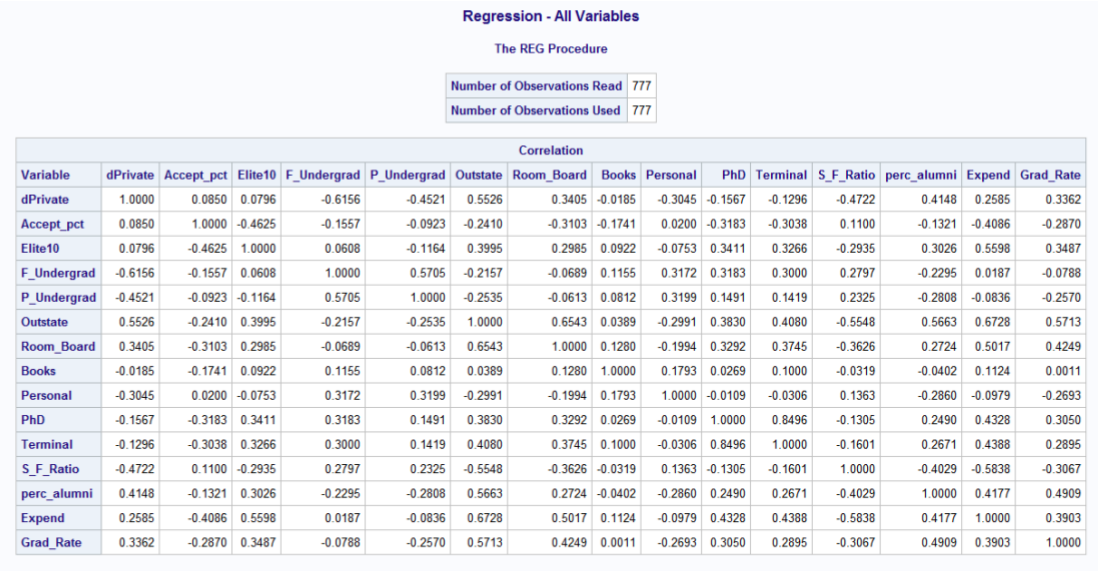
Answer the following questions.

a) Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation (This is the data exploration stage, therefore use the appropriate statics to explore your data).



Overall selected colleges, the median and mean of graduation rate is around 65% which is a high number. And these two numbers are quite similar. In that case the distribution is normal and symmetric.
The minimum graduation rate is 10%, and the maximum graduation rate is 118%, and they show a wide range of 108%.
The median which is 65% located in the middle of Q1(53%) and Q3(78%).
Transformation is not needed here because distribution is not skewed.

b) Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission).



Regression - All Variables

The REG Procedure

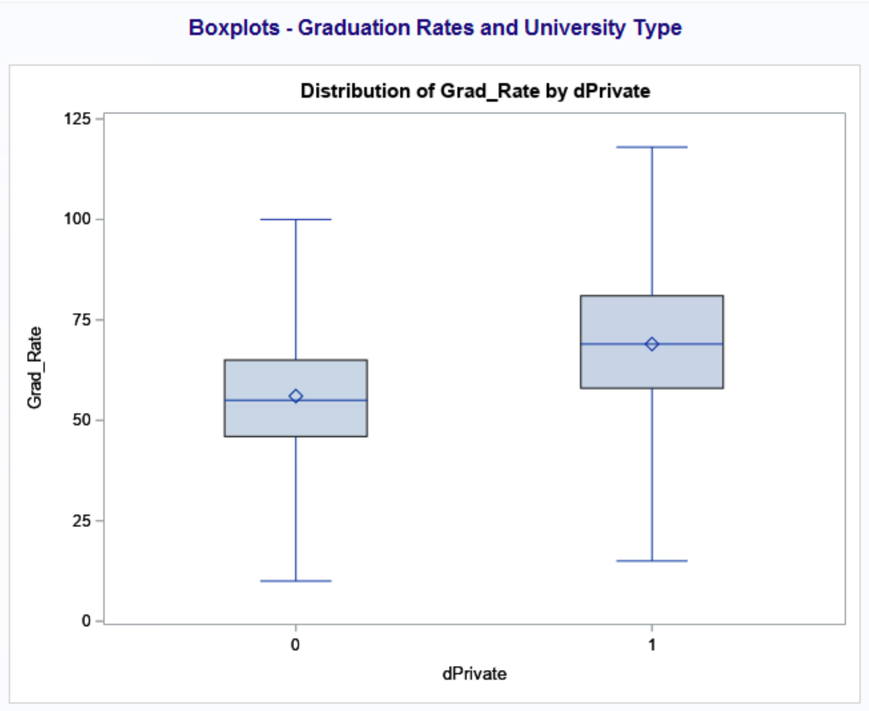
Number of Observations Read	777
Number of Observations Used	777

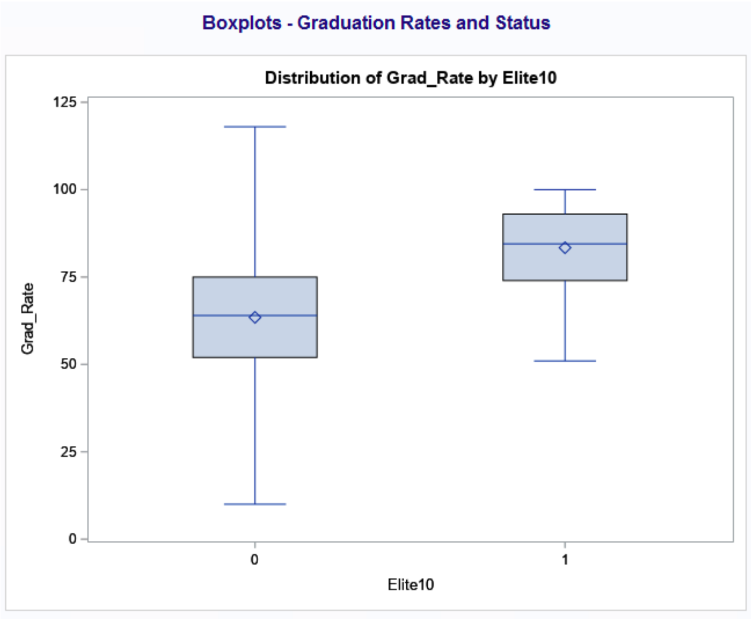
Correlation

Variable	dPrivate	Accept_pct	Elite10	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal	S_F_Ratio	perc_alumni	Expend	Grad_Rate
dPrivate	1.0000	0.0850	0.0796	-0.6156	-0.4521	0.5526	0.3405	-0.0185	-0.3045	-0.1567	-0.1296	-0.4722	0.4148	0.2585	0.3362
Accept_pct	0.0850	1.0000	-0.4625	-0.1557	-0.0923	-0.2410	-0.3103	-0.1741	0.0200	-0.3183	-0.3038	0.1100	-0.1321	-0.4086	-0.2870
Elite10	0.0796	-0.4625	1.0000	0.0608	-0.1164	0.3995	0.2985	0.0922	-0.0753	0.3411	0.3266	-0.2935	0.3026	0.5598	0.3487
F_Undergrad	-0.6156	-0.1557	0.0608	1.0000	0.5705	-0.2157	-0.0689	0.1155	0.3172	0.3183	0.3000	0.2797	-0.2295	0.0187	-0.0788
P_Undergrad	-0.4521	-0.0923	-0.1164	0.5705	1.0000	-0.2535	-0.0613	0.0812	0.3199	0.1491	0.1419	0.2325	-0.2808	-0.0836	-0.2570
Outstate	0.5526	-0.2410	0.3995	-0.2157	-0.2535	1.0000	0.6543	0.0389	-0.2991	0.3830	0.4080	-0.5548	0.5663	0.6728	0.5713
Room_Board	0.3405	-0.3103	0.2985	-0.0689	-0.0613	0.6543	1.0000	0.1280	-0.1994	0.3292	0.3745	-0.3626	0.2724	0.5017	0.4249
Books	-0.0185	-0.1741	0.0922	0.1155	0.0812	0.0389	0.1280	1.0000	0.1793	0.0269	0.1000	-0.0319	-0.0402	0.1124	0.0011
Personal	-0.3045	0.0200	-0.0753	0.3172	0.3199	-0.2991	-0.1994	0.1793	1.0000	-0.0109	-0.0306	0.1363	-0.2860	-0.0979	-0.2693
PhD	-0.1567	-0.3183	0.3411	0.3183	0.1491	0.3830	0.3292	0.0269	-0.0109	1.0000	0.8496	-0.1305	0.2490	0.4328	0.3050
Terminal	-0.1296	-0.3038	0.3266	0.3000	0.1419	0.4080	0.3745	0.1000	-0.0306	0.8496	1.0000	-0.1601	0.2671	0.4388	0.2895
S_F_Ratio	-0.4722	0.1100	-0.2935	0.2797	0.2325	-0.5548	-0.3626	-0.0319	0.1363	-0.1305	-0.1601	1.0000	-0.4029	-0.5838	-0.3067
perc_alumni	0.4148	-0.1321	0.3026	-0.2295	-0.2808	0.5663	0.2724	-0.0402	-0.2860	0.2490	0.2671	-0.4029	1.0000	0.4177	0.4909
Expend	0.2585	-0.4086	0.5598	0.0187	-0.0836	0.6728	0.5017	0.1124	-0.0979	0.4328	0.4388	-0.5838	0.4177	1.0000	0.3903
Grad_Rate	0.3362	-0.2870	0.3487	-0.0788	-0.2570	0.5713	0.4249	0.0011	-0.2693	0.3050	0.2895	-0.3067	0.4909	0.3903	1.0000

dPrivate and Elite10 are qualitative variables (dummy variables) which their scatter plots only have 0 and 1. In that case it is meaningless for association.
Overall, F_Undergrad and Books represent no relationship with Grade-Rate because points are lacking pattern.
Accept_pct, P_Undergrad, Personal and S_F_Ratio represent low negative relationship with Grade_Rate because points are lack of solid patterns. However, we can tell when one variable increases, the other variable will decrease.
PHD, Terminal and Expend represent low positive relationship with Grade_Rate because points are lack of solid patterns. However, we can tell when one variable increases, the other variable will increase.
Outstate and Room_Board represent relatively high positive relationship with Grade_Rate because points are solid patterns. However, we can tell when one variable increases, the other variable will increase.

c) Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Include the boxplots and discuss your findings. (See SAS Procedures section on D2L if you need the code to generate a boxplot).

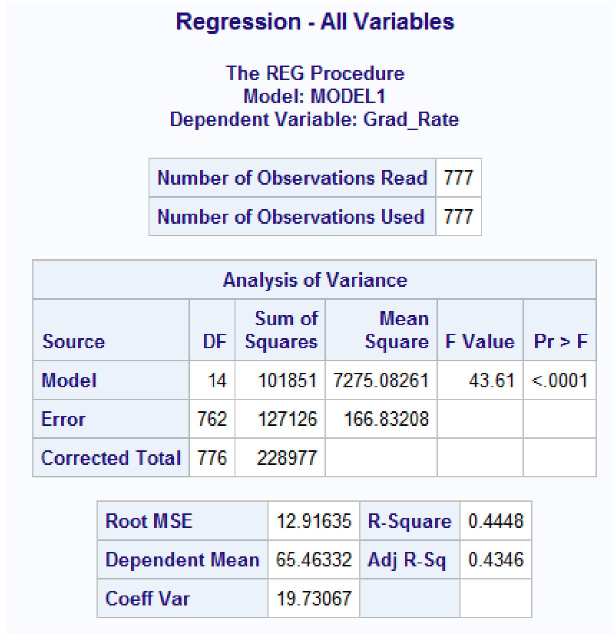




We can definitely tell that the middle of public (dPrivate = 0) and private (dPrivate = 1) in selected universities have the totally different situations. Upper quartile (75%) of public universities is under 65%. Vice versa, private universities have higher lower quartile (25%) even than the median number of public universities. 75% (Q3) of private universities has the graduation rate with 80%. There is a wide range of graduation rate of private universities from 15% to 120%. The medians of both public and private universities are quite similar to their means. Private universities have overlapping mean and median which represents the distribution of graduation rate is normal and symmetric. For public universities, the mean is higher than the median which means the distribution of graduation rate is gently skewed right.

For the second graph, the middle of non-elite (Elite10=0) and elite (Elite10=1)) in selected universities have the totally different situations. Non-elite universities have wide range of graduation rate because the box is right in the between of upper and lower extremes. Vice versa, elite universities have higher graduation rate because its box is closer to upper extreme. 75% (Q3) of elite universities are in the range of 90%. Vice versa, non-elite universities have the box between 50% to 75%. Overall, non-elite universities have longer whisker since the graduation rate is from 10% to 120%. The means for both universities are both slightly lower than its medians, showing that the distribution of graduation rate of public universitas is gently skewed left.

d) Fit a full model (with all independent variables) to predict Grad.Rate. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values. Include the relevant output.



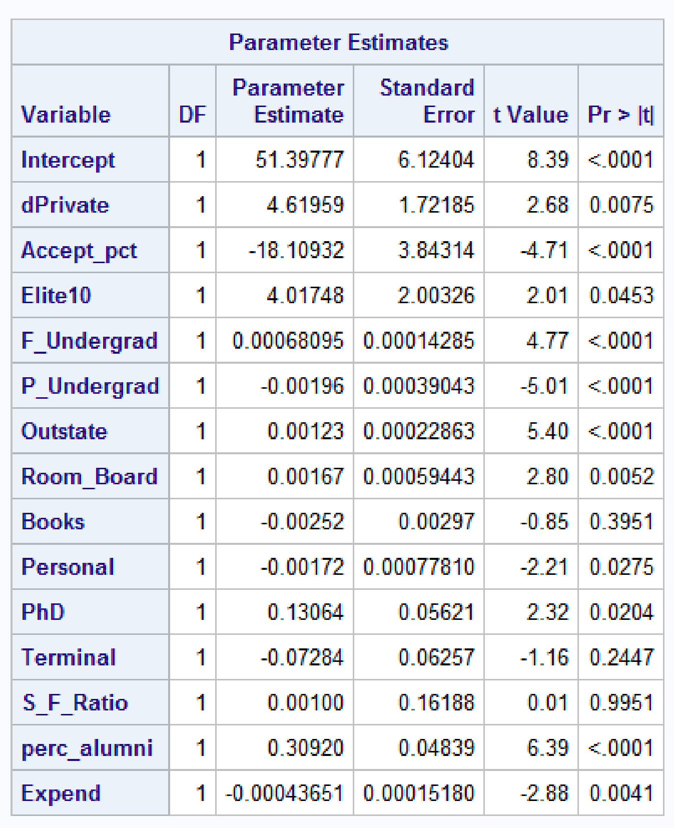
Regression - All Variables

The REG Procedure
Model: MODEL1
Dependent Variable: Grad_Rate

Number of Observations Read	777
Number of Observations Used	777

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	101851	7275.08261	43.61	<.0001
Error	762	127126	166.83208		
Corrected Total	776	228977			

Root MSE	12.91635	R-Square	0.4448
Dependent Mean	65.46332	Adj R-Sq	0.4346
Coeff Var	19.73067		



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51.39777	6.12404	8.39	<.0001
dPrivate	1	4.61959	1.72185	2.68	0.0075
Accept_pct	1	-18.10932	3.84314	-4.71	<.0001
Elite10	1	4.01748	2.00326	2.01	0.0453
F_Undergrad	1	0.00068095	0.00014285	4.77	<.0001
P_Undergrad	1	-0.00196	0.00039043	-5.01	<.0001
Outstate	1	0.00123	0.00022863	5.40	<.0001
Room_Board	1	0.00167	0.00059443	2.80	0.0052
Books	1	-0.00252	0.00297	-0.85	0.3951
Personal	1	-0.00172	0.00077810	-2.21	0.0275
PhD	1	0.13064	0.05621	2.32	0.0204
Terminal	1	-0.07284	0.06257	-1.16	0.2447
S_F_Ratio	1	0.00100	0.16188	0.01	0.9951
perc_alumni	1	0.30920	0.04839	6.39	<.0001
Expend	1	-0.00043651	0.00015180	-2.88	0.0041

Full model:

Predicted Grad_Rate = 51.39777 + 4.61959 * dPrivate – 18.10932 * Accept_pct + 4.01748 * Elite10 + 0.13064 * PhD – 0.07284 * Terminal + 0.30920 * Perc_alumni

The model initially has 14 variables. The higher the beta weight, the more important it effects on gradiuation rate. F_Undergrad, P_Undergrad, Outstate, Room_Board, Books, Personal, S_F_Ratio and Expend are excluded since they have almost nothing effects.

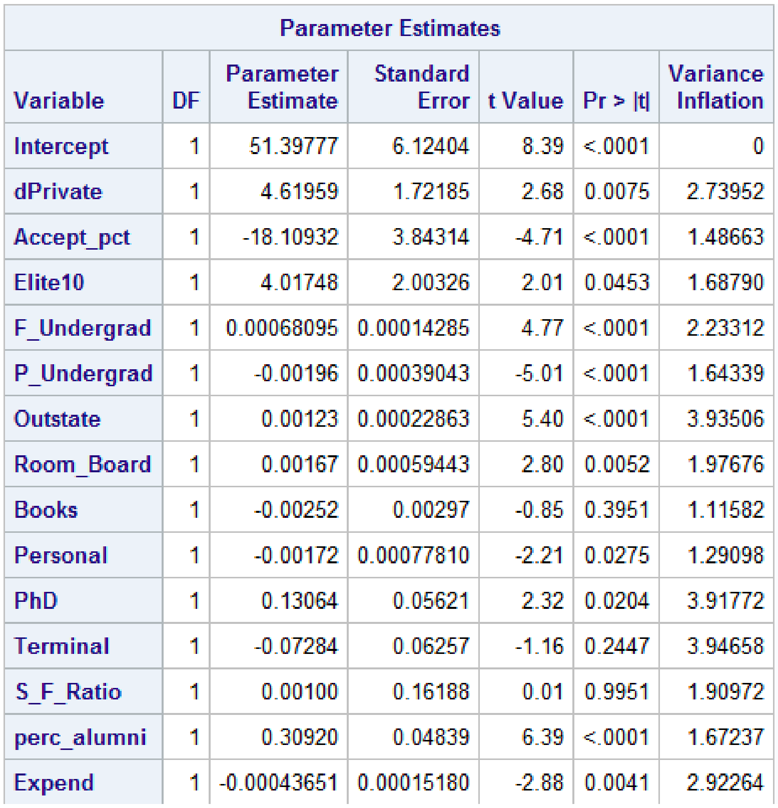
Then we use t-test on variables. The p-value for Books, Terminal and S_F_Ratio are larger than 0.05, so we cannot reject the hypothesis that Books, Terminal and S_F_Ratio do not have significant effect on graduation rate. Besides, all the other variables all have significant influence on graduation rate, which p-value are smaller than 0.05.

Null hypothesis:
H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
Alternative hypothesis:
H_a: At least one coefficient $\beta_j \neq 0$

Test statistic:
F = 43.61 and with p-value less than 0.001 (at alpha=0.05). The null hypothesis of no association between graduation rate and other variables is rejected. At least one x-variable has a significant effect on changes in graduation rate. F-test gives strong support to the fitted model.

R2 (44.48%) and adj R2 (43.46%) indicate the amount of variation in graduation rate explained by the regression model. However, adj-R2 has a relatively low rate which indicated that this is not a fairly ideally model.

e) Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. Include the relevant output and discuss your answer.



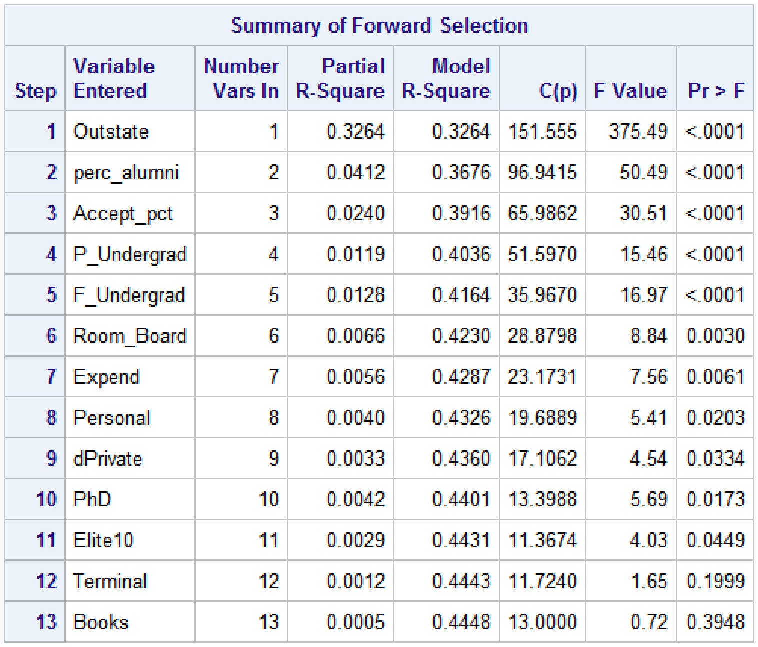
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	51.39777	6.12404	8.39	<.0001	0
dPrivate	1	4.61959	1.72185	2.68	0.0075	2.73952
Accept_pct	1	-18.10932	3.84314	-4.71	<.0001	1.48663
Elite10	1	4.01748	2.00326	2.01	0.0453	1.68790
F_Undergrad	1	0.00068095	0.00014285	4.77	<.0001	2.23312
P_Undergrad	1	-0.00196	0.00039043	-5.01	<.0001	1.64339
Outstate	1	0.00123	0.00022863	5.40	<.0001	3.93506
Room_Board	1	0.00167	0.00059443	2.80	0.0052	1.97676
Books	1	-0.00252	0.00297	-0.85	0.3951	1.11582
Personal	1	-0.00172	0.00077810	-2.21	0.0275	1.29098
PhD	1	0.13064	0.05621	2.32	0.0204	3.91772
Terminal	1	-0.07284	0.06257	-1.16	0.2447	3.94658
S_F_Ratio	1	0.00100	0.16188	0.01	0.9951	1.90972
perc_alumni	1	0.30920	0.04839	6.39	<.0001	1.67237
Expend	1	-0.00043651	0.00015180	-2.88	0.0041	2.92264

There is no VIF of any x-variables higher than 10. In that case, multicollinearity is not a problem here.

f) Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate. You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R^2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, include the outputs.

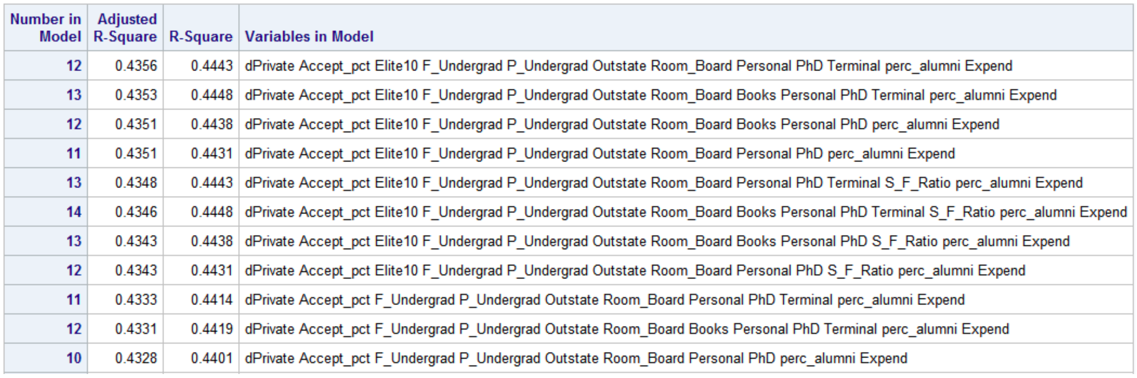
Forward Selection



Summary of Forward Selection

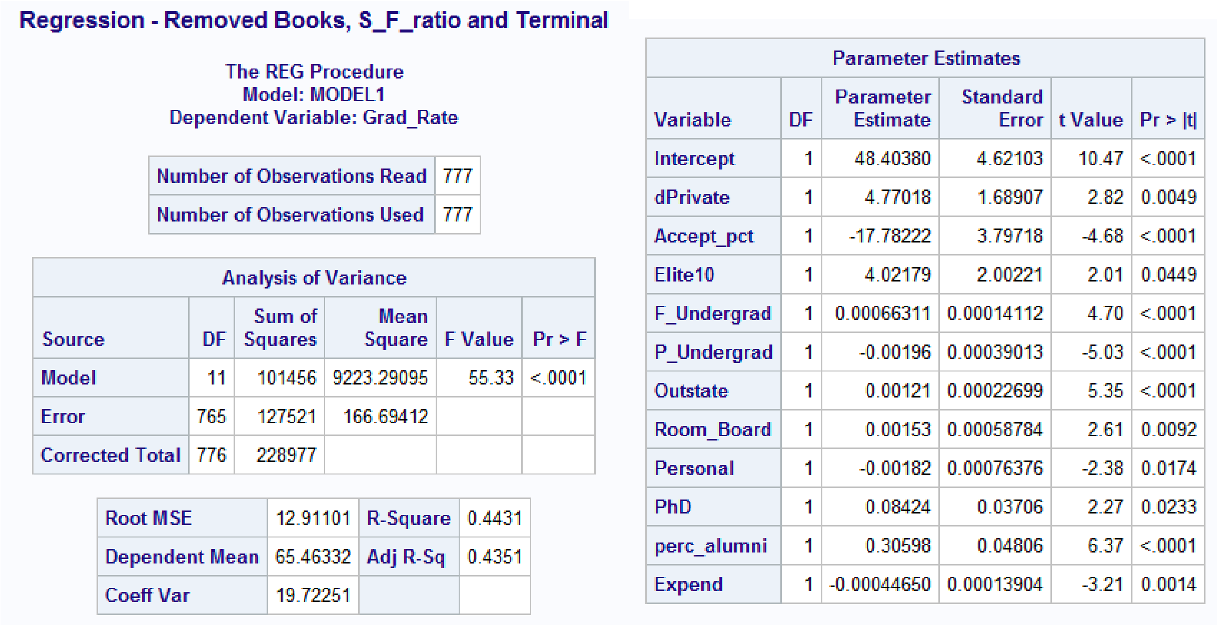
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Outstate	1	0.3264	0.3264	151.555	375.49	<.0001
2	perc_alumni	2	0.0412	0.3676	96.9415	50.49	<.0001
3	Accept_pct	3	0.0240	0.3916	65.9862	30.51	<.0001
4	P_Undergrad	4	0.0119	0.4036	51.5970	15.46	<.0001
5	F_Undergrad	5	0.0128	0.4164	35.9670	16.97	<.0001
6	Room_Board	6	0.0066	0.4230	28.8798	8.84	0.0030
7	Expend	7	0.0056	0.4287	23.1731	7.56	0.0061
8	Personal	8	0.0040	0.4326	19.6889	5.41	0.0203
9	dPrivate	9	0.0033	0.4360	17.1062	4.54	0.0334
10	PhD	10	0.0042	0.4401	13.3988	5.69	0.0173
11	Elite10	11	0.0029	0.4431	11.3674	4.03	0.0449
12	Terminal	12	0.0012	0.4443	11.7240	1.65	0.1999
13	Books	13	0.0005	0.4448	13.0000	0.72	0.3948

Adj-R2 Selection



Number in Model	Adjusted R-Square	R-Square	Variables in Model
12	0.4356	0.4443	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal perc_alumni Expend
13	0.4353	0.4448	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD Terminal perc_alumni Expend
12	0.4351	0.4438	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD perc_alumni Expend
11	0.4351	0.4431	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD perc_alumni Expend
13	0.4348	0.4443	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal S_F_Ratio perc_alumni Expend
14	0.4346	0.4448	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend
13	0.4343	0.4438	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD S_F_Ratio perc_alumni Expend
12	0.4343	0.4431	dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate Room_Board Personal PhD S_F_Ratio perc_alumni Expend
11	0.4333	0.4414	dPrivate Accept_pct F_Undergrad P_Undergrad Outstate Room_Board Personal PhD Terminal perc_alumni Expend
12	0.4331	0.4419	dPrivate Accept_pct F_Undergrad P_Undergrad Outstate Room_Board Books Personal PhD Terminal perc_alumni Expend
10	0.4328	0.4401	dPrivate Accept_pct F_Undergrad P_Undergrad Outstate Room_Board Personal PhD perc_alumni Expend

g) Fit a final regression model **M1** for Grad.Rate based on the results in f) – i.e. optimal model. Explain your choice. Write down the expression of the estimated model **M1**.



Regression - Removed Books, S_F_ratio and Terminal

The REG Procedure
Model: MODEL1
Dependent Variable: Grad_Rate

Number of Observations Read	777
Number of Observations Used	777

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	101456	9223.29095	55.33	<.0001
Error	765	127521	166.69412		
Corrected Total	776	228977			

Root MSE	12.91101	R-Square	0.4431
Dependent Mean	65.46332	Adj R-Sq	0.4351
Coeff Var	19.72251		

Parameter Estimates

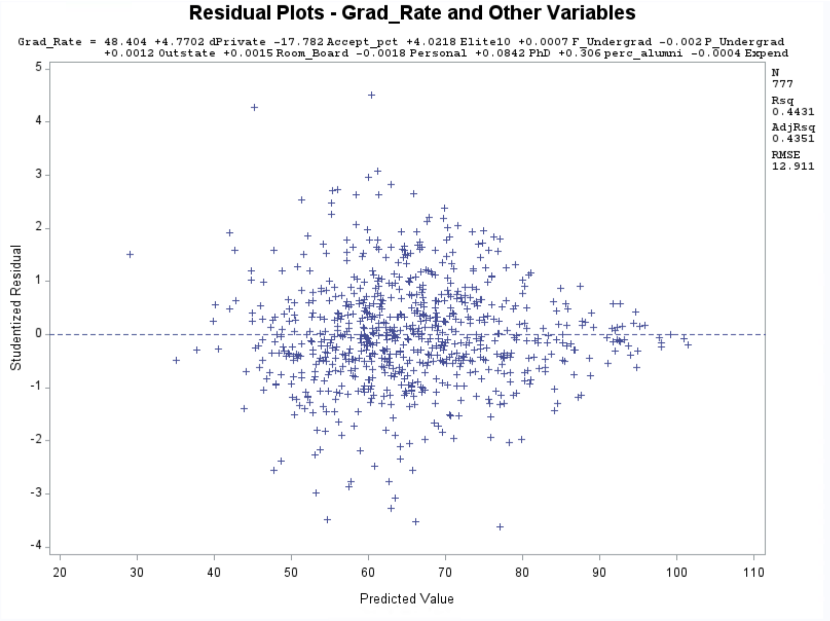
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	48.40380	4.62103	10.47	<.0001
dPrivate	1	4.77018	1.68907	2.82	0.0049
Accept_pct	1	-17.78222	3.79718	-4.68	<.0001
Elite10	1	4.02179	2.00221	2.01	0.0449
F_Undergrad	1	0.00066311	0.00014112	4.70	<.0001
P_Undergrad	1	-0.00196	0.00039013	-5.03	<.0001
Outstate	1	0.00121	0.00022699	5.35	<.0001
Room_Board	1	0.00153	0.00058784	2.61	0.0092
Personal	1	-0.00182	0.00076376	-2.38	0.0174
PhD	1	0.08424	0.03706	2.27	0.0233
perc_alumni	1	0.30598	0.04806	6.37	<.0001
Expend	1	-0.00044650	0.00013904	-3.21	0.0014

Forward selection starts without any x-variables, and then selects the good choice by each step, and ends when there is no further improvement. Adj-R2 selection is easier once we choose the best model with highest adj-R2.

We choose adj-R2 selection here to exclude estimated book costs. We should choose the highest adj-R2 and fewer variables. In that case, Terminal is removed because its p-value for t-test is 0.24 which greater than 0.05. It does not have significant effect on Grad_Rate.

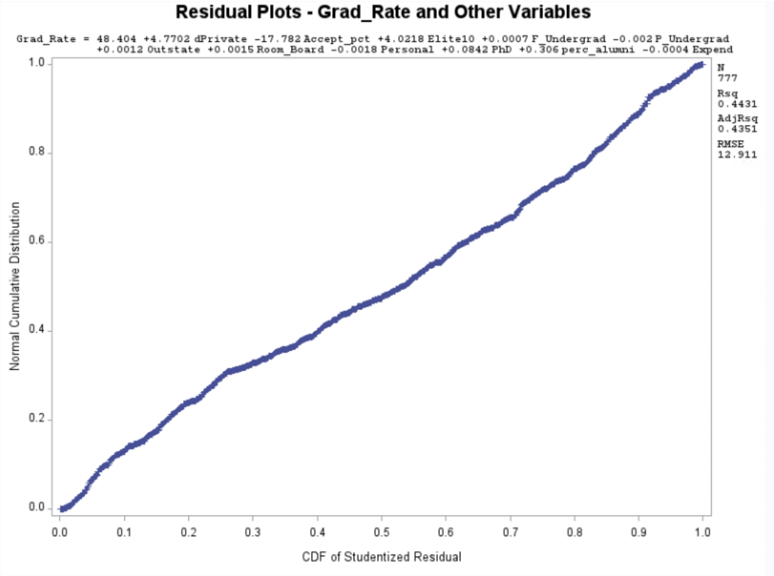
Predicted Grad_Rate = 48.40380 + 4.77018 * dPrivate – 17.78222 * Accept_pct + 4.02179 * Elite10 + 0.08424 * PhD + 0.30598 * Perc_alumni + e
(with dPrivate = 1 and Private = ‘Yes’)

h) Draw a plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis? Include the outputs and explain.



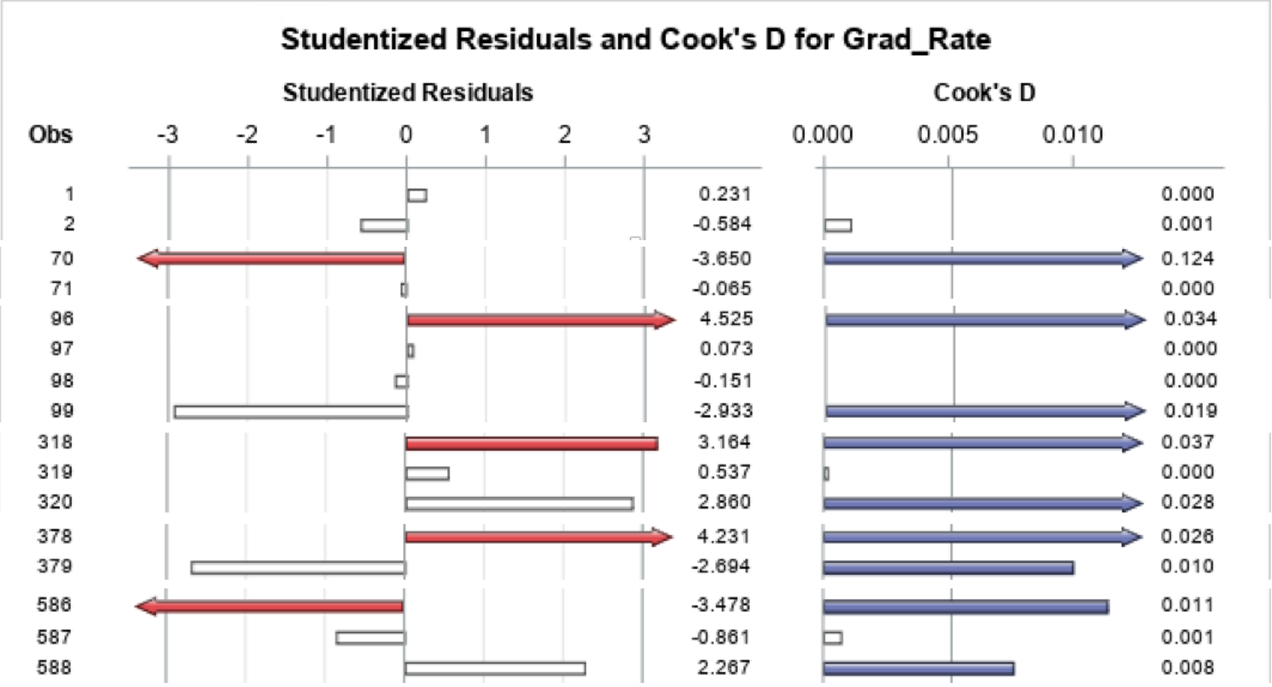
The spread of predicted values is not randomly scattered around zero line. In that case it shows no constant variance. Its kite shape violates independence assumption. There are also potential outliers greater 3 or less than -3 by y-axis.

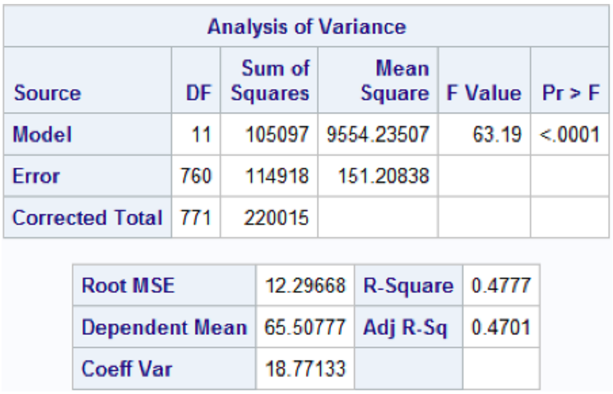
i) Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? Include the outputs and explain.



The graph shows a line which is almost 45-degree. That means it is normal distributed and linearly.

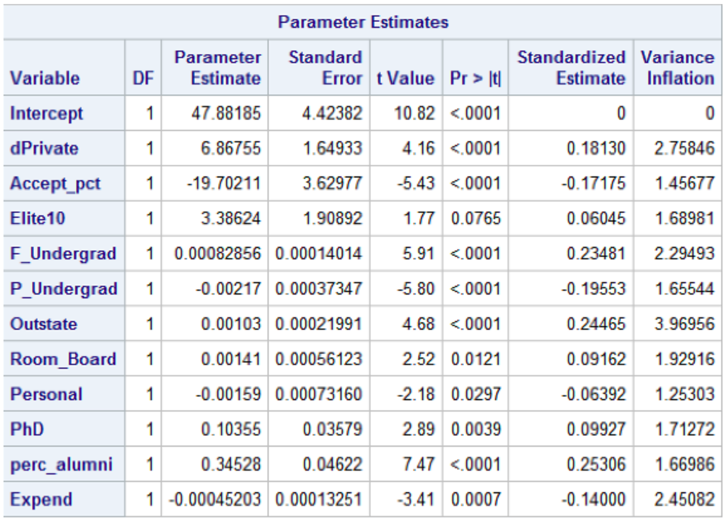
j) Are there any outliers or Influential Points? Compute appropriate statistics. Include the outputs. Take any action you think is necessary and explain why/why not you took these actions?





Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	105097	9554.23507	63.19	<.0001
Error	760	114918	151.20838		
Corrected Total	771	220015			

Root MSE	12.29668	R-Square	0.4777
Dependent Mean	65.50777	Adj R-Sq	0.4701
Coeff Var	18.77133		



Parameter Estimates

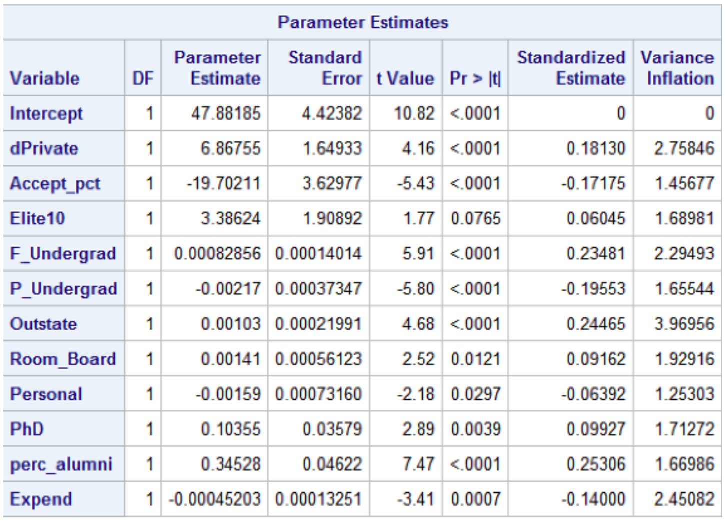
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	47.88185	4.42382	10.82	<.0001	0	0
dPrivate	1	6.86755	1.64933	4.16	<.0001	0.18130	2.75846
Accept_pct	1	-19.70211	3.62977	-5.43	<.0001	-0.17175	1.45677
Elite10	1	3.38624	1.90892	1.77	0.0765	0.06045	1.68981
F_Undergrad	1	0.00082856	0.00014014	5.91	<.0001	0.23481	2.29493
P_Undergrad	1	-0.00217	0.00037347	-5.80	<.0001	-0.19553	1.65544
Outstate	1	0.00103	0.00021991	4.68	<.0001	0.24465	3.96956
Room_Board	1	0.00141	0.00056123	2.52	0.0121	0.09162	1.92916
Personal	1	-0.00159	0.00073160	-2.18	0.0297	-0.06392	1.25303
PhD	1	0.10355	0.03579	2.89	0.0039	0.09927	1.71272
perc_alumni	1	0.34528	0.04622	7.47	<.0001	0.25306	1.66986
Expend	1	-0.00045203	0.00013251	-3.41	0.0007	-0.14000	2.45082

There are outliers marked in red as obs 40, 96, 318, 378 and 586 etc. and influential points marked in blue as 70, 96, 99, 318, 320 and 378 etc. Obs 96 is marked as both outlier and influential point. We will remove it. Then rerun, check and remove until there is no further improvement by removing observations. Overall, we remove 5 obs and see the changes of R2 and adj-R2.

k) Analyze the AdjR^2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.

R2 is 47.77% and adj-R2 is 47.01% during this time. We should check out adj-R2 because it does not increase with the addition of any x-variable that does not improve the regression model. However, adj-R2 with 47.01% is a relatively low number representing that this is not an ideally model.

l) Draw conclusions on graduation rates based on your regression analysis. What are the most important predictors in your model? Does your model show a significant difference in graduation rates between private and public universities? Do “elite” universities have higher graduation rates? Explain.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	47.88185	4.42382	10.82	<.0001	0	0
dPrivate	1	6.86755	1.64933	4.16	<.0001	0.18130	2.75846
Accept_pct	1	-19.70211	3.62977	-5.43	<.0001	-0.17175	1.45677
Elite10	1	3.38624	1.90892	1.77	0.0765	0.06045	1.68981
F_Undergrad	1	0.00082856	0.00014014	5.91	<.0001	0.23481	2.29493
P_Undergrad	1	-0.00217	0.00037347	-5.80	<.0001	-0.19553	1.65544
Outstate	1	0.00103	0.00021991	4.68	<.0001	0.24465	3.96956
Room_Board	1	0.00141	0.00056123	2.52	0.0121	0.09162	1.92916
Personal	1	-0.00159	0.00073160	-2.18	0.0297	-0.06392	1.25303
PhD	1	0.10355	0.03579	2.89	0.0039	0.09927	1.71272
perc_alumni	1	0.34528	0.04622	7.47	<.0001	0.25306	1.66986
Expend	1	-0.00045203	0.00013251	-3.41	0.0007	-0.14000	2.45082

M1:

Predicted Grad_Rate = 48.40380 + 4.77018 * dPrivate – 17.78222 * Accept_pct + 4.02179 * Elite10 + 0.08424 * PhD + 0.30598 * Perc_alumni + e
(with dPrivate = 1 and Private = ‘Yes’)

The most significant predictors are F_Undergrad, Outstate and perc_alumni by standardized estimate.

The difference of graduation rates between public and private universities are not severe. Assuming all the other variables constant, private universities will increase graduation rate by 6.87%

compared to public universities. Assuming all the other variables constant, elite universities will increase graduation rate by 3.39% compared to non-elite universities.

m) Copy and paste your FULL SAS code into the word document along with your answers.

```
TITLE "Analysis - College";

PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment5\College.csv"
out=grad replace;
delimiter=',';
getnames=yes;
RUN;

PROC PRINT;
RUN;


/*2A*/
TITLE "Histogram - Distribution of Graduation Rate";
PROC UNIVARIATE normal;
var Grad_Rate;
histogram / normal (mu=est sigma=est);
inset median mean min max range Q1 Q3 kurtosis skewness /pos = ne;
RUN;


/*2B*/
DATA grad_new;
set grad;
drop school Private;
dPrivate=(Private="Yes");
RUN;

PROC PRINT;
RUN;

PROC REG corr;
title "Correlation - All Variables";
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend;
RUN;

TITLE "GPLOTS - Y and X-variables";
PROC GPLOT data=grad_new;
plot Grad_Rate*(dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend);
RUN;


/*2C*/
TITLE "Boxplots - Graduation Rates and University Type";

PROC SORT;
by dPrivate;
RUN;
```

```
PROC BOXPLOT;
plot Grad_Rate*dPrivate;
RUN;

TITLE "Boxplots - Graduation Rates and Status";
PROC SORT;
by Elite10;
RUN;

PROC BOXPLOT;
plot Grad_Rate*Elite10;
RUN;


/*2D*/
PROC REG;
TITLE "Regression - All Variables";
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend;
RUN;

PROC REG corr;
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend;
RUN;


/*2E*/
PROC REG;
TITLE "Regression - All Variables";
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend /vif;
RUN;


/*2F*/
TITLE "Selection 1: Forward Selection";
PROC REG;
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend
/selection=forward;
RUN;

TITLE "Selection 2: Adj-R2 Selection";
PROC REG;
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Books Personal PhD Terminal S_F_Ratio perc_alumni Expend
/selection=adjrsq;
RUN;


/*2G*/
PROC REG;
TITLE "Regression - Removed Books, S_F_ratio and Terminal";
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Personal PhD perc_alumni Expend;
RUN;


/*2HI*/
```

```
PROC REGg corr;
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Personal PhD perc_alumni Expend;
RUN;

PROC REG;
TITLE "Residual Plots - Grad_Rate and Other Variables";
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Personal PhD perc_alumni Expend;
plot student.*(dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Personal PhD perc_alumni Expend predicted.);
plot npp.*student.;
RUN;


/*2JKL*/
TITLE "Final Model";
PROC REG;
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Personal PhD perc_alumni Expend /vif r influence stb;
RUN;

DATA grad_new1;
set grad_new1;
if _n_ in (112) then delete;
RUN;

TITLE "Final Model - Removed Obs";
PROC REG;
model Grad_Rate = dPrivate Accept_pct Elite10 F_Undergrad P_Undergrad Outstate
Room_Board Personal PhD perc_alumni Expend /vif r influence stb;
RUN;
```

Problem 3 [10 pts] – ONLY for GRADUATES

Select the BEST answer for the following:

1) Which of the following methods do we use to find the regression line for data in Linear Regression?

a) Minimize error

b) Maximize adj-R2

c) Maximize parameter estimates

d) Maximize standardized estimates

e) a, b, c and d

f) a, b, and c only

g) a, b only

2) Which of the following is/are not a selection method. Select all that applies.

a) Cp

b) Adj-R2

c) RMSE

d) Stepwise

e) PRESS Statistic

3) Using Cp criteria, specify which model(s) will be selected

Model #	**Predictors**				**Cp Value**
1	**X1**				**3.2**
2	**X1**	**X2**			**4.6**
3	**X1**	**X2**	**X3**		**2.1**
4	**X1**	**X2**	**X3**	**X4**	**5.1**

a) Model # 1 (k = 1, p = 2) false
b) Model # 2 (k = 2, p = 3) false
c) Model # 3 (k = 3, p = 4) false
d) Model # 4 (k = 4, p = 5) false
e) Model # 1 and 2 No
f) Model # 1, 2 and 4 No
g) Model # 3 and 4 No
h) None

4) Which of the following model(s) will be selected when using AIC as a selection criteria?

Model #	**AIC Value**
1	**1041.56**
2	**3456.40**
3	**9592.35**
4	**2467.23**

a) Model # 1
b) Model # 2
c) Model # 3
d) Model # 4
e) Model # 1 and 4
f) Model # 1, 2 and 3
g) None

5) A best model will have the following characteristics

a) Higher adj-R2
b) Higher RMSE
c) Most number of predictors
d) Predictors that make the most business sense
e) a, b, c, and d
f) a, b, c only
g) a, b, d only
h) a, c, d only
i) a, d only
j) b, c only
k) b, d only
l) none

6) Which of the following is true about residuals?

a) Lower is better
b) Higher is better
c) a or b depend on the situation
d) None of these

7) We have K independent variables (X1, X2… Xk) and dependent variable is Y. While performing regression analysis we found the correlation coefficient for X1 and Y is -0.9287. Which of the following is true for X1?

a) Association between the X1 and Y is weak

b) Association between the X1 and Y is strong

c) Association between the X1 and Y is weak and positive

d) Association between the X1 and Y is strong and positive

e) Association between the X1 and Y is weak and negative

f) Association between the X1 and Y is strong and negative

g) Association between the X1 and Y is neutral

h) Association between the X1 and Y there is no correlation

i) Correlation can’t judge the relationship

8) Based on the two statements specified below select the most appropriate answer.

Which of the following option is the correct for Pearson correlation between X1 and X2?
Statement-1. If X1 increases then X2 also decreases
Statement-2. If X1 increases then X2 behavior is unknown

a) Pearson correlation will be close to 1

b) Pearson correlation will be close to -1

c) Pearson correlation will be close to 0

d) None of these

9) Chance of overfitting is higher if you have a big dataset with nearly a million records.

a) True

b) False

10) Why shouldn’t irrelevant predictors be included in regression models?

a) The R2 will become too high

b) There are data limitations

c) It is bad research practice not to base your variables on sound theory

d) You increase the risk of making false significant results