1. Short Answer

a. List the three major assumptions of a simple linear regression model. For each assumption, give one way that you might check whether that assumption is satisfied and identify one remedy that might be used to adjust for the problem if it exists.

Assumption #1: Normality of Error Terms
Diagnostic: Examine a Normal QQ plot of the residuals for linearity; Examine a histogram; Shapiro-Wilks Test
Remedy: Transformation on Y (may use Box-Cox to find one)

Assumption #2: Constancy of Variance (among residuals)
Diagnostic: Examine residual plots (vs. Y-hat or X's) for different "vertical spreads"; Modified Levine or Breusch-Pegan Tests
Remedy: Transformations on Y may help or use Weighted Least Sqauares

Assumption #3: Independence of Error Terms
Diagnostic: Plot the residuals against time or sequence
Remedy: Account for dependency in the model (perhaps by including time or see Ch12)

b. (10 points) You are designing an experiment in which you want to relate corn yield (Y, in bushels per acre) to the amount of fertilizer used on the plot (X, in pounds per acre). Based on previous experiments, you believe that the ideal fertilizer amount is between 10 and 12 pounds per acre. You would like to obtain prediction intervals at X = 10.5, X = 11, and X = 11.5. (a) In terms of experimental design, explain what you could do to try to minimize the widths of these intervals. (b) Explain what you would do to obtain these three intervals with a *family confidence level* of 95%.

For part (a), note that $s^2\{pred\} = MSE\left[1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{(X_i-\bar{X})^2}\right]$. So there are three things we can do: (1) Add observations to increase n, (2) Make $\bar{X}=11$ so that the third term has a small numerator, and (3) Increase the spread in the X's so that SSX in the denominator is increased.

For part (b), we would want to use a Bonferroni correction, taking $\alpha = 0.5/3 = 0.0167$.

2. Omit

3. Refer to the SAS output marked OUTPUT FOR PROBLEM 2. The data are from a study of company executives. The response variable is annual `salary` (dollars), and the two explanatory variables used are `gender` (0 = female, 1 = male) and `exper` (experience in years). The variable `expgen` is the product of `gender` and `exper`.

a. Write down the linear model used in this analysis, including the distributional assumption. *(5)*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \text{ where } X_1 = \texttt{gender}, X_2 = \texttt{exper}, X_3 = X_1 X_2$$

$$\varepsilon_i \sim \text{iid } N\left(0, \sigma^2\right)$$

b. Write the estimated regression equation (2 pts). Then, write two separate fitted lines predicting `salary` from `exper`: one for females (2 pts) and one for males (2 pts).

$$\hat{Y} = 58050 + 7799X_1 + 2045X_2 + 864X_1X_2$$

females: $\hat{Y} = 58050 + 2045X_2$

males: $\hat{Y} = (58050 + 7799) + (2045 + 864)X_2 = 65849 + 2909X_2$

c. Estimate the mean salary for women with 5 years of experience.

$$\hat{Y} = 58050 + 2045 \times 5 = 58050 + 10225 = 68275$$

d. How do you justify that either gender or experience or their product, or any combination of those three variables is useful to predict salary?

The *F* test in ANOVA table gives F value = 98.09 and p-value <0.0001, so we conclude that at least one of those three variables has linear relationship with Salary.

e. Does residual show anything strange? Explain why or why not.

The residuals appear to be normal from the normal quantile plot, the variance seems to be constant from the residual scatter plots, and the residuals are approximately equally distributed above and below the zero-line. So nothing seems strange with residuals.

4. Suppose we have an incomplete ANOVA table in studying the linear model of predicting Salary by quality, experience and publication.

a. Complete the ANOVA table.

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			

b. Write down the linear model and the hypothesis for the *F* test. What is your conclusion from this test?

The conclusion from this test is that at least one of the three variables: quality, experience and publication has linear relationship with salary.

Suppose that we miss the last three columns, `t Value, Pr > |t|, 95% Confidence Limits`, in the following table.

c. Fill in the blanks for experience based on the information from the ANOVA table and the above table.

		Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	17.84693	2.00188	8.92	<.0001	13.67109	22.02277
quality	1	1.10313	0.32957	3.35	0.0032	0.41565	1.79061
experience	1	0.32152	0.03711	8.66	<.0001	0.24411	0.39893
publications	1	1.28894	0.29848	4.32	0.0003	0.66632	1.91156

d. Write down the hypothesis for the *t* test in c. What is your conclusion from this test?

The conclusion from *t* test is that given quality and publications included in the model, the experience is still significantly useful to predict salary.