APM 630 Regression Analysis
HW4 – Dummy Variable Regression
Name: Ruta Basijokaite

Project Report

1. Introduction

A study was conducted to investigate the trend in college enrollment for both male and female students from 1975 to 1993. The variables in the study included:

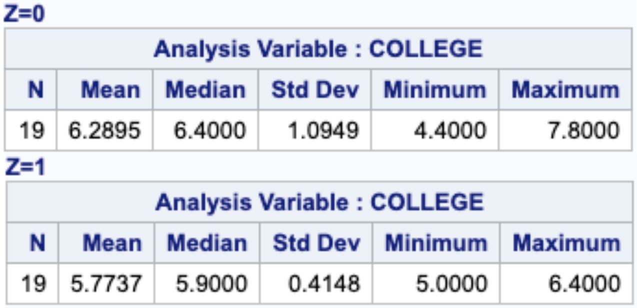
Y = college enrollment in millions (COLLEGE)
SEX = male of female
YEAR = calendar year
X = record year from 1 to 19

The purpose of the study is to (1) conduct a regression analysis to estimate the trend in college enrollment, (2) conduct a regression analysis to estimate the trend in college enrollment for male and female students, (3) perform statistical tests to examine if enrollment trend is different between male and female.

2. Methods

College admissions rates were recorded between years 1975 and 1993. Data obtained reflected separate college admission rates for male and female students. The descriptive statistics of college admission for male and female are listed in Table 1.

Table 1. Descriptive statistics of all variables (Z=0 represents female and Z=1 represents male). Answers contain four significant digits.



Z=0

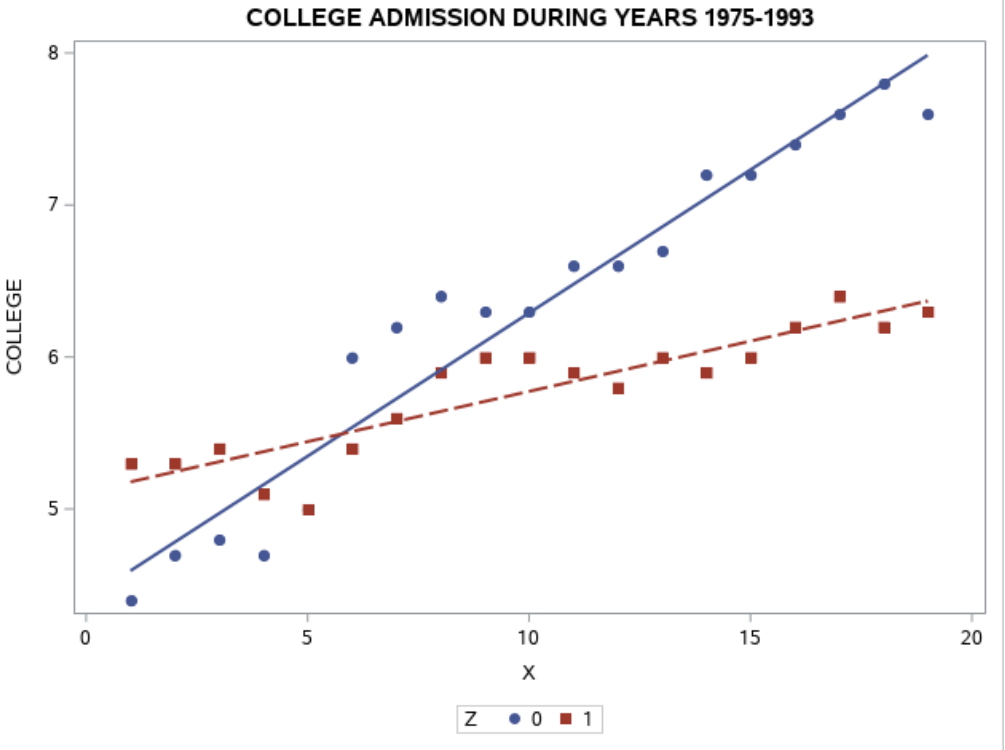
Analysis Variable : COLLEGE					
N	Mean	Median	Std Dev	Minimum	Maximum
19	6.2895	6.4000	1.0949	4.4000	7.8000

Z=1

Analysis Variable : COLLEGE					
N	Mean	Median	Std Dev	Minimum	Maximum
19	5.7737	5.9000	0.4148	5.0000	6.4000

Figure 1 illustrates the how college admission rates changed over the years for male and female students.

Figure 1. Matrix scatterplot of MPG and four predictive variables.



Often when dealing with categorical variables in the data, in this case male and female, dummy variable is used to represent these variables. Dummy variables indicate no meaningful measurement, but rather the categories of interest taken on by categorical variables. To represent categorical variable of sex, two categories were defined as follows:

Z=0 for female students
Z=1 for male students

First, overall model was developed using Statistical Analysis System (SAS) to combine the data for male and female to produce a single regression model that fits both categories. Overall model was based on the following linear regression formula:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad [1]$$

where Y is college enrollment in millions, β_0 - β_1 are regression coefficients to be estimated, and ε is the model error. Then, separate regression model was developed for male and female student enrollment that followed the same equation 1.

The single linear dummy variable regression model that was developed to incorporate the categories represented above can be defines as follows:

$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_2 XZ + \varepsilon$ [2]

where Y is college enrollment in millions, β_0 - β_2 are regression coefficients to be estimated, Z is a dummy variable and ε is the model error. Thus, for the first category (Z=0), the model is

$Y = \beta_0 + \beta_1 X + \varepsilon$ [3]

and for the second category (Z=1), the model is

$Y = \beta_0 + \beta_1 X + \beta_2 *1 + \beta_3 *1 + \varepsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon$ [4]

This allows us to incorporate the two separate regression models within a single model and allows different intercept and slope coefficients for male and female. When comparing regression line of the two levels of Z, it is important to compare their slopes, intercepts and whether they coincide. Statistical inference questions about these models can be answered after performing appropriate tests for parallelism, equal intercepts, and coincidence:

- Test of parallelism. The null hypothesis that the two regression lines are parallel is equivalent to test H_0: β_3=0. If H_0 cannot be rejected, equation 4 can be simplified:
  $Y = (\beta_0 + \beta_2) + \beta_1 X + \varepsilon$ [5]
  In that case, equation 3 and 5 would have the same slope, meaning that two lines are parallel to each other.
- Test of equal intercepts. The hypothesis that the two intercepts are equal, allowing for unequal slopes, is equivalent to test H_0: β_2=0 for the single regression model. If H_0 cannot be rejected, equation 4 can be simplified:
  $Y = \beta_0 + (\beta_1 + \beta_3) X + \varepsilon$ [6]
  In that case, equation 3 and 6 would have the same intercept, but different slopes.
- Test of coincidence. The hypothesis that the two regression lines are coincident is H_0: $\beta_2=\beta_3=0$. If H_0 cannot be rejected, the male model can be reduced to:
  $Y = \beta_0 + \beta_1 X + \varepsilon$ [7]
  In that case, two lines (equation 3 and 7) have the same coefficients i.e. two lines coincide.

3. Results and Discussion

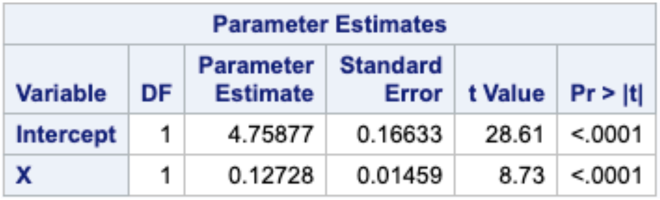
3.1 Regression Model Combining Male and Female Data

Equation 1 was edited to fit the male and female data using least-square method. Model can be represented using the following equation:

$\hat{Y} = 4.75877 - 0.12728*X$ [8]

Model R^2 value was equal to 0.6789 indicating that 67.89% of the total variation can be explained by this model. In addition, p-value analysis revealed that slope coefficient (β_1) was statistically significant at α=0.05 (Table 3). MSE value of the model was equal to 0.2426.

Table 3. Estimated regression coefficients for the combined male and female model



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.75877	0.16633	28.61	<.0001
X	1	0.12728	0.01459	8.73	<.0001

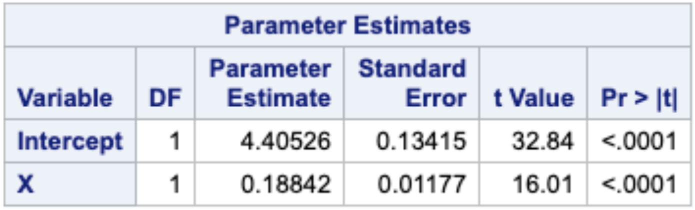
3.2 Separate Regression Models

Equation 1 was edited to fit the male and female data separately, creating two model. Model that fits the female (Z=0) enrollment data can be represented using the following equation:

$\hat{Y}$ = 4.40526 – 0.18842*X [9]

Female model R^2 value was equal to 0.9378. MSE was equal to 0.07891. Slope coefficient was statistically significant at α=0.05 (Table 4).

Table 4. Estimated regression coefficients for the female model



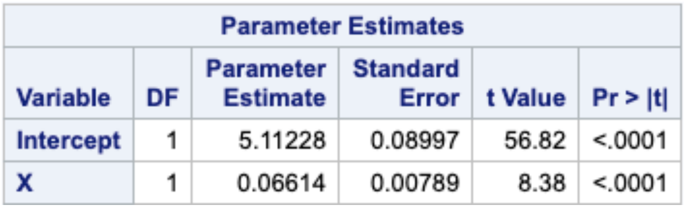
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.40526	0.13415	32.84	<.0001
X	1	0.18842	0.01177	16.01	<.0001

Model that fits the male (Z=1) enrollment data can be represented using the following equation:

$\hat{Y}$ = 5.11228 – 0.06614*X [10]

Male model R^2 value was equal to 0.8052. MSE was equal to 0.03549. Slope coefficient was statistically significant at α=0.05 (Table 5).

Table 5. Estimated regression coefficients for the male model



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.11228	0.08997	56.82	<.0001
X	1	0.06614	0.00789	8.38	<.0001

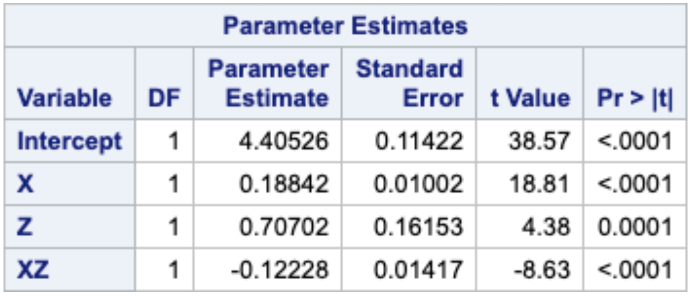
3.3 Regression Model with Dummy Variable Z

Equation 2 was edited to fit enrollment data in dummy variable model and can be represented using the following equation:

$\hat{Y}$ = 4.40526 + 0.18842*X + 0.70702*Z – 0.12228*XZ [11]

Dummy variable model R^2 value was equal to 0.9285, which is higher than the value that was produced by combined model and both separate models. MSE was equal to 0.05720. Slope coefficient was statistically significant at α=0.05 (Table 6).

Table 6. Estimated regression coefficients for the dummy variable model



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.40526	0.11422	38.57	<.0001
X	1	0.18842	0.01002	18.81	<.0001
Z	1	0.70702	0.16153	4.38	0.0001
XZ	1	-0.12228	0.01417	-8.63	<.0001

Equation 11 can be modified to manually derive linear regression model (equation 1) for male and female data following equations 3 and 4.

For female (Z=0):
$\hat{Y}_F$ = 4.40526 + 0.18842*X + 0.70702*0 – 0.12228*X*0 = 4.40526 + 0.18842*X [12]

For male (Z=1):
$\hat{Y}_M$ = 4.40526 + 0.18842*X + 0.70702*1 – 0.12228*X*1 = (4.40526 + 0.70702) + (0.18842 - 0.12228)*X = 5.11228 + 0.06614*X [13]

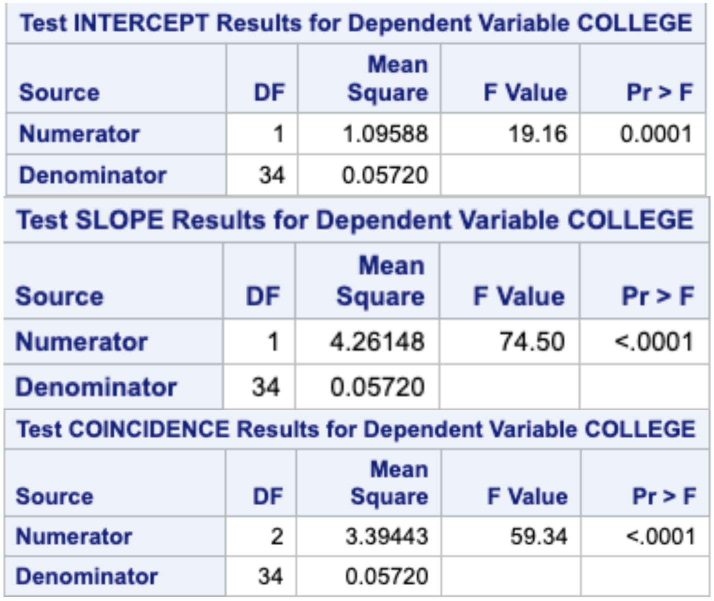
Results obtained from two straight line equation (12 and 13), derived from dummy variable model, are exactly the same as obtained by fitting separate regressions for male and female data as was obtained in equations 9 and 10.

Although at the beginning of the dataset (1975) male enrollment rate was 0.7 millions higher, over the 19 year period analyzed, female college enrollment was increasing 2.8 times faster than male enrollment.

3.4 Intercept, slopes and coincidence

Table 7 shows test results from analyzing the regression line of the two levels of Z. Intercept test revealed that H_0: β_2=0 can be rejected as p-value was lower than α=0.05 and therefore two lines do not have the same intercept. Same slope and coincidence null hypotheses were also rejected as p-value was lower than α=0.05, identifying that two lines did not have the same slope and were not coincident.

Table 7. Comparison of two lines



Test INTERCEPT Results for Dependent Variable COLLEGE				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	1.09588	19.16	0.0001
Denominator	34	0.05720		

Test SLOPE Results for Dependent Variable COLLEGE				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	4.26148	74.50	<.0001
Denominator	34	0.05720		

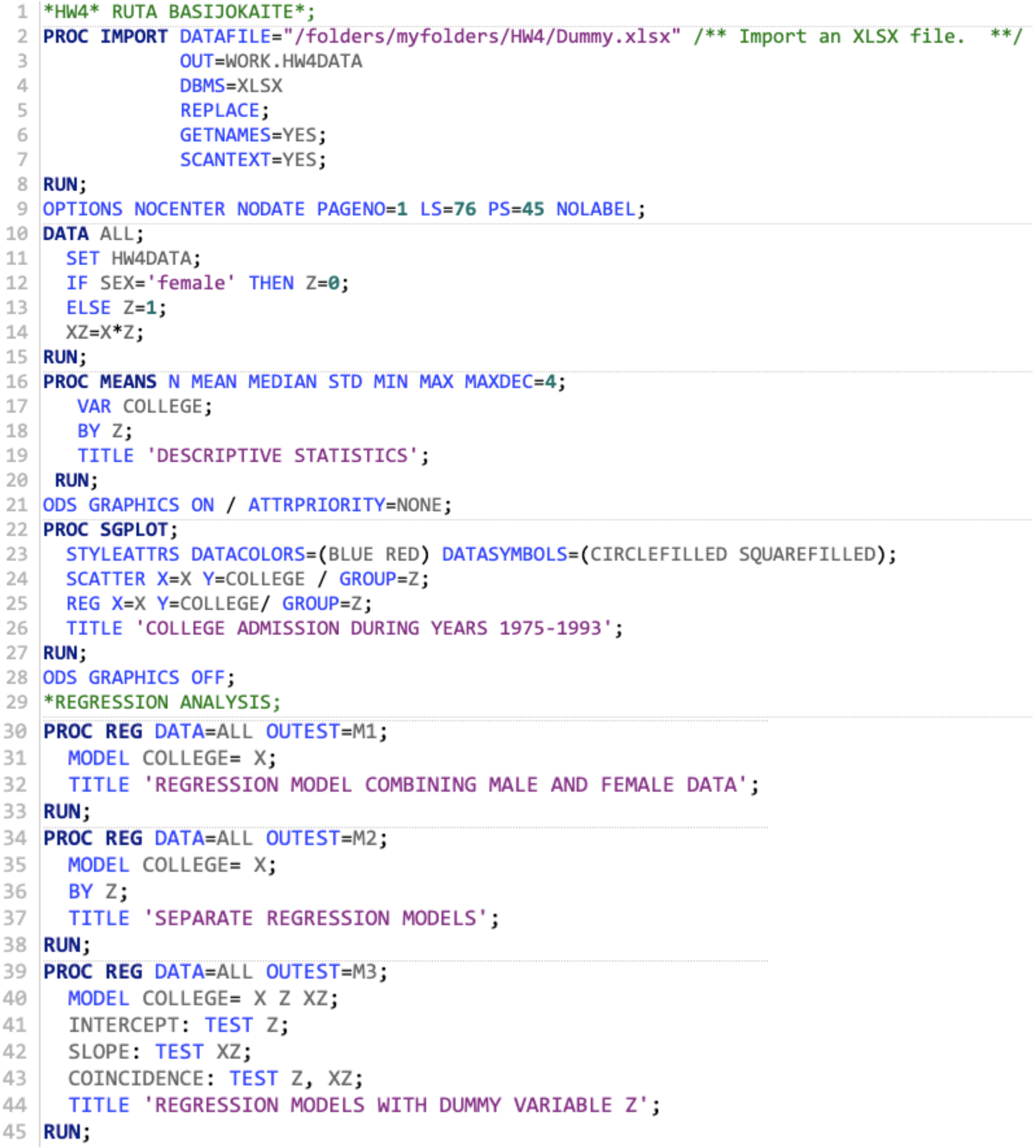
Test COINCIDENCE Results for Dependent Variable COLLEGE				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	3.39443	59.34	<.0001
Denominator	34	0.05720		

4. Summary

Trend in college enrollment for both male and female students was analyzed comparing different linear regression models. First, male and female enrollment data was combined and combined regression model was created. Then separate regression models were created for male and female enrollment. Also, dummy variable regression model was created and reduced to two regression lines for each category level, which compared to separate regression line models. Results revealed that dummy variable model produced the highest R^2 value of 0.9285. Two lines that were derived from dummy variable model did not have the same intercept, slope and were not coincident.

College enrollment trend analysis shows that during 1975 female students enrollment rate was lower than male. However, over the 19-year period analyzed, female enrollment rate was increasing 2.8 times faster than male enrollment. As college enrollment data was collected between 1975 and 1993, trends that were observed from this data can only be applied for that 19 year period. In order to evaluate more recent college enrollment trends, data after 1993 should be collected deriving enrollment trends over much longer period of time.

5. SAS programs



```
*HW4* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW4/Dummy.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW4DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
            SCANTEXT=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ALL;
  SET HW4DATA;
  IF SEX='female' THEN Z=0;
  ELSE Z=1;
  XZ=X*Z;
RUN;
PROC MEANS N MEAN MEDIAN STD MIN MAX MAXDEC=4;
   VAR COLLEGE;
   BY Z;
   TITLE 'DESCRIPTIVE STATISTICS';
 RUN;
ODS GRAPHICS ON / ATTRPRIORITY=NONE;
PROC SGPLOT;
  STYLEATTRS DATACOLORS=(BLUE RED) DATASYMBOLS=(CIRCLEFILLED SQUAREFILLED);
  SCATTER X=X Y=COLLEGE / GROUP=Z;
  REG X=X Y=COLLEGE/ GROUP=Z;
  TITLE 'COLLEGE ADMISSION DURING YEARS 1975-1993';
RUN;
ODS GRAPHICS OFF;
*REGRESSION ANALYSIS;
PROC REG DATA=ALL OUTEST=M1;
  MODEL COLLEGE= X;
  TITLE 'REGRESSION MODEL COMBINING MALE AND FEMALE DATA';
RUN;
PROC REG DATA=ALL OUTEST=M2;
  MODEL COLLEGE= X;
  BY Z;
  TITLE 'SEPARATE REGRESSION MODELS';
RUN;
PROC REG DATA=ALL OUTEST=M3;
  MODEL COLLEGE= X Z XZ;
  INTERCEPT: TEST Z;
  SLOPE: TEST XZ;
  COINCIDENCE: TEST Z, XZ;
  TITLE 'REGRESSION MODELS WITH DUMMY VARIABLE Z';
RUN;
```