MATH 5531 Statistical Methods I HW 6

Russell Land

December 3, 2018

33.2	3.5	9.0	6.1
40.2	5.3	20.0	6.4
38.7	5.1	18.0	7.4
46.8	5.8	33.0	6.7
41.4	4.2	31.0	7.5
37.5	6.0	13.0	5.9
39.0	6.8	25.0	6.0
40.7	5.5	30.0	4.0
30.1	3.1	5.0	5.8
52.9	7.2	47.0	8.3
38.2	4.5	25.0	5.0
31.8	4.9	11.0	6.4
43.3	8.0	23.0	7.6
44.1	6.5	35.0	7.0
42.8	6.6	39.0	5.0
33.6	3.7	21.0	4.4
34.2	6.2	7.0	5.5
48.0	7.0	40.0	7.0
38.0	4.0	35.0	6.0
35.9	4.5	23.0	3.5
40.4	5.9	33.0	4.9
36.8	5.6	27.0	4.3
45.2	4.8	34.0	8.0
35.1	3.9	15.0	5.0

A researcher in a scientific foundation wished to evaluate the relation between intermediate and senior level annual salaries of bachelor's and master's level mathematicians (Y, in thousand dollars) and an index of work quality (X1), the number of years of experience (X2), and an index of publication (X3). The data for a sample of 24 bachelor's and master's level mathematicians are given above.Assume that the regression model for the three predictors variables with independent normal error terms is appropriate.

1) Obtain the least squares estimated regression equation.

2) Calculate the coefficient of determination and interpret the measure.

3) Test whether there is a significant regression relation. State the alternatives. decision rule and conclusion. What does your test imply about the regression coefficients? What is the p-value of the test?

4) Obtain a 95% confidence interval for each of the regression coefficients.

5) Obtain a plot of the residuals against each predictor variable.

6) Prepare a normal probability plot.

7) Analyze your plots and summarize your findings.

To answer questions 1-7 we will be using the SAS Studio Software, University Edition. In this document we will implement a LaTeX package named StatRep. This will allow us to display our SAS code in our TeX document neatly, and will also allow us to compile our SAS results straight into the LaTeX document. Once we present our data below, I will follow up with an analysis of the data. We will begin by creating a data set in the program, and then run the actually procedures for each test. Below shows the data set `Evaluations` being created:

```
title 'Mathematician Evaluations';

data Evaluation;

/*  Bachelor and Master Level Mathematician Data
Y represents annual salaries in thousands of dollars
X1 represents an idex in quality of work
X2 represents the number of years of experience
X3 represents an index of publication success */

input Y X1 X2 X3;
datalines;
33.2 3.5  9.0 6.1
40.2 5.3 20.0 6.4
38.7 5.1 18.0 7.4

... more data lines ...

36.8 5.6 27.0 4.3
45.2 4.8 34.0 8.0
35.1 3.9 15.0 5.0

run;
```

Now that our data set is stored as `Evaluation`, we can run the appropriate procedures, including `glm` and `reg`. We are going to use the `glm` procedure because it uses the method of least squares to fit general linear models. The `reg` procedure will help us get the data concerning the residuals. Below is the SAS code for our analyses:

```
proc print;

proc glm;
model Y=X1 X2 X3 / clparm ;

proc reg;
model Y=X1 X2 X3 / cli ;

run;
```

Figure 1: Statistics for Regression Analysis

Mathematician Evaluations

Obs	Y	X1	X2	X3
1	33.2	3.5	9	6.1
2	40.2	5.3	20	6.4
3	38.7	5.1	18	7.4
4	46.8	5.8	33	6.7
5	41.4	4.2	31	7.5
6	37.5	6.0	13	5.9
7	39.0	6.8	25	6.0
8	40.7	5.5	30	4.0
9	30.1	3.1	5	5.8
10	52.9	7.2	47	8.3
11	38.2	4.5	25	5.0
12	31.8	4.9	11	6.4
13	43.3	8.0	23	7.6
14	44.1	6.5	35	7.0
15	42.8	6.6	39	5.0
16	33.6	3.7	21	4.4
17	34.2	6.2	7	5.5
18	48.0	7.0	40	7.0
19	38.0	4.0	35	6.0
20	35.9	4.5	23	3.5
21	40.4	5.9	33	4.9
22	36.8	5.6	27	4.3
23	45.2	4.8	34	8.0
24	35.1	3.9	15	5.0

Mathematician Evaluations

The GLM Procedure

Number of Observations Read	24
Number of Observations Used	24

Mathematician Evaluations

The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	628.0425418	209.3475139	68.56	<.0001
Error	20	61.0670415	3.0533521		
Corrected Total	23	689.1095833			

R-Square	Coeff Var	Root MSE	Y Mean
0.911383	4.424225	1.747384	39.49583

Figure 1: *continued*

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	306.7653205	306.7653205	100.47	<.0001
X2	1	264.1167925	264.1167925	86.50	<.0001
X3	1	57.1604289	57.1604289	18.72	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	34.3992592	34.3992592	11.27	0.0031
X2	1	230.9566896	230.9566896	75.64	<.0001
X3	1	57.1604289	57.1604289	18.72	0.0003

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	17.84740401	1.99574171	8.94	<.0001	13.68435977	22.01044826
X1	1.10282189	0.32856347	3.36	0.0031	0.41745050	1.78819327
X2	0.32175047	0.03699494	8.70	<.0001	0.24458038	0.39892057
X3	1.28747999	0.29756455	4.33	0.0003	0.66677122	1.90818877

Mathematician Evaluations

The REG Procedure

Model: MODEL1

Dependent Variable: Y

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	628.04254	209.34751	68.56	<.0001
Error	20	61.06704	3.05335		
Corrected Total	23	689.10958			

Root MSE	1.74738	**R-Square**	0.9114
Dependent Mean	39.49583	**Adj R-Sq**	0.8981
Coeff Var	4.42422		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	17.84740	1.99574	8.94	<.0001
X1	1	1.10282	0.32856	3.36	0.0031
X2	1	0.32175	0.03699	8.70	<.0001
X3	1	1.28748	0.29756	4.33	0.0003

Figure 1: *continued*

Mathematician Evaluations

The REG Procedure

Model: MODEL1

Dependent Variable: Y

Output Statistics						
Obs	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**95% CL Predict**		**Residual**
1	33.2	32.4567	0.7491	28.4909	36.4224	0.7433
2	40.2	38.3672	0.4243	34.6164	42.1181	1.8328
3	38.7	38.7907	0.6346	34.9128	42.6685	−0.0907
4	46.8	43.4877	0.4639	39.7164	47.2589	3.3123
5	41.4	42.1096	0.8082	38.0936	46.1256	−0.7096
6	37.5	36.2432	0.6681	32.3409	40.1455	1.2568
7	39.0	41.1152	0.5917	37.2669	44.9635	−2.1152
8	40.7	38.7154	0.7396	34.7574	42.6734	1.9846
9	30.1	30.3423	0.8575	26.2821	34.4025	−0.2423
10	52.9	51.5961	0.9379	47.4592	55.7330	1.3039
11	38.2	37.2913	0.5040	33.4977	41.0848	0.9087
12	31.8	35.0304	0.6250	31.1592	38.9015	−3.2304
13	43.3	43.8551	0.9892	39.6665	48.0436	−0.5551
14	44.1	45.2894	0.5458	41.4707	49.1081	−1.1894
15	42.8	44.1117	0.7526	40.1430	48.0804	−1.3117
16	33.6	34.3495	0.6794	30.4387	38.2603	−0.7495
17	34.2	34.0183	0.9034	29.9150	38.1216	0.1817
18	48.0	47.4495	0.6687	43.5467	51.3523	0.5505
19	38.0	41.2448	0.7774	37.2554	45.2343	−3.2448
20	35.9	34.7165	0.7936	30.7133	38.7198	1.1835
21	40.4	41.2805	0.5990	37.4273	45.1337	−0.8805
22	36.8	38.2466	0.6440	34.3620	42.1313	−1.4466
23	45.2	44.3803	0.8283	40.3465	48.4141	0.8197
24	35.1	33.4121	0.5801	29.5715	37.2526	1.6879

Sum of Residuals	0
Sum of Squared Residuals	61.06704
Predicted Residual SS (PRESS)	83.38706

Figure 2: Residual Analysis

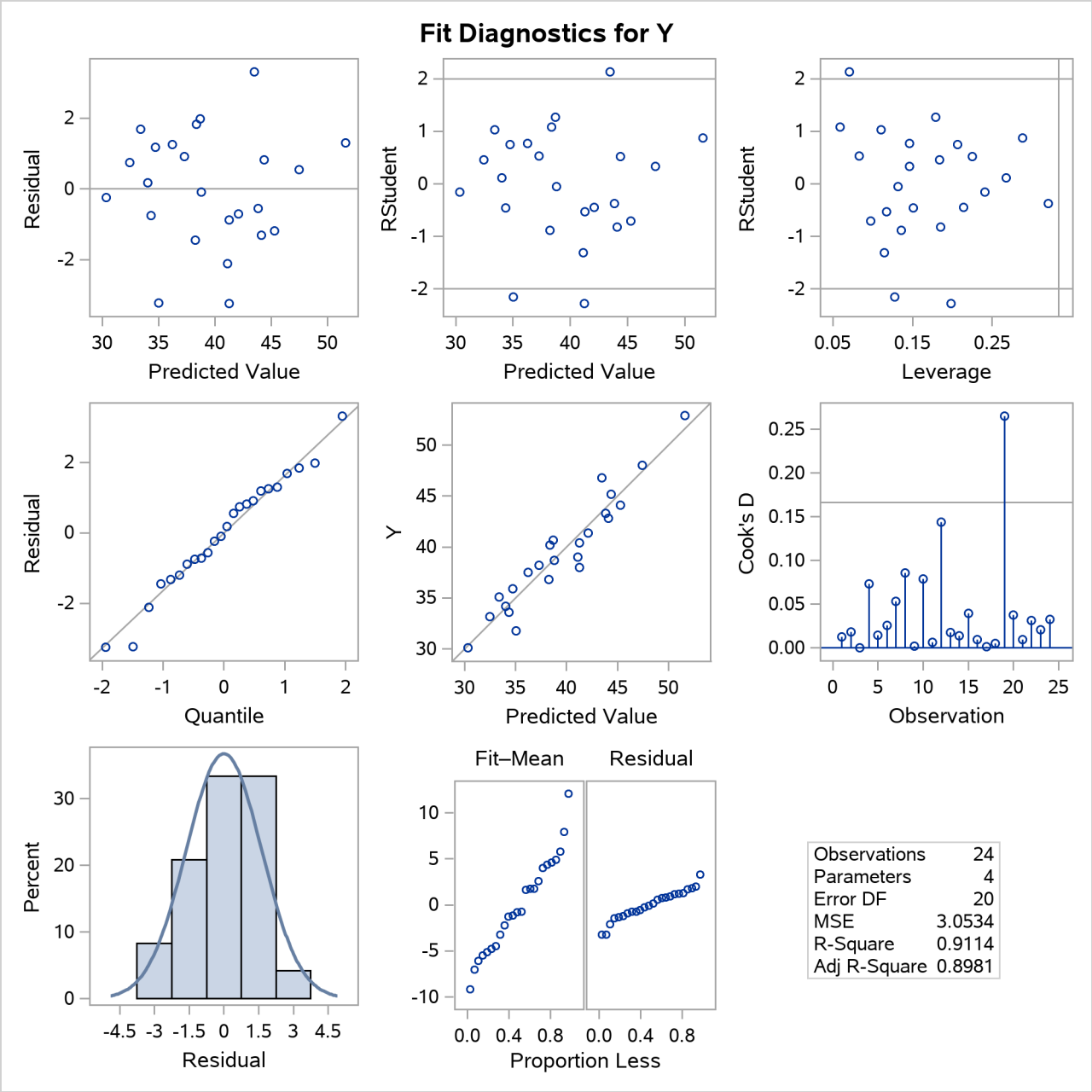
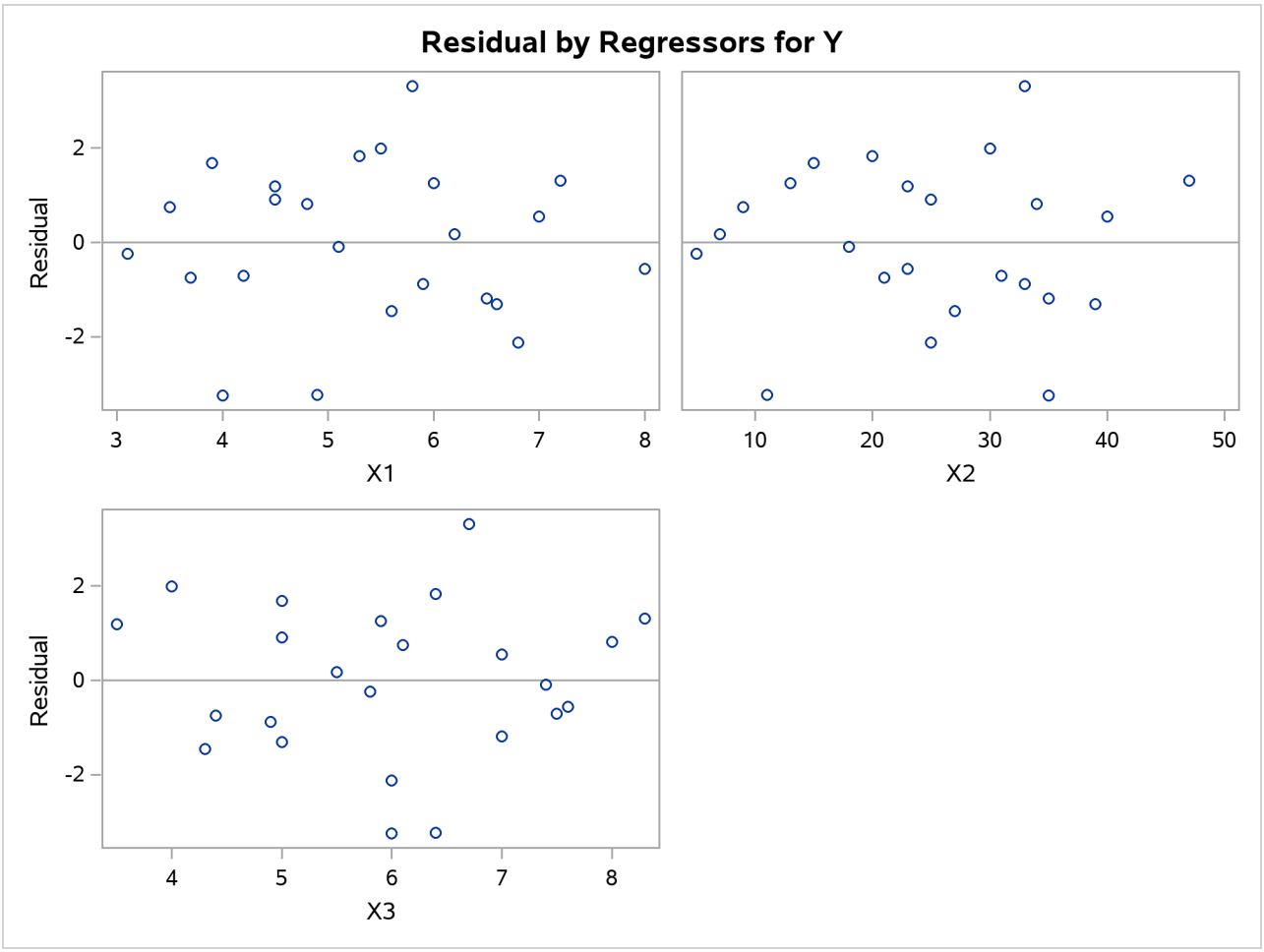
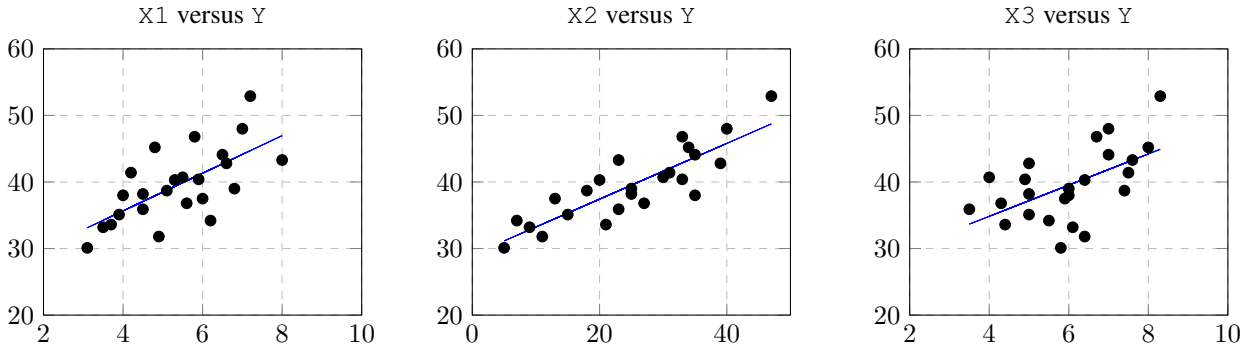


Figure 2: *continued*



Answers to questions 1-7:

Before we can start a regression analysis, we have to check our assumptions for linearity. We can make scatter plots to check this:



From here we can see that the scatter plots show a linear trend. We can assume linearity.

From the SAS results, we can see that our least squares estimated regression equation can be obtained from one of the tables. We just need to take the parameters from the table and these will correspond to β_0, β_1, β_2, and β_3. The regression equation is:

$$Y = 17.84740 + 1.10282(X1) + 0.32175(X2) + 1.28748(X3)$$

where Y is the dependent variable of Salary and $X1$, $X2$, and $X3$ are the independent variables for quality of work, age, and quality of publications, respectively.

The coefficient of determination is calculated using the correlation coefficient. From looking at the scatter plots, we can see there will be positive correlation between each plot. The second graph appears to have the tightest fitting data as well. Our SAS results compute our R^2 value. The coefficient of determination is always between 0 and 1 and tells us what percentage of the data can be predicted confidently from our regression analysis. In other words, an R^2 value of 1 tells us that we are 100% confident that our regression analysis can predict future value. The coefficient of determination for our analysis was 0.9114 or 91.14%. This implies that we are 91.14% confident that our regression analysis can predict the salary.

Now we want to test if there is significant regression relation. We can first state the hypotheses:

$$H_0 : \beta_j = 0 \quad \text{for } j = 0, 1, 2, 3$$
$$\text{vs}$$
$$H_A : \beta_j \neq 0 \quad \text{for } j = 0, 1, 2, 3$$

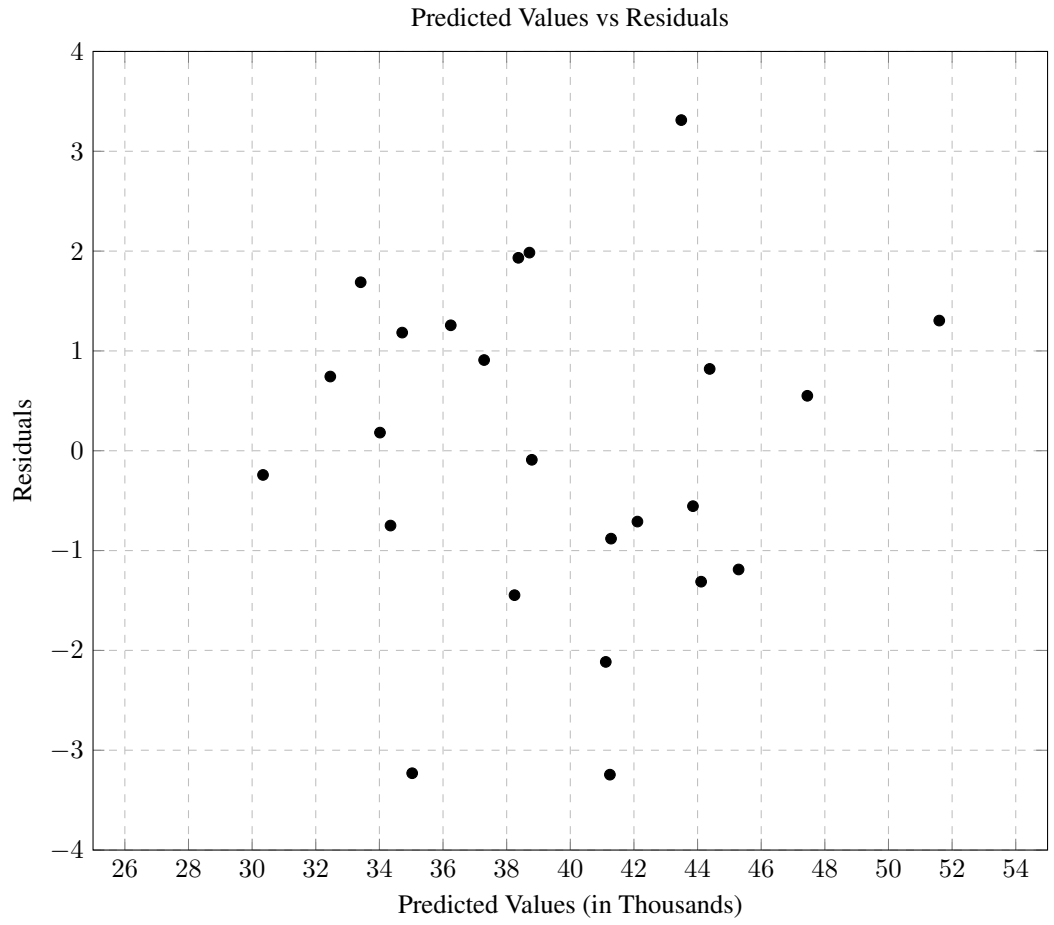
From our SAS results and the `glm` procedure we obtained an F-value of 68.56 for an ANOVA test. This is our test statistic for our test of significance. This leads us to a p-value of < 0.0001 according to our results. The real value is $3.3 \cdot 10^{-8}$ which is nearly 0. Thus, our decision rule is rejecting the null hypothesis in favor of the alternative and say that there is significant regression relation for our model. In other words, our regression model is "good" at predicting values of Y.

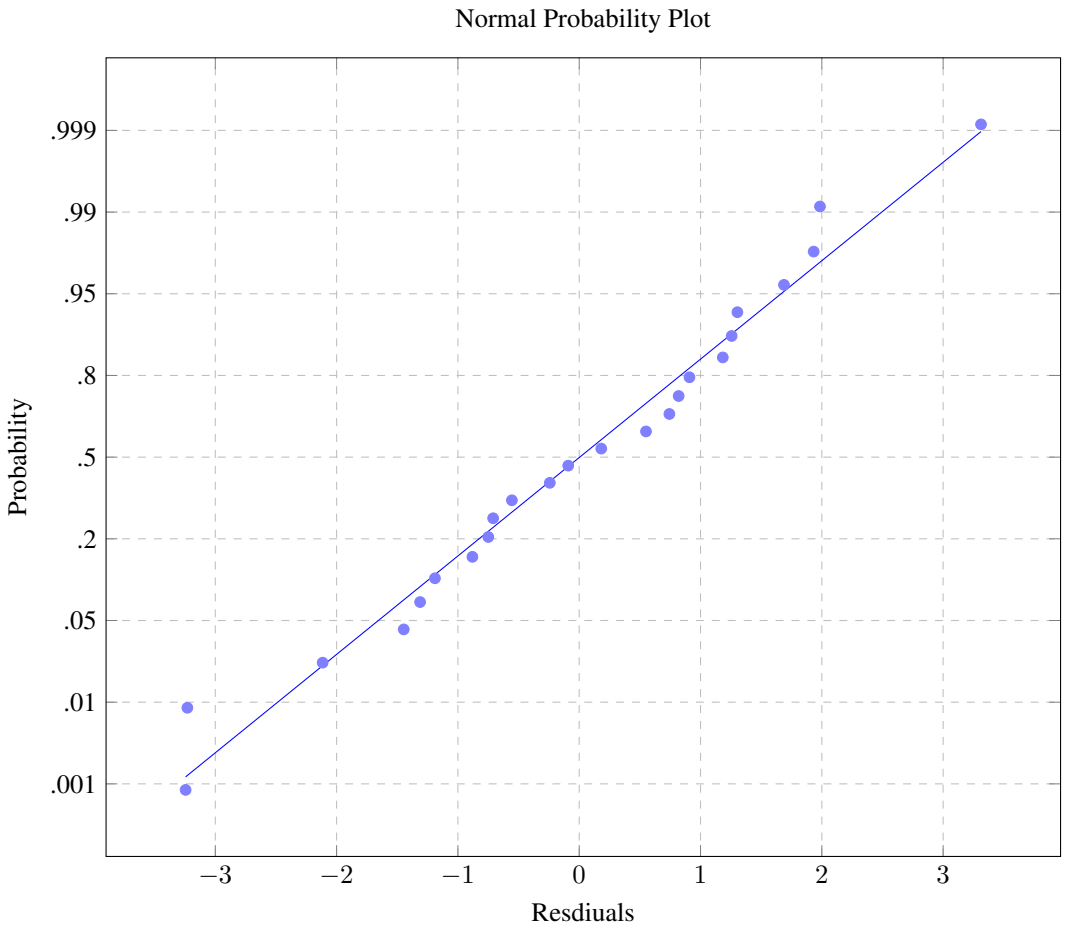
Next, we need to obtain 95% confidence intervals for each of the regression coefficients. We can use the SAS procedure `glm`, with the `clparm` to obtain the confidence intervals for the parameters. The values from the results are

Parameter	Estimate	95% Confidence Limits	
Intercept	17.84740401	13.68435977	22.01044826
X1	1.10282189	0.41745050	1.78819327
X2	0.32175047	0.24458038	0.39892057
X3	1.28747999	0.66677122	1.90818877

These values can be computed from the equation on page 576 as well. Since we were given the standard error values, we just multiple s_ϵ and $t_{\alpha/2,n-2}$ together and add/subtract from the parameter values. We obtain the same results.

Next, we are asked to obtain a plot of the residuals against each predictor variable. For this we simply find the predicted values using our regression equation, then find the residuals by subtracting them from the actual salaries (in thousands). The table of values can be found in the SAS results. Below is a scatter plot of the values followed by a normal probability plot.


Predicted Values vs Residuals
Residuals
Predicted Values (in Thousands)
4
3
2
1
0
−1
−2
−3
−4
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54


Normal Probability Plot
Probability
Resdiuals
.999
.99
.95
.8
.5
.2
.05
.01
.001
−3
−2
−1
0
1
2
3

From our plots above we can see that the residuals in fact follow a normal distribution trend. The points are tightly packed against the trend line. That conclusion leads us to believe we will not need to transform the data, and we can use our linear model for future predictions. The residual data and conclusions can also be found from the SAS data above. I used the `reg` procedure with the `cli` option to obtain the data.