Project Report

Name: Yifang Zhang

Email:yzhan377@syr.edu

1. Introduction

1.1 Background

A group of high-tech companies agreed to share employee salary information in an effort to establish salary ranges for technical positions in research and development for a study. The variables in the research included:

(1) One dependent variable:

Y = Current salary for each employee

(2) One independent variable:

X = years of experience since last degree for each employee

(3) One coded variable:

Z = The highest academic degree obtained which 1=B.S., 2=M.S., 3=Ph.D.

1.2 Purposes

The purposes of the study are:

(1) Compute descriptive statistics of Y and X for each level of Z;

(2) Find the scatterplot of Y and X for each level of Z;

(3) Test the differences between the three levels of Z on both Y and X;

(4) Fit a linear regression model with Y and X by each level of Z;

(5) Fit the dummy variable model and derive the model $Y=\beta_0+\beta_1*X$ by each level of Z from the dummy variable model and compare them.

(6) Test the regression line of the three levels of Z on intercepts, slopes and coincidence.

2. Data and Methods

2.1 Description of the variables

A group of high-tech companies offers employee salary information. Data obtained from every employee contented one dependent variable Y, one independent variable, and one coded variable Z, as defined above. The descriptive statistics of Y and X for each level of Z is shown in Table 1.

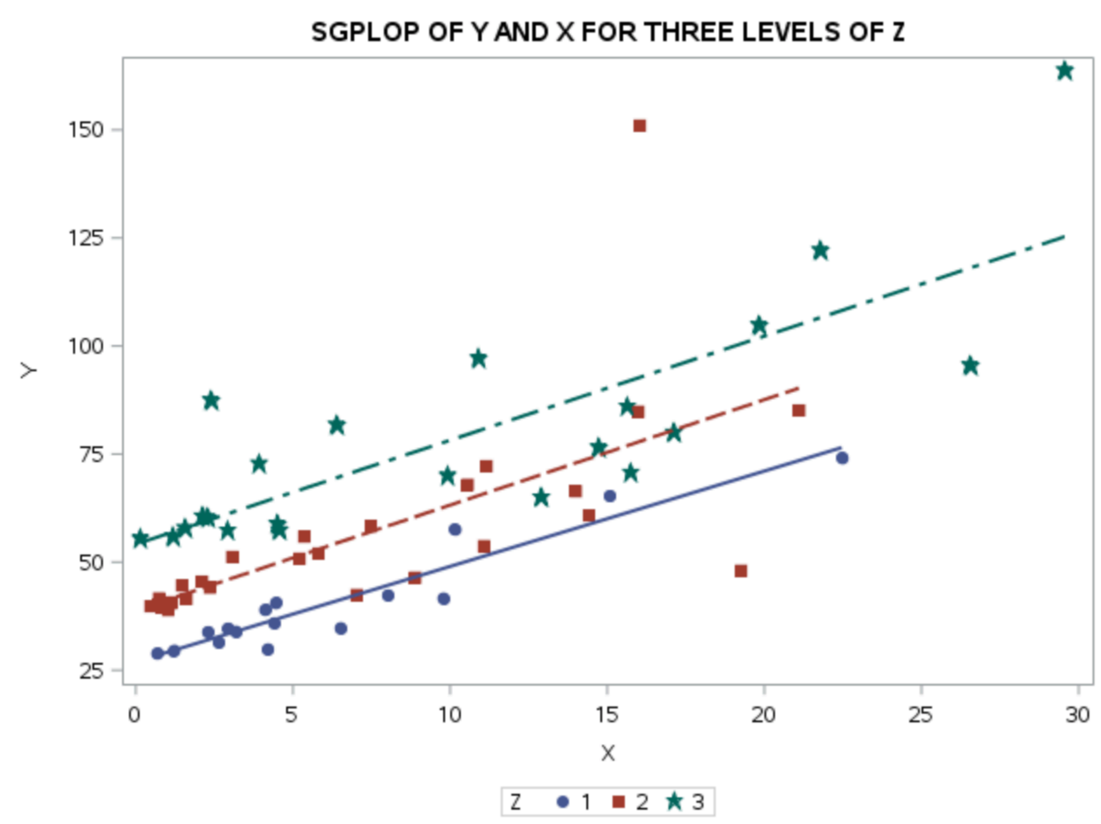
Among the total 65 employees, there are 16 employees have B.S. degree, 27 employees have M.S. degree and 22 employees have Ph.D. degree. Intuitively, the mean of Y and X in level Z=3 is the highest, while the mean of Y and X in level Z=1 is the lowest, which indicates that with a higher degree, the average of employee's' salary and years of experience is higher.

Table 1. Descriptive statistics of Y and X for each level of Z.

Z=1							
Variable	**N**	**Mean**	**Std Dev**	**Std Error**	**Coeff of Variation**	**Minimum**	**Maximum**
Y	16	40.981	13.33	3.332	32.527	29	74.2
X	16	6.363	5.733	1.433	90.111	0.65	22.46
Z=2							
Variable	**N**	**Mean**	**Std Dev**	**Std Error**	**Coeff of Variation**	**Minimum**	**Maximum**
Y	27	55.9	23.204	4.466	41.51	39.1	151.2
X	27	7.006	6.383	1.228	91.108	0.44	21.08
Z=3							
Variable	**N**	**Mean**	**Std Dev**	**Std Error**	**Coeff of Variation**	**Minimum**	**Maximum**
Y	22	78.918	26.19	5.584	33.187	55.5	163.7
X	22	10.289	8.787	1.873	85.408	0.14	29.54

Figure 1 draws the scatterplot of Y and X for the three levels of Z, in which blue and symbol circle was used to represent the scatterplot of Y and X for the level Z=1, red and symbol square was used to represent the scatterplot of Y and X for the level Z=2, and green and symbol star was used to represent the scatterplot of Y and X for the level Z=3. Note that in level Z=2 and Z=3, there are outliers.

Figure 1. The scatterplot of Y and X for the three levels of Z.



2.2. ANOVA Test on both Y and X

Table 2 exhibits the ANOVA test on the differences between the three levels of Z on variable Y. Since the p-value < 0.001, we can reject the H_0 for Y, which is $\mu_{b.s.} = \mu_{m.s.} = \mu_{phd}$, and this means the differences of average salary are statistically significant for the employees between B.S., M.S., and Ph.D. degrees.

Table 2. ANOVA test on the differences between the three levels of Z on Y.

Dependent Variable: Y					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	14115.00074	7057.50037	14.08	<.0001
Error	62	31069.0371	501.1135		
Corrected Total	64	45184.03785			

Table 3 exhibits the ANOVA test on the differences between the three levels of Z on variable X. Since the p-value = 0.1732 > 0.05, we do not reject the H_0 for X, which is $\mu_{b.s.} = \mu_{m.s.} = \mu_{phd}$, and this means the differences of average working experience are not statistically significant for the employees between B.S., M.S., and Ph.D. degrees.

Table 3. ANOVA test on the differences between the three levels of Z on X.

Dependent Variable: X					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	184.704964	92.352482	1.8	0.1732
Error	62	3173.905211	51.19202		
Corrected Total	64	3358.610175			

2.3 Methods

First, in this research, least-squares method was applied to fit the following two linear regression model by using Statistical Analysis System (SAS):

(1) Simple linear regression model by each level of Z (=1,2,3)

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

Where Y and X are defined as above in the background. β_0, β_1 are regression coefficients to be estimated, and ε is the model random error.

(2) The dummy variable model:

$$Y = \beta_0 + \beta_1*X + \beta_2*Z_1 + \beta_3*Z_2 + \beta_4*(XZ_1) + \beta_5*(XZ_2) + \varepsilon \quad [2]$$

Where Z1 and Z2 are two dummy variables defined as: if Z=1, then Z1=0, Z2=0; if Z=2, then Z1=1, Z2=0; if Z=3, then Z1=0, Z2=1. And XZ1=X*Z1, XZ2=X*Z2. β_0, β_1, β_2, β_3, β_4, β_5 are regression coefficients to be estimated, and ε is the model random error.

Second, we derive the model $Y = \beta_0 + \beta_1*X$ by each level of Z (=1,2,3) from the dummy variable model, that is equation [2].

Finally, we test the regression line of the three levels of Z if: (1) they have the same intercepts, (2) they have the same slopes, or (3) they are coincident by F-test.

3. Results and Discussion

3.1 Separate regression models by each level of Z (=1,2,3)

Equation [1] was fit the data using least-square method by each level of Z (=1,2,3) as follows:

(1) Z=1

$$Y = 26.944 + 2.206*X \qquad [1\text{-}1]$$

The model R^2 was 0.9005 and adjusted R^2 was 0.8934, indicating that 90.05% of the total variation in Y can be explained by the variable X by the level of Z=1. And both the intercept coefficient (β_0) and slope coefficient (β_1) are statistically significant at $\alpha = 0.05$ with p-value < 0.001. Table 4 gives the parameter estimates by the level of Z=1.

Table 4. Parameter Estimates by the level of Z=1.

Parameter Estimates for Z=1

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	26.94391	1.6552	16.28	<.0001	23.39386	30.49395
X	1	2.20626	0.19602	11.26	<.0001	1.78583	2.62669

(2) Z=2

$$Y = 38.813 + 2.439*X \qquad [1-2]$$

The model R^2 was 0.4501 and adjusted R^2 was 0.4281, indicating that only 45.01% of the total variation in Y can be explained by the variable X by the level of Z=2. And both the intercept coefficient (β_0) and slope coefficient (β_1) are statistically significant at $\alpha = 0.05$. Table 5 gives the parameter estimates by the level of Z=2.

Table 5. Parameter Estimates by the level of Z=2.

Parameter Estimates for Z=2

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	38.81349	5.06686	7.66	<.0001	28.3781	49.24888
X	1	2.43887	0.53916	4.52	0.0001	1.32844	3.54929

(3) Z=3

$$Y = 54.148 + 2.407*X \quad [1-3]$$

The model R^2 was 0.6525 and adjusted R^2 was 0.6351, indicating that 65.25% of the total variation in Y can be explained by the variable X by the level of Z=3. And both the intercept coefficient (β_0) and slope coefficient (β_1) are statistically significant at $\alpha = 0.05$. Table 6 gives the parameter estimates by the level of Z=3.

Table 6. Parameter Estimates by the level of Z=3.

Parameter Estimates for Z=3

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	54.14841	5.26482	10.28	<.0001	43.16619	65.13062
X	1	2.40749	0.39289	6.13	<.0001	1.58793	3.22705

3.2 The Dummy Variable Model

Equation [2] was fit the data using least-square method as follows:

$$Y=26.944+2.206*X+11.870*Z_1+27.205*Z_2+0.233*(XZ_1)+0.201*(XZ_2) \quad [2]$$

The model R^2 was 0.7130 and adjusted R^2 was 06886. The slope coefficients of X (β_1) and Z_2 (β_3) were statistically significant at $\alpha = 0.05$. Table 7 shows the parameter estimates of the dummy variable model.

Table 7. Parameter Estimates of the dummy variable model.

Parameter Estimates for the dummy variable model

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	26.94391	5.63802	4.78	<.0001	15.66225	38.22556
X	1	2.20626	0.66771	3.3	0.0016	0.87018	3.54235
Z1	1	11.86959	7.07918	1.68	0.0989	-2.29582	26.03499
Z2	1	27.2045	7.49199	3.63	0.0006	12.21306	42.19594
XZ1	1	0.2326	0.80831	0.29	0.7745	-1.38481	1.85002
XZ2	1	0.20123	0.7625	0.26	0.7928	-1.32453	1.72698

3.3 Model Comparison

Dependent on the dummy variable model, we can derive the model $Y=\beta_0+\beta_1*X$ and as follows:

- Z=1 Y= 26.944+2.206*X [3-1]
- Z=2 Y= (26.944+11.870) + (2.206+0.233) *X= 38.814 + 2.439*X [3-2]
- Z=3 Y= (26.944+27.205) + (2.206+0.201) *X= 54.149 + 2.407*X [3-3]

Then we can compare them with the three models obtained in 3.1. Table 8 shows the comparison. We can find that there are just slightly differences between separate regression models by each level of Z (=1,2,3) and the dummy variable models. Actually, we can say that they are the same models.

Table 8. Comparison between separate models and dummy variable models.

MODEL COMPARISON

	Separate Rgression Models	Dummy Variable Models
Z=1	Y = 26.944 + 2.206*X	Y = 26.944+2.206*X
Z=2	Y = 38.813 + 2.439*X	Y= 38.814 + 2.439*X
Z=3	Y = 54.148 + 2.407*X	Y= 54.149 + 2.407*X

3.4 Testing the Regression line of the three levels of Z

We can test the regression line of the three levels of Z as following:

(1) Slopes:

We have the H0: $\beta_4 = \beta_5 = 0$, comparing the following two models:

- $Y = \beta_0 + \beta_1 * X + \beta_2 * Z_1 + \beta_3 * Z_2 + \beta_4 * (XZ_1) + \beta_5 * (XZ_2)$ [a]
- $Y = \beta_0 + \beta_1 * X + \beta_2 * Z_1 + \beta_3 * Z_2$ [b]

Table 9 shows the test slopes results for dependent variable Y. The p-value = 0.9564 > 0.05, thus we cannot reject H0, which means that these two models do have same slopes.

Table 9. Test SLOPE Results for Dependent Variable Y.

Test SLOPE Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	9.8182	0.04	0.9564
Denominator	59	219.82651		

(2) Intercepts:

We have the H0: $\beta_2 = \beta_3 = 0$, comparing the following two models:

- $Y = \beta_0 + \beta_1{*}X + \beta_2{*}Z_1 + \beta_3{*}Z_2 + \beta_4{*}(XZ_1) + \beta_5{*}(XZ_2)$ [a]
- $Y = \beta_0 + \beta_1{*}X + \beta_4{*}(XZ_1) + \beta_5{*}(XZ_2)$ [c]

Table 10 shows the test intercepts results for dependent variable Y. The p-value = 0.0022 < 0.05, thus we can reject H0, which means that these two models don't have same intercepts.

Table 10. Test INTERCEPTS Results for Dependent Variable Y.

Test INTERCEPT Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1492.04321	6.79	0.0022
Denominator	59	219.82651		

(3) Coincident:

We have the H0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, comparing the following two models:

- $Y = \beta_0 + \beta_1{*}X + \beta_2{*}Z_1 + \beta_3{*}Z_2 + \beta_4{*}(XZ_1) + \beta_5{*}(XZ_2)$ [a]
- $Y = \beta_0 + \beta_1{*}X$ [d]

Table 11 shows the test coincidence results for dependent variable Y. The p-value < 0.0001 < 0.05, thus we can reject H0, which means that these two models are not coincident.

Table 11. Test CONINCIDENCE Results for Dependent Variable Y.

Test CONINCIDENCE Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	4	1840.59094	8.37	<.0001
Denominator	59	219.82651		

4. Summary

In this research, since the data has a coded variable Z (=1,2,3), which separately represented B.S., M.S., and Ph.D. degrees, we applied the dummy variable models to derive the model $Y = \beta_0 + \beta_1 * X$ by each level of Z (=1,2,3) and compare them with the separate regression model by each level of Z (=1,2,3). The result shows that the differences between them are slightly.

By conducting ANOVA, we find that differences of average salary are statistically significant for the employees between B.S., M.S., and Ph.D. degrees, while differences of average working experience are not statistically significant for the employees between B.S., M.S., and Ph.D. degrees.

Finally, by testing the regression line of the three level of Z, we find that they have the same slopes, but they don't have the same intercepts and they are not coincident.

5. SAS programs

5.1 Descriptive statistics and scatterplot

```
OPTIONS NODATE LS=76 PS=45 PAGENO=1 NOLABEL;
PROC IMPORT
DATAFILE='/home/yifangz02120/sasuser.v94/Salary2.xls'
OUT=Salary2
DBMS=XLS
REPLACE;
RUN;
DATA ALL;
SET Salary2;
RUN;
*===CALCULATE MEAN FOR DIFFERENT DEGREE===;
PROC SORT DATA=ALL;
BY Z;
PROC MEANS N MEAN STD STDERR CV MIN MAX MAXDEC=3;
VAR Y X;
BY Z;
RUN;
*===SGPLOP OF Y AND X FOR THREE LEVELS OF Z===;
TITLE 'SGPLOP OF Y AND X FOR THREE LEVELS OF Z';
PROC SGPLOT DATA=ALL;
STYLEATTRS DATACOLORS=(BLUE RED GREEN ) DATASYMBOLS=(Circlefilled
SquareFilled StarFilled);
SCATTER X=X Y=Y / GROUP=Z;
REG X=X Y=Y / GROUP=Z;
RUN;
```

5.2 ANOVA

```
*===ANOVA===;
PROC GLM DATA=ALL;
CLASS Z;
MODEL Y X = Z;
MEANS Z / LSD;
RUN;
```

5.3 Fit the Separate Regression Models

```
*=== SEPARATE REGRESSION MODELS===;
TITLE 'SEPARATE REGRESSION MODELS';
PROC REG DATA=ALL;
MODEL Y = X / CLB CLM;
BY Z;
RUN;
```

5.4 Create Two Dummy Variables, Fit the Dummy Variable Models and Testing.

```
*=== Create dummy variables===;
DATA ONE;
SET ALL;
IF Z^=2 THEN Z1=0;
ELSE IF Z=2 THEN Z1=1;
IF Z^=3 THEN Z2=0;
ELSE IF Z=3 THEN Z2=1;
RUN;
DATA DUMMY;
SET ONE;
XZ1 = X*Z1;
XZ2 = X*Z2;
```

```
RUN;
*===REGRESSION MODEL WITH DUMMY VARIABLE===;
TITLE 'Regression Model with Dummy Variable';
PROC REG DATA=DUMMY;
MODEL Y = X Z1 Z2 XZ1 XZ2 / CLB CLM;
SLOPE: TEST XZ1,XZ2;
INTERCEPT: TEST Z1,Z2;
CONINCIDENCE: TEST Z1,Z2,XZ1,XZ2;
RUN;
```