**DSC423: Data Analysis And Regression / DSC 324: Data Analysis & Statistical Software II**

**Assignment-5** | **Total Points: 25 pts for DSC 423/25 pts for DSC 324**

**Due Date: 05/07/2019 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.

1. 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.

|  |
| --- |
|  |

The REG Procedure

Model: MODEL1

Dependent Variable: Balance

|  |  |
| --- | --- |
| **Number of Observations Read** | 102 |
| **Number of Observations Used** | 102 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 3 | 7155379617 | 2385126539 | 481.48 | <.0001 |
| **Error** | 98 | 485464527 | 4953720 |  |  |
| **Corrected Total** | 101 | 7640844145 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2225.69532 | **R-Square** | 0.9365 |
| **Dependent Mean** | 24888 | **Adj R-Sq** | 0.9345 |
| **Coeff Var** | 8.94289 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Standardized Estimate** |
| **Intercept** | **1** | -17732 | 3801.66282 | -4.66 | <.0001 | 0 |
| **Age** | **1** | 367.82141 | 64.59824 | 5.69 | <.0001 | 0.16436 |
| **Education** | **1** | 1300.30871 | 249.97314 | 5.20 | <.0001 | 0.15027 |
| **Wealth** | **1** | 0.11647 | 0.00468 | 24.89 | <.0001 | 0.80124 |

We take this fitted regression model from assignment 4:

Balance = -17732+367.82 age+1300.31 education +.116 wealth

The sample in the new census is as follow for median: 34 years, 13 years education, wealth 140,000. Home value and income are also given but we took these variables out of the model last week so we do not need to account for this.

Balance = -17732+367.82(34)+1300.31(13)+.116(140000) =

Balance = -17732+12505.88+16904.03+16240

Balance = 27917.91 AVG Bank balance for new census

|  |
| --- |
| confidence interval |

The UNIVARIATE Procedure

Variable: Balance

|  |  |  |  |
| --- | --- | --- | --- |
| **Basic Confidence Limits Assuming Normality** | | | |
| **Parameter** | **Estimate** | **95% Confidence Limits** | |
| **Mean** | 24888 | 23179 | 26596 |
| **Std Deviation** | 8698 | 7646 | 10088 |
| **Variance** | 75651922 | 58461065 | 101764664 |

We are dealing with a value of 27917.91 for the census data. We can see that based on the CI of 95%, 27917 us about $1400 above the mean of 26596 and well above the estimate of 24888. This data shows that the new census sample is above the mean of the bankingfull sample that we have been dealing with.

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

**proc** **import** datafile = "S:\Assignment5\bankingfull.txt" out= banking replace;

deimiter= '09'x;

getames = yes;

dataow= **2**;

**run**;

title "print bankingFull dataset";

**proc** **print**;

**run**;

title "scatterplot";

**PROC** **GPLOT**;

PLOT Balance\*(Age Education Income HomeVal Wealth);

**RUN**;

title "Correlation";

**Proc** **CORR**;

var Balance Age Education Income HomeVal Wealth;

**run**;

title "regression model";

**PROC** **reg**;

model balance= age Education Income HomeVal Wealth;

**run**;

title "regression model take away homeVal";

**PROC** **reg**;

model balance= age Education Income Wealth /vif tol;

**run**;

**PROC** **reg**;

model balance= age Education Wealth / stb;

**run**;

title "confidence interval";

ods select BasicIntervals;

**proc** **univariate** data = banking cibasic;

var balance;

**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.

1. 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)

Based on the histogram, an argument can be made for the dataset to be symmetrical. The median sits at 65 and that’s exactly where the mu for this histogram is. The inner quartile is from 53 to 78% for graduation rate with a max of 118%. This 118% for graduation rate is odd because the max percentage for graduation should only be at 100%.

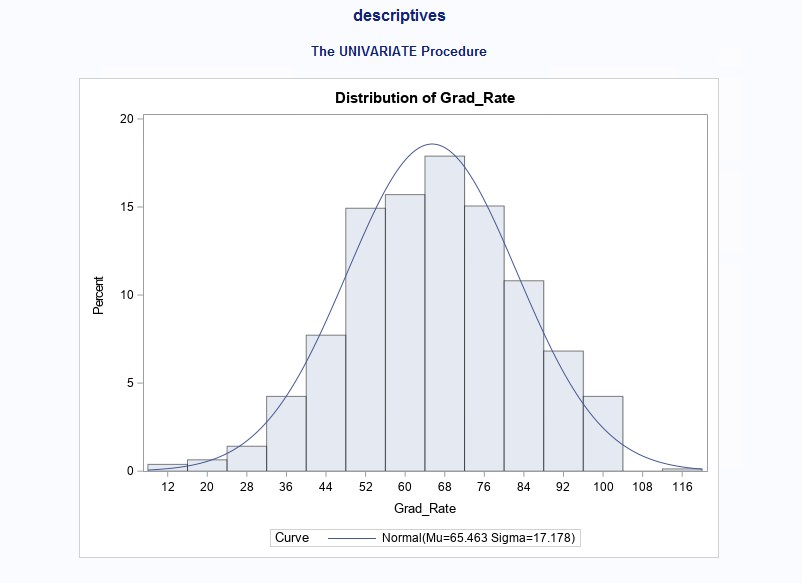
The MEANS Procedure

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Analysis Variable : Grad\_Rate** | | | | | | | | | |
| **Mean** | **Std Dev** | **Std Error** | **Lower 95% CL for Mean** | **Upper 95% CL for Mean** | **Minimum** | **25th Pctl** | **50th Pctl** | **75th Pctl** | **Maximum** |
| 65.4633205 | 17.1777099 | 0.6162469 | 64.2536119 | 66.6730290 | 10.0000000 | 53.0000000 | 65.0000000 | 78.0000000 | 118.0000000 |

The UNIVARIATE Procedure

Fitted Normal Distribution for Grad\_Rate

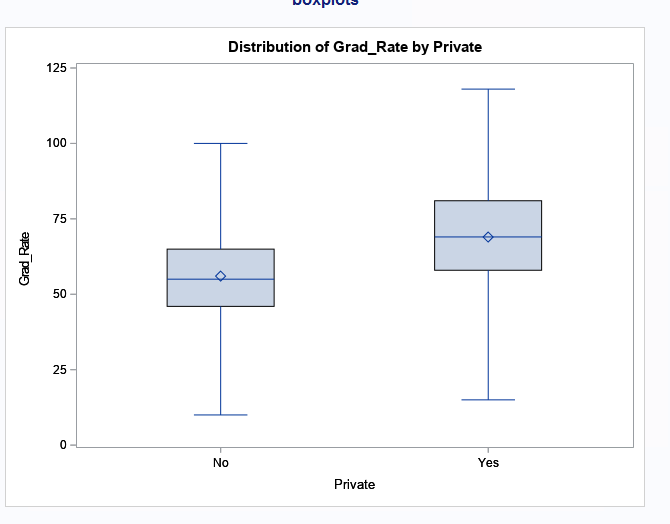
|  |  |  |
| --- | --- | --- |
| **Parameters for Normal Distribution** | | |
| **Parameter** | **Symbol** | **Estimate** |
| **Mean** | Mu | 65.46332 |
| **Std Dev** | Sigma | 17.17771 |

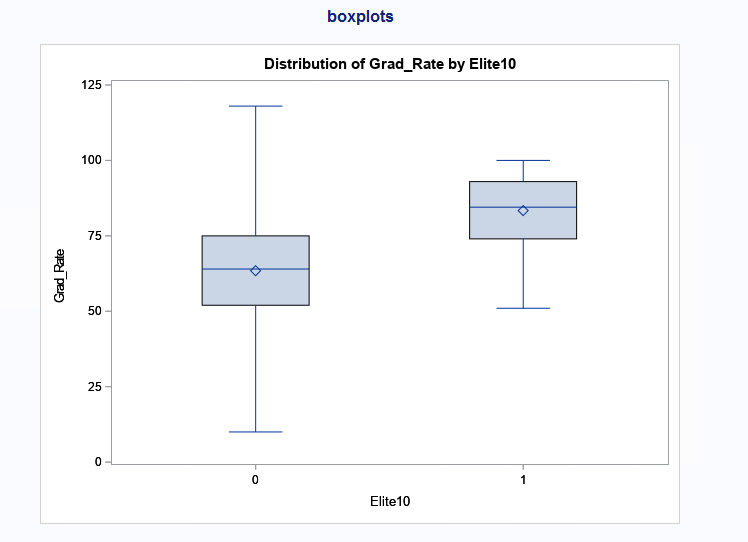


1. Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission).

Looking through the scatterplots, we can assume which independent variables may have a linear association with grad rate. Outstate, Room.board, perc.alumni are the only variables that seem to show any sort of trend at all. This makes sense, usually the alumni that donate are the ones who graduated from that school. Students from out of state are also the same students that usually room and board.

1. 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).





It seems that the graduation rate from private schools was higher than in public schools and the graduation rate for elite 10 schools was also higher. This makes sense. Elite 10 schools are much more difficult to get into and students in these schools may be more educationally inclined, which is what potentially got them into this school in the first place.

1. 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.

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** | 13 | 100650 | 7742.33001 | 46.03 | <.0001 |
| **Error** | 763 | 128327 | 168.18731 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 12.96870 | **R-Square** | 0.4396 |
| **Dependent Mean** | 65.46332 | **Adj R-Sq** | 0.4300 |
| **Coeff Var** | 19.81064 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 54.34320 | 6.04926 | 8.98 | <.0001 |
| **Accept\_pct** | **1** | -16.87896 | 3.83115 | -4.41 | <.0001 |
| **Elite10** | **1** | 3.77156 | 2.00928 | 1.88 | 0.0609 |
| **F\_Undergrad** | **1** | 0.00053711 | 0.00013295 | 4.04 | <.0001 |
| **P\_Undergrad** | **1** | -0.00202 | 0.00039119 | -5.17 | <.0001 |
| **Outstate** | **1** | 0.00145 | 0.00021445 | 6.78 | <.0001 |
| **Room\_Board** | **1** | 0.00185 | 0.00059295 | 3.12 | 0.0019 |
| **Books** | **1** | -0.00207 | 0.00297 | -0.70 | 0.4858 |
| **Personal** | **1** | -0.00174 | 0.00078122 | -2.23 | 0.0264 |
| **PhD** | **1** | 0.11700 | 0.05620 | 2.08 | 0.0377 |
| **Terminal** | **1** | -0.09216 | 0.06240 | -1.48 | 0.1401 |
| **S\_F\_Ratio** | **1** | -0.06436 | 0.16068 | -0.40 | 0.6889 |
| **perc\_alumni** | **1** | 0.32644 | 0.04815 | 6.78 | <.0001 |
| **Expend** | **1** | -0.00044976 | 0.00015234 | -2.95 | 0.0032 |

This is the first test we use, so we would have to remove variables one by one based on if they are significant or not and then re run the tests. S.f.ratio is the first x variable we can remove. The r2 value is pretty low as well here on the first run.

After the next few tests, books, terminal and phD are removed because they all have a value well above .05.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | 52.80789 | 4.19840 | 12.58 | <.0001 |
| **Accept\_pct** | **1** | -16.92629 | 3.77110 | -4.49 | <.0001 |
| **Elite10** | **1** | 3.99191 | 2.00587 | 1.99 | 0.0469 |
| **F\_Undergrad** | **1** | 0.00056402 | 0.00012251 | 4.60 | <.0001 |
| **P\_Undergrad** | **1** | -0.00200 | 0.00039072 | -5.13 | <.0001 |
| **Outstate** | **1** | 0.00150 | 0.00020944 | 7.18 | <.0001 |
| **Room\_Board** | **1** | 0.00177 | 0.00058651 | 3.01 | 0.0027 |
| **Personal** | **1** | -0.00180 | 0.00076782 | -2.34 | 0.0193 |
| **perc\_alumni** | **1** | 0.33231 | 0.04759 | 6.98 | <.0001 |
| **Expend** | **1** | -0.00041557 | 0.00013914 | -2.99 | 0.0029 |

This is what’s left that we can use in our final model – for now. The adjusted r2 value is at .4289 and that’s not a very good value to be at, suggesting that this may not be a great model to use.

1. 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.

Based on what we went over in class, any values above .9 would be said to have an issue with collinearity. Upon running the code and checking the Pearson correlation coefficient table, no value seems to be above this .9 value. No variables are even close to this, suggesting there are no issues with multicollinearity.

1. 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-R2, Cp, stepwise. Make sure to include the o/p of the 2 selection methods. No need to discuss the models, include the outputs.

|  |
| --- |
| 1. The SAS System |

The REG Procedure

Model: MODEL1

Dependent Variable: Grad\_Rate

|  |  |
| --- | --- |
| **Number of Observations Read** | 777 |
| **Number of Observations Used** | 777 |

|  |
| --- |
|  |

|  |
| --- |
| Forward Selection: Step 1 |

|  |
| --- |
| Variable Outstate Entered: R-Square = 0.3264 and C(p) = 116.9640 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 74732 | 74732 | 375.49 | <.0001 |
| **Error** | 775 | 154245 | 199.02634 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 39.99511 | 1.40840 | 160500 | 806.42 | <.0001 |
| **Outstate** | 0.00244 | 0.00012588 | 74732 | 375.49 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1, 1 |

|  |
| --- |
| Forward Selection: Step 2 |

|  |
| --- |
| Variable perc\_alumni Entered: R-Square = 0.3676 and C(p) = 64.4687 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 84177 | 42088 | 224.97 | <.0001 |
| **Error** | 774 | 144800 | 187.08073 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 38.44671 | 1.38276 | 144629 | 773.08 | <.0001 |
| **Outstate** | 0.00184 | 0.00014808 | 28998 | 155.00 | <.0001 |
| **perc\_alumni** | 0.34158 | 0.04807 | 9444.93193 | 50.49 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.472, 5.888 |

|  |
| --- |
| Forward Selection: Step 3 |

|  |
| --- |
| Variable Accept\_pct Entered: R-Square = 0.3916 and C(p) = 34.7465 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 3 | 89675 | 29892 | 165.87 | <.0001 |
| **Error** | 773 | 139302 | 180.21021 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 54.08112 | 3.13907 | 53490 | 296.82 | <.0001 |
| **Accept\_pct** | -18.64409 | 3.37543 | 5497.99332 | 30.51 | <.0001 |
| **Outstate** | 0.00168 | 0.00014843 | 22998 | 127.62 | <.0001 |
| **perc\_alumni** | 0.34299 | 0.04718 | 9522.95870 | 52.84 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5355, 12.208 |

|  |
| --- |
| Forward Selection: Step 4 |

|  |
| --- |
| Variable P\_Undergrad Entered: R-Square = 0.4036 and C(p) = 20.9704 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 92409 | 23102 | 130.59 | <.0001 |
| **Error** | 772 | 136568 | 176.90186 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 58.50805 | 3.30769 | 55350 | 312.88 | <.0001 |
| **Accept\_pct** | -20.84219 | 3.39071 | 6684.00364 | 37.78 | <.0001 |
| **P\_Undergrad** | -0.00131 | 0.00033369 | 2734.25255 | 15.46 | <.0001 |
| **Outstate** | 0.00159 | 0.00014879 | 20154 | 113.93 | <.0001 |
| **perc\_alumni** | 0.31060 | 0.04747 | 7573.86804 | 42.81 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5717, 21.251 |

|  |
| --- |
| Forward Selection: Step 5 |

|  |
| --- |
| Variable F\_Undergrad Entered: R-Square = 0.4164 and C(p) = 6.0000 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 95350 | 19070 | 110.03 | <.0001 |
| **Error** | 771 | 133627 | 173.31646 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 54.94348 | 3.38641 | 45624 | 263.24 | <.0001 |
| **Accept\_pct** | -18.62280 | 3.39914 | 5202.24424 | 30.02 | <.0001 |
| **F\_Undergrad** | 0.00049754 | 0.00012078 | 2941.24711 | 16.97 | <.0001 |
| **P\_Undergrad** | -0.00214 | 0.00038646 | 5306.87039 | 30.62 | <.0001 |
| **Outstate** | 0.00164 | 0.00014781 | 21340 | 123.13 | <.0001 |
| **perc\_alumni** | 0.32067 | 0.04705 | 8051.32378 | 46.45 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5832, 36.555 |

|  |
| --- |
| All variables have been entered into the model. |

|  |
| --- |
|  |
|  |
|  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | 116.964 | 375.49 | <.0001 |
| **2** | perc\_alumni | 2 | 0.0412 | 0.3676 | 64.4687 | 50.49 | <.0001 |
| **3** | Accept\_pct | 3 | 0.0240 | 0.3916 | 34.7465 | 30.51 | <.0001 |
| **4** | P\_Undergrad | 4 | 0.0119 | 0.4036 | 20.9704 | 15.46 | <.0001 |
| **5** | F\_Undergrad | 5 | 0.0128 | 0.4164 | 6.0000 | 16.97 | <.0001 |

|  |
| --- |
| The SAS System |

The REG Procedure

Model: MODEL1

Dependent Variable: Grad\_Rate

|  |
| --- |
| The SAS System |

The REG Procedure

Model: MODEL1

Dependent Variable: Grad\_Rate

|  |  |
| --- | --- |
| **Number of Observations Read** | 777 |
| **Number of Observations Used** | 777 |

|  |
| --- |
|  |

|  |
| --- |
| Backward Elimination: Step 0 |

|  |
| --- |
| All Variables Entered: R-Square = 0.4164 and C(p) = 6.0000 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 95350 | 19070 | 110.03 | <.0001 |
| **Error** | 771 | 133627 | 173.31646 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 54.94348 | 3.38641 | 45624 | 263.24 | <.0001 |
| **Accept\_pct** | -18.62280 | 3.39914 | 5202.24424 | 30.02 | <.0001 |
| **F\_Undergrad** | 0.00049754 | 0.00012078 | 2941.24711 | 16.97 | <.0001 |
| **P\_Undergrad** | -0.00214 | 0.00038646 | 5306.87039 | 30.62 | <.0001 |
| **Outstate** | 0.00164 | 0.00014781 | 21340 | 123.13 | <.0001 |
| **perc\_alumni** | 0.32067 | 0.04705 | 8051.32378 | 46.45 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5832, 36.555 |

|  |
| --- |
| All variables left in the model are significant at the 0.1000 level. |

|  |
| --- |
| The SAS System |

The REG Procedure

Model: MODEL1

Dependent Variable: Grad\_Rate

|  |
| --- |
| The SAS System |

The REG Procedure

Model: MODEL1

Dependent Variable: Grad\_Rate

|  |  |
| --- | --- |
| **Number of Observations Read** | 777 |
| **Number of Observations Used** | 777 |

|  |
| --- |
|  |

|  |
| --- |
| Forward Selection: Step 1 |

|  |
| --- |
| Variable Outstate Entered: R-Square = 0.3264 and C(p) = 116.9640 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 74732 | 74732 | 375.49 | <.0001 |
| **Error** | 775 | 154245 | 199.02634 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 39.99511 | 1.40840 | 160500 | 806.42 | <.0001 |
| **Outstate** | 0.00244 | 0.00012588 | 74732 | 375.49 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1, 1 |

|  |
| --- |
| Forward Selection: Step 2 |

|  |
| --- |
| Variable perc\_alumni Entered: R-Square = 0.3676 and C(p) = 64.4687 |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 84177 | 42088 | 224.97 | <.0001 |
| **Error** | 774 | 144800 | 187.08073 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 38.44671 | 1.38276 | 144629 | 773.08 | <.0001 |
| **Outstate** | 0.00184 | 0.00014808 | 28998 | 155.00 | <.0001 |
| **perc\_alumni** | 0.34158 | 0.04807 | 9444.93193 | 50.49 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.472, 5.888 |

|  |
| --- |
| Forward Selection: Step 3 |

|  |
| --- |
| Variable Accept\_pct Entered: R-Square = 0.3916 and C(p) = 34.7465 |
|  |
|  |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 3 | 89675 | 29892 | 165.87 | <.0001 |
| **Error** | 773 | 139302 | 180.21021 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 54.08112 | 3.13907 | 53490 | 296.82 | <.0001 |
| **Accept\_pct** | -18.64409 | 3.37543 | 5497.99332 | 30.51 | <.0001 |
| **Outstate** | 0.00168 | 0.00014843 | 22998 | 127.62 | <.0001 |
| **perc\_alumni** | 0.34299 | 0.04718 | 9522.95870 | 52.84 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5355, 12.208 |

|  |
| --- |
| Forward Selection: Step 4 |

|  |
| --- |
| Variable P\_Undergrad Entered: R-Square = 0.4036 and C(p) = 20.9704 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 4 | 92409 | 23102 | 130.59 | <.0001 |
| **Error** | 772 | 136568 | 176.90186 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 58.50805 | 3.30769 | 55350 | 312.88 | <.0001 |
| **Accept\_pct** | -20.84219 | 3.39071 | 6684.00364 | 37.78 | <.0001 |
| **P\_Undergrad** | -0.00131 | 0.00033369 | 2734.25255 | 15.46 | <.0001 |
| **Outstate** | 0.00159 | 0.00014879 | 20154 | 113.93 | <.0001 |
| **perc\_alumni** | 0.31060 | 0.04747 | 7573.86804 | 42.81 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5717, 21.251 |

|  |
| --- |
| Forward Selection: Step 5 |

|  |
| --- |
| Variable F\_Undergrad Entered: R-Square = 0.4164 and C(p) = 6.0000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 5 | 95350 | 19070 | 110.03 | <.0001 |
| **Error** | 771 | 133627 | 173.31646 |  |  |
| **Corrected Total** | 776 | 228977 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Parameter Estimate** | **Standard Error** | **Type II SS** | **F Value** | **Pr > F** |
| **Intercept** | 54.94348 | 3.38641 | 45624 | 263.24 | <.0001 |
| **Accept\_pct** | -18.62280 | 3.39914 | 5202.24424 | 30.02 | <.0001 |
| **F\_Undergrad** | 0.00049754 | 0.00012078 | 2941.24711 | 16.97 | <.0001 |
| **P\_Undergrad** | -0.00214 | 0.00038646 | 5306.87039 | 30.62 | <.0001 |
| **Outstate** | 0.00164 | 0.00014781 | 21340 | 123.13 | <.0001 |
| **perc\_alumni** | 0.32067 | 0.04705 | 8051.32378 | 46.45 | <.0001 |

|  |
| --- |
| Bounds on condition number: 1.5832, 36.555 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | 116.964 | 375.49 | <.0001 |
| **2** | perc\_alumni | 2 | 0.0412 | 0.3676 | 64.4687 | 50.49 | <.0001 |
| **3** | Accept\_pct | 3 | 0.0240 | 0.3916 | 34.7465 | 30.51 | <.0001 |
| **4** | P\_Undergrad | 4 | 0.0119 | 0.4036 | 20.9704 | 15.46 | <.0001 |
| **5** | F\_Undergrad | 5 | 0.0128 | 0.4164 | 6.0000 | 16.97 | <.0001 |

1. 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**.

I went with forward selection because this was where the highest r2 value was represented. The expression would include all the variables from above because they are all less than .05.

1. 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.

There are problems in the regression analysis based on this. The output is shown below.

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

There is evidence of this because it does not show a fully linear line. There is an s shape shown for this.

1. 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?
2. Analyze the AdjR2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.

The r2 value is pretty low compared to the other hw assignments we have had. It is in the high 40s and this suggest that this model may not be the best to use.

1. 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.

There were some surprising factors at play here, as I stated in the other problems above. Elite universities do tend to have higher graduation rates, along with accept\_pct. The students that are more inclined to study and di well are the ones that graduate more often.

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

**proc** **import** datafile = "S:\Assignment5\College.csv" out= college replace;

Delimiter = ',' MISSOVER FIRSTOBS = **2**;

getnames = yes;

**run**;

**proc** **print**;

**run**;

title "descriptives";

**proc** **means** mean std stderr clm min p25 p50 p75 max;

var grad\_rate;

**run**;

title "Histogram";

**proc** **univariate** normal;

var Grad\_rate;

histogram / normal (mu=est sigma =est);

**run**;

**proc** **gplot** ;

plot grad\_rate\*(School Private Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal

S\_F\_Ratio perc\_alumni Expend);

**run**;

title "boxplots";

**proc** **sort**;

by private;

**run**;

**proc** **boxplot**;

plot grad\_rate\*private;

**run**;

**proc** **sort**;

by elite10;

**run**;

**proc** **boxplot**;

plot grad\_rate\*elite10;

**run**;

**proc** **corr**;

var grad\_rate Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend ;

**run**;

**proc** **reg**;

model grad\_rate = Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Books Personal PhD Terminal S\_F\_Ratio perc\_alumni Expend;

**run**;

**proc** **print**;

**run**;

title "remove s.f.ratio, books, terminal, phd";

**proc** **reg**;

model grad\_rate = Accept\_pct Elite10 F\_Undergrad P\_Undergrad Outstate Room\_Board Personal perc\_alumni Expend;

**run**;

**proc** **reg**;

model grad\_rate = Accept\_pct F\_Undergrad P\_Undergrad

Outstate perc\_alumni/selection=forward;

**Run;**

**proc** **reg**;

model grad\_rate = Accept\_pct F\_Undergrad P\_Undergrad

Outstate perc\_alumni/selection=backward;

**run**;