

Stat 111 Spring 2011 Week 9 - More Multiple Regression

1. I have a cauldron of coins (pennies, nickels and dimes) in my office. I had $n = 26$ students each draw a small handful of coins and record $X_1 =$ number of coins, $X_2 =$ number of pennies, and $Y_1 =$ total value in cents. They also recorded $Y_2 =$ weight in grams of each handful (data in the file coins.txt).
 - a) Discuss the interpretation of simple and multiple regression coefficients for regressions of Y_1 or Y_2 on X_1 and/or X_2 . Note that it would be reasonable to fit a no-intercept model for some of these regressions.
 - b) Show other less contrived real data examples of *Simpson's paradox*.
 - c) Show how to get the multiple regression coefficients from simple regression output. Fit a simple regression of Y on X_2 and save the residuals. Then fit a regression of X_1 on X_2 and save the residuals. Finally, fit a simple regression of the first set of residuals on the second set of residuals. The fitted slope is the slope for X_1 in the multiple regression. An analogous procedure gives the coefficient for X_2 in the multiple regression. JMP's Leverage Plots graph these "leverage residuals," after recentering them to the average values of the corresponding variables.
 - d) Show how to find a confidence interval for a mean response, and a prediction interval for an individual response, and explain the difference.

2. The ANOVA table and F tests for regression.
 - a) Show that a 1-sample t analysis is equivalent to linear regression on a constant, and that a pooled 2-sample t analysis is equivalent to linear regression on a variable that takes values 0 and 1 (an "indicator variable," or "dummy variable"). Use the data for height and gender as an example.
 - b) Describe the ANOVA table in the context of regression on indicator variables. Describe the whole-model F -test and derive equivalent tests based on R^2 and adjusted R^2 .
 - c) Analysis of Covariance (ANCOVA) refers to a multiple regression with a numeric response Y , and with categorical and numeric variables as X 's, along with possible interactions.
 - d) Explain the "Extra Sum of Squares F test" for multiple coefficients, such as categorical predictors with more than two categories (or interactions with categorical predictors).
 - e) Stat 11 students recorded the Centennial Conference men's basketball data found in the file CCBB.txt. Each row represents one season for one of the ten teams. There are many variables (columns) recorded for each season and team, making this a good example for methods of variable selection. Find a good model for predicting a team's winning percentage from statistics other than points scored (if you include points scored, no other statistics will seem important at all). There is a stepwise regression procedure that automates the choice of variables. Discuss what it means for a predictor to be important and what it means to "find a good model."

3. Swarthmore Biology professor Sara Hiebert Burch came to me with this problem. Theory predicts that, within a certain range of temperature values, the metabolic rate Y_i of a hamster will decrease linearly with temperature until a lower critical temperature θ is reached. At that point, the decrease remains linear, but with a less negative slope. For each hamster, measurements of metabolic rate Y_i are made at $n = 7$ temperature values x_i ($i = 1, \dots, n$). Consider the following model:

$$Y_i = \mu_i + \epsilon_i; \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2); \quad \mu_i = \begin{cases} \beta_0 + \beta_1 x_i, & x_i < \theta. \\ \beta_0 + \beta_1 x_i + \delta(x_i - \theta), & x_i \geq \theta. \end{cases}$$

- Write out the likelihood function for the five unknown parameters.
- For a given value of θ , find expressions for the maximum likelihood estimates of β_0 , β_1 , δ and σ^2 . You may express the estimate of σ in terms of the other estimates, and all will depend on θ .
- The table below gives temperatures (degrees C) and metabolic rate measurements for one of the hamsters.

Temp	5.3	11.8	19.3	23.8	25.6	27.6	31
MR	6.7	4.4	3.3	2.12	2.02	1.82	1.74

Graph the metabolic rate values against temperature and draw in the best fitting lines (based on the MLE) for $\theta = 13.59$ and for $\theta = 23.24$.

- Make a graph of the maximized likelihood function against a grid of θ values between $t_1 = 12$ and $t_2 = 27$ (imagine you have theoretical justification for setting those boundaries). If you scale the likelihood values to have a maximum value of 1.0, this is a graph of the GLR statistic for testing different values of θ . Which of the values from part c is more consistent with the data according to this criteria?

Problems to turn in:

- Suppose $Y_i \stackrel{\text{indep}}{\sim} N(X_i' \beta, \sigma^2)$; $X_i' = (1, x_{i1}, \dots, x_{iq})$.
 - Treat σ as known and assume an improper q -dimensional uniform (constant) prior density for β . Derive the conditional posterior density for $\beta | y, \sigma^2$.
 - Let $\hat{\beta}$ be the usual least squares estimate for β . Find the posterior distribution of $\hat{\beta} - \beta | y, \sigma^2$, and compare this to the sampling distribution of $\hat{\beta} - \beta | \beta, \sigma^2$.
 - Find the posterior density for $\beta | y$, assuming a “non-informative” joint prior density $p(\beta, \sigma^2) \propto 1/\sigma^2$, $\sigma^2 > 0$, $\beta \in \mathbb{R}^q$.
- Find the sampling distribution of $\hat{\beta}$, the MLE for the *weighted* regression model (\mathbf{W} is a known symmetric $n \times n$ positive definite matrix, and \mathbf{X} is $n \times p$ and is also known):

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{W}^{-1}).$$