Engineering Statistics II, HW 5

Due Start of class, Friday, Oct. 3.

Instructions: This homework is due at the beginning of class Friday, Oct 3. 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.

0: Be sure to have read the *Residual Analysis and Model Refinement*, *Multiple Regression Analysis*, *Coding Quantitative Independent Variables*, and *Models with Two Quantitative Independent Variables* handouts plus sections 11.1-11.3, 11.4.1, 11.7, 11.9, 12.1, and 12.6.1 plus **new** sections 11.6 and 12.3.

1: When fitting polynomial models with observational data, why should one always standardize each predictor, i.e., replace each predictor x_j with $(x_j - \bar{x}_j)/S_{x_j}$, before generating higher order terms, e.g., x_j^2, x_j^3, etc.?

2: The data set for this problem consists of water pressure measurements (`p`) at varying temperatures (`t`). This data set provides an example of why it is better to use the `bctrans17` macro (available on the course website) than the Box-Cox option in Minitab's regression routine.

i. Open the data set for this problem on the course website and create a scatter plot of pressure (`p`) vs. temperature (`t`). Include a copy of this graph in your hw.

ii. Fit the quadratic model

$$\mathrm{p} \;=\; \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon$$

by squaring the values for `t` in column `c2` and storing them in column `c3`. Ask Minitab to use the optimal value of λ via the `options` submenu. Note the value Minitab uses and the confidence interval it provides for λ.

iii. Now determine the optimal value for λ using the `bctran17` macro by installing the macro (it's available on the course website) and issuing the following command in the Minitab session (upper) window after the prompt (`MTB >`):

```
%bctrans17 c1 c2 c3.
```

and wait for the macro to finish. Note that when using this macro the first column you specify is the response variable (here `p`) and the remaining columns contain the predictor variables. Note that the log-likelihood plot, the left plot in the `bctrans17 output` (include a copy of this output in your hw) has two peaks and Minitab's regression has picked the greater of the two. But why the peak near zero? Recall that when λ is 0 we

log transform y to get $\tilde{y}) = \ln(y)$. So the second peak in the `bctrans17` log-likelihood plot is telling us that $\ln p = f(t)$ is a good model, info not provided by Minitab's regression routine. This is important because a major goal of regression modeling is to provide insight into relationships between variables. The theoretical model relating pressure and temperature is the Clausius-Clapeyron relationship which states $\ln(p) \propto -1/t$. Thus by blindly using the Box-Cox transformation with a quadratic model we correctly identified the correct transformation for pressure. There are advanced methods we will not consider which can also identify transformations for the predictors.

iv. Use Minitab's regression routine and `bctrans17` to determine λ for the model

$$p^{\lambda} \quad = \quad \beta_0 + \beta_1 t^{-1}$$

by creating a column in Minitab containing the inverse values of temperature in `c2` and then using both routines. Notice that $\lambda = 0$ for the above model yields the correct, Clausius-Clapeyron, model. Which output do you feel is more informative - that from Minitab's regression routine or that from the `bctrans17` macro? Include copies of relevant output from both routines in your hw.

3: The data set for this problem consists of the salaries and years of experience of a simple random sample of 50 social workers. Analyze it as follows:

i. Open the data set for this problem on the course website and construct a regression model to predict salary using years of experience which adequately meets the regression assumptions. If necessary, use higher order terms and/or variable transformations.

ii. Taking into account how the data was collected and using residual analysis, demonstrate that your model adequately satisfies all regression assumptions. Be sure to include copy of all relevant Minitab output (e.g., graphs) in your hw.

iii. Demonstrate, using t-tests at $\alpha = 0.05$, that all terms in your final model are statistically significant. Be sure to control multicollinearity/VIF's. Be sure to include a copy of Minitab's regression coefficient table for your final model.

iv. Construct a 90% CI for the intercept term (β_0) of your model.

v. Use your model to construct a 95% prediction interval for the salary of a randomly selected social worker with 15 years of experience. (Be glad that you are not pursuing social work as a career.)

4: The data for this problem (provided on the website) consists of the results of a designed experiment to maximize the percent yield (y) of a chemical process as a function of time and temperature. Analyze this data as follows:

i. Code the predictor variables.

ii. Model the relationship between the yield and time and temperature using a full second order polynomial model.

iii. Analyze the residuals to determine if the regression assumptions are met.

iv. Assuming the regression assumptions are met, eliminate any unecessary terms from the model using $\alpha = 0.10$. Refit your new, refined model.

v. Using the shortcut procedure in the handout, show that your model indicates that y achieves a maximum.

vi. Using calculus techniques, estimate values of time and temperature which maximize the yield by computing the values of time and temp which maximize your model.

vii. Compute a point estimate of the yield maximum. Also, compute 95% confidence and prediction intervals for the yield at the values you computed in part vi. What does the prediction interval say about the next measurement of yield at the values computed in part vi?

5: The data for this problem consists of the response variable, `thick`, polysilicon wafer thickness in angstroms, and two predictor variables: oxide thickness in angstroms (`o`) and deposition time in seconds (`t`). Analyze this data as follows:

i. Code the predictor variables and model the relationship between polysilicon thickness and the two predictor variables using a full second order model.

ii. Assume all regression assumptions are met (do not check them) and use coefficient hypothesis tests ($\alpha = 0.05$ to eliminate nonsignificant terms.

iii. Using your refined model from part ii, determine the amount of time needed to maximize thickness when oxide thickness is 1000 angstroms. Note that with oxide thickness set to 1000 angstroms, wafer thickness is a function just of time. Use calculus methods to demonstrate that this function does in fact have a maximum.

iv. Use your refined model to estimate the average thickness when oxide thickness is 1000 angstroms and time is set to the maximizing value you determined in part iii.