**Homework 2: Linear and Multiple Linear Regression Models**

**Goals:**

Use R to perform single and multiple variable linear regressions. Learn how to analyze the quality of the models.

**Description:**

Students are divided into groups of 3-4. Please see WISE for your group division. Please read the textbook and the multiple linear regression example posted on WISE under HW2 before you start working on this homework.

Using a simulated fake dataset to perform the single and multiple linear regression, then use the given dataset to build single and multiple linear regression models, perform model quality analysis, write a report to justify the parameter selections and model results.

In regression, the goal is to find the parameters

**Part 1: Experiment on fake data set (All groups)**

1. Simulate a fake dataset with the code in P.70 of the textbook. Then perform the following tasks:
   1. Build a regression model and see it recovers the true values of betas.
   2. Simulate another fake variable x\_2 that has a gamma distribution with parameters you pick. Now make the truth be that y is a linear combination of both x\_1 and x\_2. Fit a model that only depends on x\_1. Fit a model that only depends on x\_2. Fit a model that uses both. Vary the sample size and make a plot of mean square error (plotting this for different sizes in loop) of the training set and of the test set versus sample size.
   3. Create a new variable z, that is equal to x\_1^2. Include this as one of the predictors in your model. See what happens when you fit a model that depends on x\_1 only and then also on z. Vary the sample size and make a plot of mean square error of the training set and f the test set versus sample size.
   4. Play around more by
      1. Changing parameter values (the true betas)
      2. Changing the distribution of the true error, and
      3. Including more predictors in the model with other kinds of probability distributions. (rnorm() means randomly generate values from a normal distribution. rbinom() does the same for binomial. Look up these functions online and try to find more)
   5. Create scatter plots of all pairs of variables and histograms of single variables.

**Part 2: Experiment on real data sets**

**Team 1, 2, 3:**

1. A group of researchers are interested in developing a model that describes the gas mileage (in mpg) of sports utility vehicles. They decide to use engine size (in cubic cm), horsepower and weight of the car (in pounds) as explanatory variables. From a random sample of 11 SUVs they obtain the following data:

Engine HP Weight MPG

3471 260 4420 23

2979 225 4586 21

4195 275 4787 20

4701 235 4379 19

3471 240 4439 22

3960 195 3786 21

4701 235 3786 20

4701 265 3786 19

3311 230 3860 24

4664 235 5390 17

4605 302 4834 19

1. Find the least-squares regression equation using MPG as the response variable and engine size, weight and horse power as explanatory variables. Write down the regression equation.
2. Explain in context what the coefficient corresponding to horse power

means.

1. Conduct the F-test for the overall fit of the regression. Comment on the results.
2. Test each of the individual regression coefficients. Do the results indicate

that any of the explanatory variables should be removed from the model?

1. Determine the regression model with the explanatory variable(s)

identified in part (d) removed.

1. Going back to the original model containing all three explanatory

variables, construct a 99% confidence interval for the mean gas mileage for SUVs with Engine = 2,000, HP = 250 and Weight = 4,000.

1. Construct a 99% prediction interval for the mileage of a particular SUV

with Engine = 2,000, HP = 250 and Weight = 4,000.

**Team 4, 5, 6:**

1. A paper company is interested in making its operations more efficient. They collect data on the total manufacturing cost per month (in dollars), the total production of paper per month (in tons), the total number of machine hours per month, the total variable overhead cost per month (in thousands of dollars) and the total number of labor hours each month. The data can be found on WISE papercompany.txt .
   1. Fit a multiple regression using cost as the response variable, and the other four variables as explanatory variables. Write down the regression equation.
   2. Conduct the F-test for the overall fit of the regression. What conclusions can you draw?
   3. What proportion of the variation in cost has been explained by the regression?
   4. Perform a partial F-test to determine whether the variables associated with overhead and labor hours can be removed from the model. Comment on the results of the test. If so, remove the explanatory variable and rebuild the model.

**Team 7, 8:**

1. Researchers were interested in determining factors that affected the water consumption in the 48 contiguous states. For each state the researchers measured the per capita consumption of water (in gallons per day), the per capita income (in $1000), the average annual rainfall (in inches) and the average cost of 1000 gallons of water (in dollars). The data can be found on WISE Water.txt .
   1. Fit a multiple regression model using water consumption as the response variable and the other three variables as explanatory variables.
   2. What proportion of the variation in water consumption has been explained by the regression model?
   3. Conduct an F-test for the overall fit of the regression model. Comment on the results.
   4. Test each of the individual regression coefficients. Do the results indicate that any of the explanatory variables can be removed from the model? If so, remove the explanatory variable and rebuild the model.

**Deadline:** Mar 14 11:59PM.

**Deliverables and submission:** Each team will develop R code collaboratively and write a report to describe their findings. Each team will also need to make a powerpoint presentation to report their findings.

The team leaders need to upload the zipped R code, written report and powerpoint slides to WISE dropbox before the deadline. The zip file must have the homework number and team number in it. For example, “HW2\_Team\_1.zip”. During Mar 15 class, each team will give a 5-10 minute presentation about their findings.