Mathematics 376 – Mathematical Statistics
Lab Project 5 – A Regression Case-Study
May 4 and 7, 2012

*Background*

In class, we have discussed the basic methods for obtaining the least-squares estimators for the coefficients $\beta_i$ in simple linear models

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \tag{1}$$

and techniques for hypothesis testing concerning the $\beta_i$ (see page 585 for the case $k = 1$ and models of the form $Y = \beta_0 + \beta_1 x + \epsilon$, and page 618 in the text for the general case).

In this lab project, we will work through a "case-study" of how these methods might be used.

*The "Story"*

Utility companies, which must plan the operation and expansion of electricity generating facilities, are vitally interested in predicting customer demand over both short and long periods of time. A short-term study was conducted to investigate the effect of

- $x_1$ = each month's *daily mean temperature, in degrees Fahrenheit*, and
- $x_2$ = the *cost per kilowatt-hour (in dollars)* per household.

The company officials expected the demand for electricity to rise in cold weather (due to heating), fall when the weather was moderate, then increase again in hot weather (due to air conditioning). They also expected demand to decrease as the cost increased. Because of the expectation about the dependence on $x_1$, a model just involving $x_1$ to the first power was thought to be inappropriate, so they set up a model of the form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2 x_1 + \beta_5 x_2 x_1^2 + \varepsilon \tag{2}$$

The data they had to work with was as follows:

- With $x_2 = .08$ (i.e. 8 cents per kilowatt-hour):

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ (*temp*) | 31 | 34 | 39 | 42 | 47 | 56 | 62 | 66 | 68 | 71 | 75 | 78 |
| $y$ (*demand*) | 55 | 49 | 46 | 47 | 40 | 43 | 41 | 46 | 44 | 51 | 62 | 73 |

- With $x_2 = .1$ (i.e. 10 cents per kilowatt-hour):

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ (*temp*) | 32 | 36 | 39 | 42 | 48 | 56 | 62 | 66 | 68 | 72 | 75 | 79 |
| $y$ (*demand*) | 50 | 44 | 42 | 42 | 38 | 40 | 39 | 44 | 40 | 44 | 50 | 55 |

(Note: The data were from different locations and months so the temperatures don't match up exactly.)

A) Start by formatting the data in a way appropriate for entry into `R`. You may use a data file if you want.
B) Compute the least squares estimators for the coefficients $\beta_i$ in the model (2).
C) What does the model predict about the demand in months with average temperatures between 20 and 90 degrees Fahrenheit (in increments of 10 degrees Fahrenheit) and costs between .06 and .12 per kilowatt-hour (in increments of .02)? Report all these estimated values, and a 95% prediction interval for each. (See Section 11.13 in the text and the notes and `R` example from class on 4/30 for the needed formulas.) Note: You will need to decide how to organize all of these predicted demand values!
D) Now, we want to consider a different question. Namely: *Did including the $x_2$ terms in (2) really give us any added predictive power?* Think of (2) as the *complete* model as in class on 5/2, and consider the *reduced model*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \tag{3}$$

where the $x_2$ terms have been removed (i.e. $\beta_3 = \beta_4 = \beta_5 = 0$). Is there sufficient evidence to indicate that the model (2) gives significantly better predictive value about $Y =$ demand than the reduced model (3)? Test using the procedure outlined in class on 5/2 and in §11.14 in the book. Recall that the needed information can be found in the `R` analysis of variance table ( generated by the `anova` command) for the linear model if you know where and how to look. Explain which entries there are relevant and how you are drawing your conclusion. Also what does this say in real world terms?

*Assignment*

Individual writeups, due at the end of class on Monday, May 7.