Statistics 512: Solution to Homework#4

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 95%, 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 $X = 5$. (Note that some observations have been omitted from the output.)

```
                  Analysis of Variance
                        Sum of       Mean
Source           DF    Squares     Square   F Value  Pr > F
Model             1      16183      16183   805.62   <.0001
Error            16  321.39597   20.08725
Corrected Total  17      16504

Root MSE        4.48188    R-Square 0.9805
Dependent Mean 64.00000    Adj R-Sq 0.9793
Coeff Var       7.00294

Parameter Estimates
                Parameter  Standard
Variable   DF   Estimate     Error  t Value   Pr > |t|
Intercept   1   -2.32215   2.56435    -0.91     0.3786
x           1   14.73826   0.51926    28.38     <.0001

Output Statistics
          Dep Var  Predicted      Std Error
Obs  x         y      Value   Mean Predict    95% CL Mean   Residual
3    5   78.0000    71.3691         1.0878  69.0630 73.6752   6.6309
4    1   10.0000    12.4161         2.1021   7.9598 16.8724  -2.4161
6    4   62.0000    56.6309         1.0878  54.3248 58.9370   5.3691
8    3   39.0000    41.8926         1.3125  39.1103 44.6750  -2.8926
10   2   33.0000    27.1544         1.6737  23.6064 30.7024   5.8456
```

Solution: To obtain a joint coverage probability of at least 95% for $g = 5$ intervals, we use the Bonferroni correction and construct individual confidence intervals with coverage probability $1 - \alpha/g = 1 - 0.05/5 = 1 - 0.01 = 0.99$. Thus we would construct **99%** confidence intervals for each value of X.
From the SAS output we see that $df_E = n-2 = 16$, and $s\{\hat{Y}_h\} = 1.09$ when $X_h = 5$. To obtain the 99% CI we use $t^c = t(1-\alpha/2g, n-2) = t(0.995, 16) = 2.92$. Thus the 99% CI for the mean when $X = 5$ is $71.3691 \pm 2.921 \times 1.0878 = 71.3691 \pm 3.1775 = [\mathbf{68.2}, \mathbf{74.5}]$.

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

Solution: The design matrix $\mathbf{X}$ has the form $\begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix}$. The transpose $\mathbf{X}'$ has the form $\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 & 10 \end{bmatrix}$. As a result, the matrix $\mathbf{X}'\mathbf{X}$ can be calculated as

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 & 10 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix} = \begin{bmatrix} 5 & 30 \\ 30 & 220 \end{bmatrix}.$$

Using the formula from Chapter 5 (Equation (5.22) on page 191), we have that

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{5 \times 220 - 30^2} \begin{bmatrix} 220 & -30 \\ -30 & 5 \end{bmatrix} = \begin{bmatrix} \frac{11}{10} & \frac{-3}{20} \\ \frac{-3}{20} & \frac{1}{40} \end{bmatrix} = \begin{bmatrix} 1.1 & -0.15 \\ -0.15 & 0.025 \end{bmatrix}.$$

The next part of the formula $\mathbf{X}'\mathbf{Y}$ can be calculated as

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 & 10 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 1+2+3+7+9 \\ 2+8+18+56+90 \end{bmatrix} = \begin{bmatrix} 22 \\ 174 \end{bmatrix}.$$

Finally, we can calculate the least squares estimate $\mathbf{b}$ as

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1.1 & -0.15 \\ -0.15 & 0.025 \end{bmatrix} \begin{bmatrix} 22 \\ 174 \end{bmatrix} = \begin{bmatrix} 24.2 - 26.1 \\ -3.3 + 4.35 \end{bmatrix} = \begin{bmatrix} -1.9 \\ 1.05 \end{bmatrix}.$$

You can check this last calculation using `proc reg` in SAS.

For the following 5 problems, consider the data given in the file `CH06PR18.DAT`, which describes a data set ($n = 24$) used 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 (X_1), number of years of experience (X_2), and an index of publication success (X_3).

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 also zero (give null and alternative hypotheses, test statistic with degrees of freedom, p-value, and brief conclusion in words).

   Solution: The fitted regression equation is $salary = 17.847 + 1.103 \times quality + 0.322 \times experience + 1.289 \times publications$. The R^2 value for this model is 0.911 .

Testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (all regression coefficients are zero) vs H_A : at least one regression coefficient is nonzero, we obtain an F-statistic of 68.12, which with $df = (3, 20)$ gives a p-value of 1.12×10^{-10}. We reject H_0 and conclude that at least one of the explanatory variables has a significant linear relationship with salary.

4. Give 95% 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?

   Solution: The 95% confidence intervals are $[0.415, 1.791]$, $[0.244, 0.399]$, and $[0.666, 1.912]$ for the coefficients on quality, experience, and publications, respectively. For the rest of this problem, we refer to the coefficients as β_1, β_2, and β_3, respectively.

   The following table summarizes the hypothesis tests for the individual coefficients:

Coefficient	Hypotheses	Test Statistic (t)	df	p	Conclusion
quality	$H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$	3.35	20	0.0032	Reject H_0
experience	$H_0 : \beta_2 = 0$ vs. $H_A : \beta_2 \neq 0$	8.66	20	3.35×10^{-8}	Reject H_0
publications	$H_0 : \beta_3 = 0$ vs. $H_A : \beta_3 \neq 0$	4.32	20	0.000333	Reject H_0

   We conclude that quality, experience, and publications has a significant linear relationship with salary. That each of these null hypotheses was rejected is consistent with the finding that none of the confidence intervals contain 0.

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

   Solution: There do not appear to be any obvious deviations from a random scatter of points. The assumptions of linearity and constant variance do not seem to be violated in any obvious way. See Figures 1 and 2.

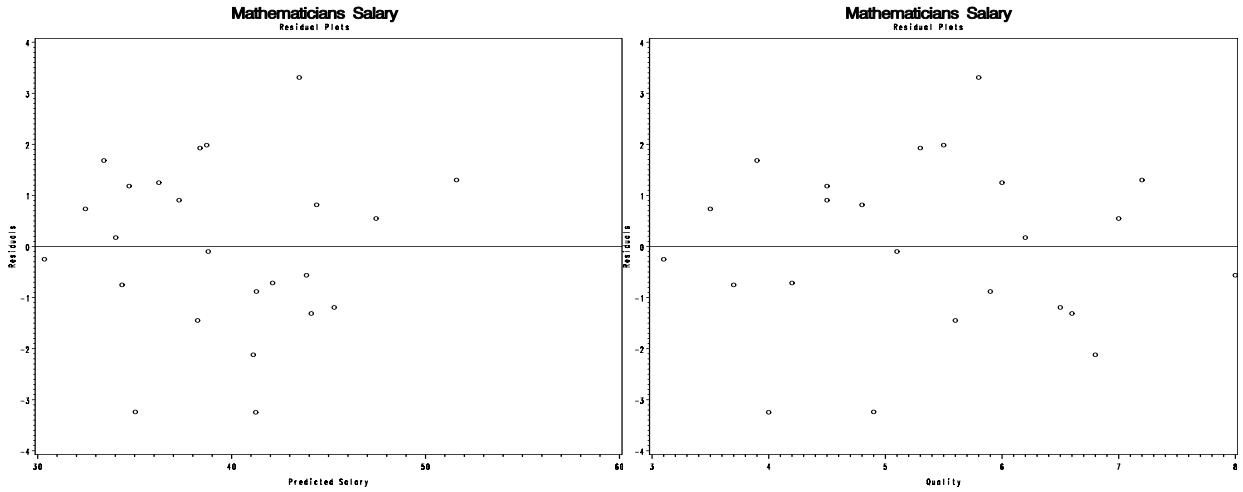

Figure 1: Residual plots for Problem 5

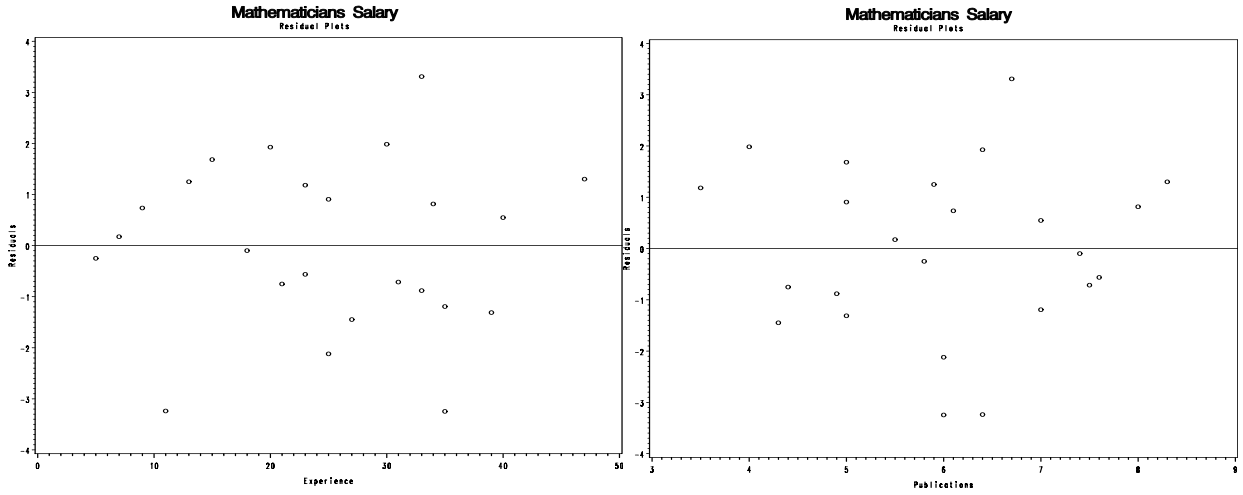

Figure 2: Second set of residual plots for Problem 5

6. Examine the assumption of normality for the residuals using a qqplot and histogram. State your conclusions.

   Solution: The residuals appear to be fairly normal. See Figure 3.

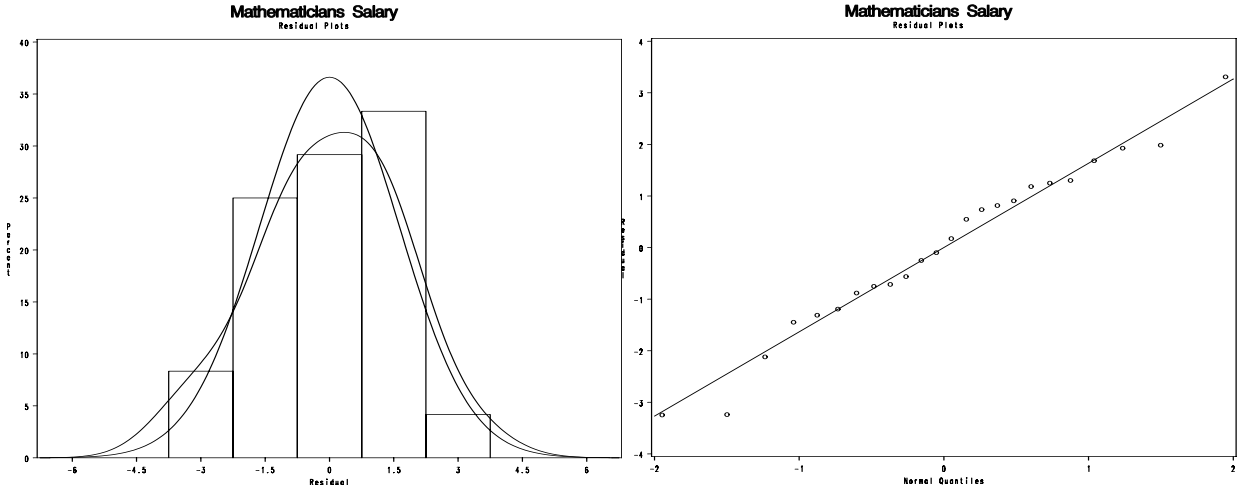

Figure 3: Histogram and qqplot for Problem 6

7. Predict the salary for a mathematician with quality index equal to 6.2, 8 years of experience, and publication index equal to 5.9 . Provide a 95% prediction interval with your prediction.

   Solution: The model predicts a salary of 34.86 (i.e., \$34.863), with a 95% interval of $[30.79, 38.94]$ (i.e., $[\$30,786, \$38.940]$).