STAT 512: Homework 4

Name: Lei Nie

1. *Consider the following SAS output giving 5 confidence intervals for the mean of Y. If you wanted to guarantee that joint coverage of the five confidence intervals was at least 90%, what confidence level would you use when forming each interval, using the Bonferroni correction? Compute this adjusted confidence interval for the mean of Y when X = 4. (Note that some observations have been omitted from the output.)*

   In this case, to guarantee the joint coverage of the five confidence intervals was at least 90%, we need to calculate the individual confidence intervals with probability $1-\frac{\alpha}{5}=0.99$. From the SAS output, we get the degree freedom is 16 and Standard Error of Mean predict when $X=4$ is 1.0878. Thus

$$t-critic = t_{16}(0.995) = 2.921$$

The 99% Confidence Interval for $X = 4\ is\ 56.6309 \pm 2.921 \times 1.0878 = (53.4534, 59.9084)$.

2. *Based on the following small data set, construct the design matrix, X, its transpose X', and the matrices $X'X$, $(X'X)^{-1}$, $X'Y$, and $b=(X'X)^{-1}X'Y$.*

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \quad X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 6 & 15 \\ 15 & 55 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 0.5238 & -0.1429 \\ -0.1429 & 0.0571 \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 \\ 4 \\ 7 \\ 9 \\ 10 \\ 8 \end{bmatrix} \quad X'Y = \begin{bmatrix} 39 \\ 125 \end{bmatrix}$$

$$b = (X'X)^{-1}X'Y = \begin{bmatrix} 2.571429 \\ 1.571429 \end{bmatrix}$$

3. *Run the multiple linear regression with quality, experience, and publications as the explanatory variables and salary as the response variable. Summarize the regression results by giving the fitted regression equation, the value of R^2, and the results of the significance test for the null hypothesis that the three regression coefficients for the explanatory variables are all zero (give null and alternative hypotheses, test statistic with degrees of freedom, p-value, and brief conclusion in words).*

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	627.81700	209.27233	68.12	<.0001
Error	20	61.44300	3.07215		
Corrected Total	23	689.26000			

Root MSE	1.75276	**R-Square**	0.9109
Dependent Mean	39.50000	**Adj R-Sq**	0.8975
Coeff Var	4.43735		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	17.84693	2.00188	8.92	<.0001
X1	**1**	1.10313	0.32957	3.35	0.0032
X2	**1**	0.32152	0.03711	8.66	<.0001
X3	**1**	1.28894	0.29848	4.32	0.0003

Given that Y represents *salary*, X_1 represents *quality*, X_2 represents *experience* and X_3 represents *publications*, the fitted regression equation:

$$\hat{Y} = 17.8469 + 1.1031X_1 + 0.3215X_2 + 1.2889X_3$$

The R^2 of the fitted model is 91.09%.

The Hypothesis that the three regression coefficients for the explanatory variables are all zero is

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

H_a: At least one of the β_i's is not equal to zero.

Since

$$F - statistic = 68.12 \text{ with } df = 3 \text{ \& } 20 \text{ and } P - value < 0.0001,$$

We can conclude that we are 95% confident that at least one of the regression coefficients is not equal to zero.

4. *Give 85% confidence intervals (do not use a Bonferroni correction) for regression coefficients of quality, experience, and publications based on the multiple regression. Describe the results of the hypothesis tests for the individual regression coefficients (give null and alternative hypotheses, test statistic with degrees of freedom, p-value, and a brief conclusion in words). What is the relationship between these results and the confidence intervals?*

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**85% Confidence Limits**	
Intercept	**1**	17.84693	2.00188	8.92	<.0001	14.85005	20.84381
X1	**1**	1.10313	0.32957	3.35	0.0032	0.60975	1.59651
X2	**1**	0.32152	0.03711	8.66	<.0001	0.26597	0.37707
X3	**1**	1.28894	0.29848	4.32	0.0003	0.84211	1.73577

The 85% confidence interval for quality (X_1) is (0.60975, 1.59651);
The 85% confidence interval for experience (X_2) is (0.26597, 0.37707);
The 85% confidence interval for publications (X_3) is (0.84211, 1.73577).

Hypothesis test for the regression coefficient of quality (X_1):

$$H_0: \beta_1 = 0 \; vs. H_a: \beta_1 \neq 0$$
$$Test\ Statistic\ (t) = 3.35 \text{ with DF} = 1 \text{ and P} - \text{value} = 0.0032$$

Thus, with 85% confidence interval, we reject the null hypothesis and conclude that *quality* has a significant linear relationship with *salary*. The conclusion is consistent with the fact that its 85% confidence interval doesn't contain zero.

Hypothesis test for the regression coefficient of experience (X_2):

$$H_0: \beta_2 = 0 \; vs. H_a: \beta_2 \neq 0$$
$$Test\ Statistic\ (t) = 8.66 \text{ with DF} = 1 \text{ and P} - \text{value} < 0.0001$$

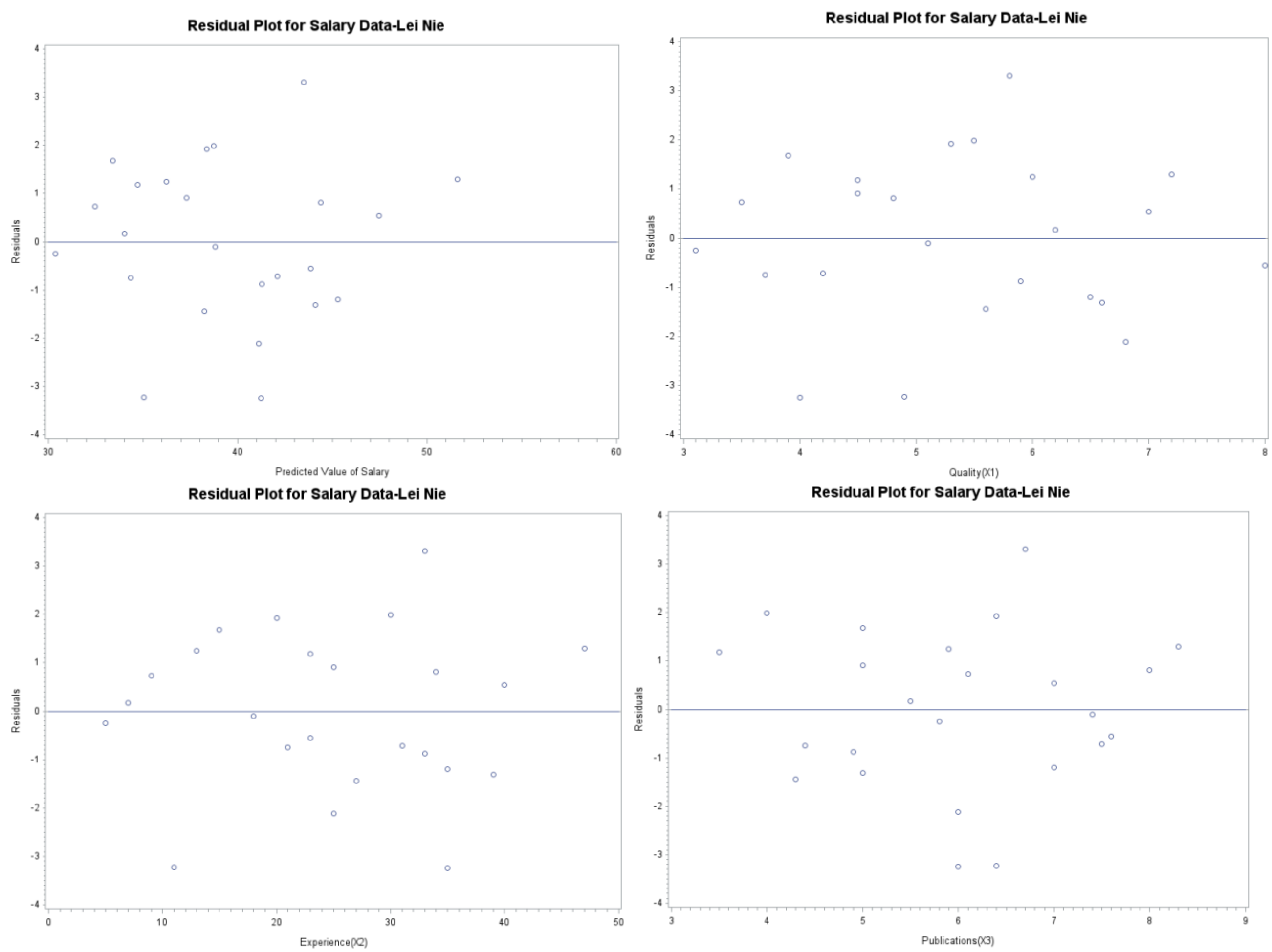
Thus, with 85% confidence interval, we reject the null hypothesis and conclude that *experience* has a significant linear relationship with *salary*. The conclusion is consistent with the fact that its 85% confidence interval doesn't contain zero.

Hypothesis test for the regression coefficient of experience (X_3):

$$H_0: \beta_3 = 0 \; vs. H_a: \beta_3 \neq 0$$
$$Test\ Statistic\ (t) = 4.32 \text{ with DF} = 1 \text{ and P} - \text{value} = 0.0003$$

Thus, with 85% confidence interval, we reject the null hypothesis and conclude that *publication* has a significant linear relationship with *salary*. The conclusion is consistent with the fact that its 85% confidence interval doesn't contain zero.

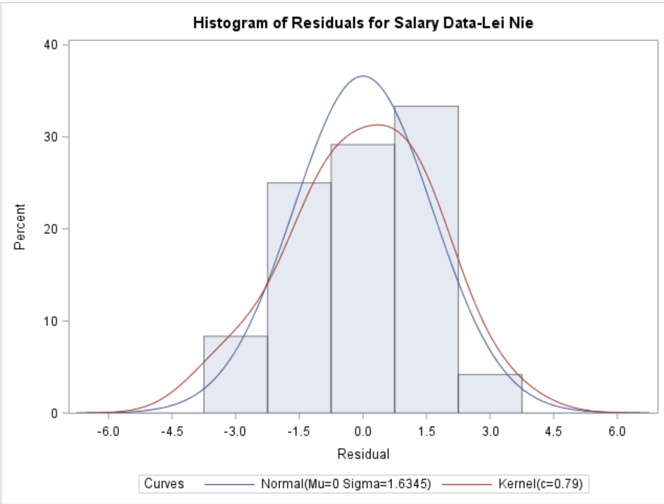
5. *Plot the residuals versus the predicted salary and each of the explanatory variables (i.e., 4 residual plots). Are there any unusual patterns?*

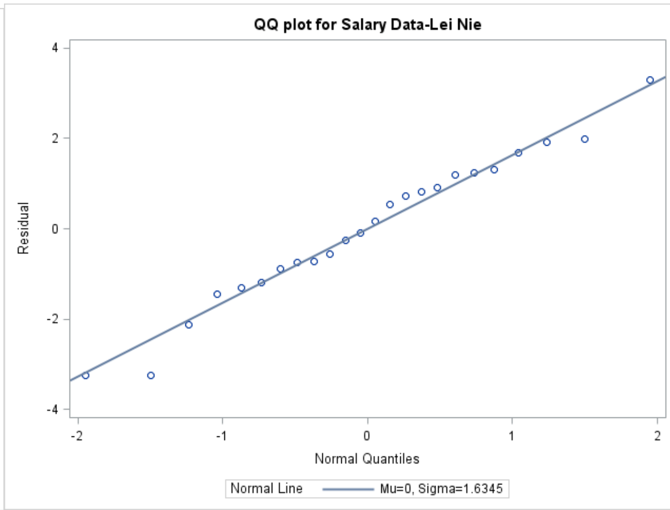


There are no obvious patterns on the residual plots. The linearity and constant variance assumptions are not violated.

6. *Examine the assumption of normality for the residuals using a qq-plot and histogram. State your conclusions.*

   There seems to be no violation to the normality assumption based on the histogram and the QQ-plot.





7. *Predict the salary for a mathematician with quality index equal to 5.2, 15 years of experience, and publication index equal to 6.5. Provide an 85% prediction interval with your prediction.*

   The predicted salary is $36.7841, and the 85% confidence interval is ($34.0382, $39.5300).

Appendix: SAS code

```
*Question 2;
data data2;
input X Y;
datalines;
0 1
;
proc print data=data2;run;
proc reg data=data2;model Y=X;run;
*Question 3;
data data3;
input Y X1 X2 X3;
datalines;
33.2    3.5    9.0    6.1
   ;
proc print data=data3;
proc reg data=data3;
model Y= X1 X2 X3;run;
*Question 4;
proc reg data=data3;
model Y= X1 X2 X3/ clb alpha=0.15;
output out=out5 r=res p=pred;
run;
proc print data=out5;run;

*Question 5;
title1 'Residual Plot for Salary
Data-Lei Nie';
symbol1 v=circle i=rl;
axis1 label=('Predicted Value of
Salary');
axis2 label=(angle=90 'Residuals');
axis3 label=('Quality(X1)');
axis4 label=('Experience(X2)');
axis5 label=('Publications(X3)');
proc gplot data=out5;
plot res*pred/ haxis=axis1
vaxis=axis2 vref=0;
plot res*X1/ haxis=axis3
vaxis=axis2 vref=0;
plot res*X2/ haxis=axis4
vaxis=axis2 vref=0;
plot res*X3/ haxis=axis5
vaxis=axis2 vref=0;
run;

*Question 6;
ods graphics on;
title1 'Histogram of Residuals for
Salary Data-Lei Nie';
proc univariate data=out5 plot
normal;
var res;
histogram res / normal  kernel(L=2)
odstitle=title1;
label res='Residual';
run;
title1 'QQ plot for Salary Data-Lei
Nie';
proc univariate data=out5 plot
normal;
var res;
qqplot res / normal (L=1 mu=est
sigma=est) odstitle=title1;
label res='Residual';
run;
ods graphics off;*Question 5;
title1 'Residual Plot for Salary
Data-Lei Nie';
symbol1 v=circle i=rl;
axis1 label=('Predicted Value of
Salary');
axis2 label=(angle=90 'Residuals');
axis3 label=('Quality(X1)');
axis4 label=('Experience(X2)');
axis5 label=('Publications(X3)');
proc gplot data=out5;
plot res*pred/ haxis=axis1
vaxis=axis2 vref=0;
plot res*X1/ haxis=axis3
vaxis=axis2 vref=0;
plot res*X2/ haxis=axis4
vaxis=axis2 vref=0;
plot res*X3/ haxis=axis5
vaxis=axis2 vref=0;
run;

*Question 6;
title1 'Plot for Normality
Diagnosis for Salary Data-Lei Nie';
proc univariate data=out5 plot
normal;
var res;
histogram res / normal  kernel(L=2);
qqplot res / normal (L=1 mu=est
sigma=est);run;

*Question 7;
data a1;
X1=5.2; X2=15; X3=6.5; output;
data data7;
set data3 a1;
proc reg data=data7;
model Y=X1 X2 X3/cli alpha=0.15;
id X1 X2 X3;run;
```