Engineering Statistics II, HW 4

Due Start of class, Friday, Sept. 26

Instructions: This homework is due at the beginning of class Friday, Sept. 26. A subset of these problems will be graded.

Additional Instructions: Be sure to include copies of all *relevant* Minitab output in your hw. For additional instructions see those for hw3 which is posted on the course website.

1: Suppose that a random sample of 50 bottles of a particular brand of cough syrup is selected, and the alcohol content of each bottle is determined. Let μ denote the average alcohol content for the population of all bottles of the brand under study. Suppose that the resulting 95% confidence interval is (7.8,8.4). Answer the following questions:

i. Would a 90% confidence interval calculated from this sample have been narrower or wider than the given interval? Explain your reasoning.

ii. Consider the following statement: There is a 95% chance that μ is between 7.8 and 8.4. Is this statement correct? Why or why not?

iii. Consider the following statement: We can be highly confident that 95% of all bottles of this type of cough syrup have an alcohol content that is between 7.8 and 9.4. Is this statement correct? Why or why not?

iv. Consider the following statement: If the process of selecting a sample of size 50 and then computing the corresponding 95% interval is repeated 100 times, 95 of the resulting intervals will include μ. Is this statement correct? Why or why not?

2: Testing whether or not a person has a disease is essentially a hypothesis test. Suppose that the null hypothesis is that the person does not have the disease and that the alternate hypothesis is that the person has the disease. The **sensitivity** of a diagnostic test is the probability that the test is positive (declares disease present) when the person does in fact have the disease. The **specificity** of a diagnostic test is the probability that the test is negative (declares disease not present) when the person in fact does not have the disease. Suppose the sensitivity of a certain test is 0.99 and the specificity is 0.95. Answer the following questions concerning the "before-test-administered" error probabilities:

i. What is the probability of a type I error, α, using this test?

ii. What is the probability of a type II error, β, using this test?

Note: Ignorance of statistics can be fatal! The fact that most people do not know and/or understand hypothesis testing and that both types of error have nonzero probability has caused problems. For example, back in the 80's when the life expectancy of those diagnosed with AIDS was measured in months, testing centers mandated that individuals get their

test results in person from a counselor. They did this because, in the absence of counseling about the accuracy of the test, many people committed suicide upon getting a positive - very possibly false positive - result.

3: A manufacturer of handheld calculators receives very large shipments of printed circuits from a supplier. It is too costly and time-consuming to inspect all incoming circuits, so when each shipment arrives, a sample is selected for inspection. Information from the sample is then used to test H_o: $\pi = 0.05$ versus H_a: $\pi > 0.05$, where π is the true proportion of defectives in the shipment. If the null hypothesis is not rejected, the shipment is accepted, and the circuits are used in the production of calculators. If the null hypothesis is rejected, the entire shipment is returned to the supplier due to inferior quality. (A shipment is defined to be of inferior quality if it contains more than 5% defectives.)

i. In this context, define type I and type II errors.

ii. From the calculator manufacturer's point of view, which type of error would be considered more serious? What Greek symbol denotes the probability of this type of error?

iii. From the printed circuit supplier's point of view, which type of error would be considered more serious? What Greek symbol denotes the probability of this type of error?

iv. How can the probabilities of both type I and type II errors be reduced in order to satisfy both the printed circuit manufacturer and calculator manufacturer requirements?

4: Suppose response variable Y and predictor variable x obey the simple linear model with the following values for the three parameters: $\beta_0 = 1$, $\beta_1 = 2$, and $\sigma^2 = 9$.

i. Recall that the simple linear model completely determines the distribution of Y for any given value of x. What is the distribution of Y when $x = 1$?

ii. Use your answer from i to compute the probability that the next value of Y observed when $x = 1$ is between 0 and 6.

iii. Changing the value of x from 1 to 2 has what effect on the distribution of Y?

5: Suppose our bivariate data consists of the following observations:

```
{ (1,1), (2,1), (3,2), (4,2), (5,4) }
```

i. Manually compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the matrix formula $\hat{\beta} = (X'X)^{-1}X'\vec{Y}$. Show all your work for credit.

ii. Verify your estimates and compute the residuals using Minitab's regression procedure. Include a copy of Minitab output in your hw.

iii. Using the residuals from ii, manually compute $\hat{\sigma}^2$. Show your work.

iv. See next page ...

iv. Manually compute the covariance matrix $\sigma^2(X'X)^{-1}$ and determine the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

v. Manually compute the t-test statistic for testing $H_o : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Show your work. Also, compute the p-value using Minitab's `Graph -> Probability Distribution Plot`. At $\alpha = 0.05$ what do you conclude?

vi. Check your work on part v by comparing it with your Minitab output from part ii.

6: In this problem you will model the relationship between the flexural strength (ability to resist failure in bending, in MPa) of concrete beams, strength, and the modulus of elasticity (moe) the ratio of stress to strain, i.e., deformation in per unit length, in GPa). You will then use your regression model to meet a design specification. Do the following:

i. Open the data set for this problem on the course website and fit the linear model

$$\text{strength}_i = \beta_0 + \beta_1 \text{moe}_i + \epsilon_i$$

ii. Verify that this model satisfies the regression asumption by analyzing the residuals. Be sure to include copies of relevant plots and other Minitab output in your hw. Since the order of the data is not known you can't check assumption 3 using the residuals. Further, since no description of how the data was collected we have to assume this assumption is met.

iii. Suppose for a certain application we need 95% of the beams to have a flex strength of at least 7 GPa. Use your model from part i to determine the smallest moe which will ensure this specification is met. Hint: compute an appropriate interval - CI or PI, you decide which is appropriate - with an appropriate level - again, you determine what level is needed.

7: A research engineer is investigating the use of a windmill to generate electricity. She has collected data on the DC (direct current!) output from the windmill and the corresponding wind speed. Your goal is to construct a model to predict DC output using wind speed. If possible you should construct a model which incorporates the fact that the DC output saturates (approaches an upper limit or asymptote) as the wind speed increases. Analyze the data using Minitab as follows:

i. Download the data for this problem from the course website.

ii. Construct a scatterplot of DC vs. speed.

iii. Based on the scatterplot, it appears a quadratic polynomial model might be appropriate:

$$\text{DC}_i \quad = \quad \beta_0 + \beta_1 \text{speed}_i + \beta_2 \text{speed}_i^2 + \epsilon_i$$

Fit this model and compute the residuals.

iv. See next page ...

iv. To the extent you can, check and comment on the regression assumptions using appropriate plots and procedures. Since the order of the data is unknown and no info was provide on how it was collected, you must assume that the random errors are independent.

v. Irrespective of your conclusions above, note that since the coefficient on the quadratic term is negative, the model inappropriately predicts decreasing DC output with increasing wind speed at some point (presumably before the windmill blows over). Devise a linear model which possesses a horizontal asymptote and which satisfies the random error assumptions. Demonstrate that it satisfies the four assumptions using appropriate plots and procedures. Hint: use a model of the form

$$\mathrm{DC}_i \quad = \quad \beta_0 + \beta_1 x_i + \epsilon_i$$

in which x is an appropriately transformed version of speed which provides your model with an asymptote as wind speed goes to infinity.

8: (*A Phat application*) Recall that previously we analyzed data for the study described below:

> The most accurate method for measuring intraabdominal fat requires a CT which is expensive and exposes the patient to x-ray radiation. Despres, et al, investigated the possibility of estimating intraabdominal fat area using a person's waist size, clearly a cheaper, safer measurement. They collected the waistsize (x) and the CT-measured intraabdominal fat area (y) of 109 randomly selected men between the ages of 18 and 42 years. They then investigated predicting intraabdominal fat using the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

Recall that although the data satisfied the first regression assumption they failed the second assumption - the data are heteroscedastic. According to your *Residual Analysis and Model Refinement* handout one way to achieve homoscedasticity is to transform the response variable, i.e., replace the response variable y with $\tilde{y} = f(y)$ for a judiciously chosen function f. When variance is an increasing function of the predictor variable, using $f(y) = \ln(y)$ often solves the problem. In other words using the response variable $\tilde{y} = \ln(y)$ yields a model with homoscedastic residuals. Try this approach as follows:

i. Open the data set for this experiment:

> `http://www.rose-hulman.edu/~inlow/waistdeepfat.MTW`

and create a column containing $\tilde{y} = \ln(y)$ using Minitab's `Calc -> Calculator` and the `LN` function.

ii. Fit the linear model $\tilde{y}_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and compute the residuals using Minitab.

iii. Plot the residuals vs. the predictor x. What do you conclude about the first assumption? Because the model $y = \beta_0 + \beta_1 x + \epsilon$ satisfied the first assumption (which implies y *is* a linear function of x) $\tilde{y} = \ln(y)$ is **not** a linear function of x. One way to solve this

problem is to use a polynomial model. Often a second order polynomial model is good enough so fit the following 2nd order model

$$\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

using Minitab and verify that it satisfies the four regression assumptions.

iv. Use your model from part iii to construct a 95% CI for $\mu_{\tilde{y}|x=100}$. Transform this interval to create a 95% CI for $\text{median}_{y|x=100}$.

v. Use your model from part iii to construct a 95% PI for $\tilde{y}|x = 100$. Transform this interval to create a 95% PI for $y|x = 100$.