**University of California, Davis**
**Department of Biomedical Engineering**

**Fall 2015** **Probability and Statistics** **BIM 105**
**David M. Rocke** **for Biomedical Engineers** **November 19, 2015**

## Homework Assignment 7

*Due December 3, 2015*

**Always show your work.**
**Use MATLAB for all problems except number 4.**

1. Ozone ($O_3$) is a major component of air pollution in many cities. Atmospheric ozone levels are influenced by many factors, including weather. In one study, the mean percent relative humidity ($x$) and the mean ozone levels ($y$) were measured for 120 days in a western city. Mean ozone levels were measured in ppb. The following output (from MATLAB) describes the fit of a linear model to these data. Assume that the required assumptions are satisfied sufficiently well that we can do the usual analysis.

```
>> fitlm(Humidity,Ozone)

Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate         SE             tStat      pValue
    (Intercept)    29.7240         2.051          14.490      0.0000
    x1             -0.1268       0.03825          -3.315      0.0006


Number of observations: 120, Error degrees of freedom: 118
Root Mean Squared Error: 6.26
R-squared: 0.960,  Adjusted R-Squared 0.890
```

   (a) What is the slope of the least-squares line?

   (b) Find a 95% confidence interval for the slope.

   (c) Perform a test of the null hypothesis that the slope is equal to $-0.1$. What is the P-value?

2. The article "Computation of Equilibrium Oxidation and Reduction Potentials for Reversible and Dissociative Electron-Transfer Reactions in Solution" (P. Winget, C. Cramer, and D. Truhlar, Theoretical Chemistry Accounts, 2004:217–227) presents several models for estimating aqueous one-electron potentials. The data set `HW7-2.csv` presents true potentials, measured experimentally in volts relative to the normal hydrogen electrode, for phenol and 23 substituted phenols, along with the corresponding value from the Austin model. Although the model values are not close to the true values, it is thought that they follow a linear model $y = \beta_0 + \beta_1 x + \epsilon$, where $y$ is the true value and $x$ is the model value. Using MATLAB for the computations when you can, do the following:

   (a) Compute the least-squares line for predicting the true potential from the model value.

   (b) Compute 95% confidence intervals for $\beta_0$ and $\beta_1$.

   (c) Two molecules differ in their model value by 0.5. By how much do you estimate that their true potentials will differ?

   (d) A molecule has a model value of 4.3. Find a 95% confidence interval for its true mean potential.

   (e) Can you conclude that the true mean potential of molecules whose model value is 4.5 is greater than 0.93? Explain.

3. Two radon detectors were placed in different locations in the basement of a home. Each provided an hourly measurement of the radon concentration, in units of pCi/L. The data are presented in the file `HW7-3`.

   (a) Compute the least-squares line for predicting the radon concentration at location 2 from the concentration at location 1.

   (b) Plot the residuals versus the fitted values. Does the linear model seem appropriate?

   (c) Divide the data into two groups: points where $R_1 < 4$ in group one, points where $R_1 \geq 4$ in group 2. Compute the least-squares line and the residual plot for each group. Does the line describe either group well? Which one?

   (d) Explain why it might be a good idea to fit a linear model to part of these data, and a nonlinear model to the other.

4. In an experiment to determine the factors affecting tensile strength in steel plates, the tensile strength (in kg/mm$^2$), the manganese content (in parts per thousand), and the thickness (in mm) were measured for a sample of 20 plates. The following MINITAB output presents the results of fitting the model Tensilestrength $= \beta_0 + \beta_1$Manganese $+ \beta_2$Thickness. (The MINITAB output is similar to what the MATLAB output would have been).

```
The  regression  equation  is
Strength  =  26.641  +  3.3201  Manganese  -  0.4249  Thickness

Predictor      Coef     StDev       T      P
Constant     26.641   2.72340    9.78  0.000
Manganese    3.3201   0.33198   10.00  0.000
Thickness   -0.4249   0.12606   -3.37  0.004

S  =  0.8228  R-Sq  =  86.2%  R-Sq(adj)  =  84.6%

Analysis  of  Variance
Source            DF     SS        MS       F      P
Regression         2   72.01   36.005   53.19  0.000
Residual  Error   17  11.508   0.6769
Total             19  83.517
```

(a) Predict the strength for a specimen that is 10 mm thick and contains 8.2 ppt manganese.

(b) If two specimens have the same thickness, and one contains 10 ppt more manganese, by how much would you predict their strengths to differ?

(c) If two specimens have the same proportion of manganese, and one is 5 mm thicker than the other, by how much would you predict their strengths to differ?

(d) Find a 95% confidence interval for the coefficient of Manganese.

(e) Find a 99% confidence interval for the coefficient of Thickness.

(f) Can you conclude that $\beta_1 \neq 3$? Perform the appropriate hypothesis test.

(g) Can you conclude that $\beta_2 \neq -0.1$? Perform the appropriate hypothesis test.

5. The article “Multiple Linear Regression for Lake Ice and Lake Temperature Characteristics” (S. Gao and H. Stefan, Journal of Cold Regions Engineering, 1999:59–77) presents data on maximum ice thickness in mm ($y$), average number of days per year of ice cover ($x_1$), average number of days the bottom temperature is lower than 8°C ($x_2$), and the average snow depth in mm ($x_3$) for 13 lakes in Minnesota. The data are given in the file `HW7-5.csv`

   (a) Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ For each coefficient, find the P-value for testing the null hypothesis that the coefficient is equal to 0.

   (b) If two lakes differ by 2 in the average number of days per year of ice cover, with other variables being equal, by how much would you expect their maximum ice thicknesses to differ?

   (c) Do lakes with greater average snow depth tend to have greater or lesser maximum ice thickness? Why might this be so?

6. The article “Vehicle-Arrival Characteristics at Urban Uncontrolled Intersections (V. Rengaraju and V. Rao, Journal of Transportation Engineering, 1995:317–323) presents data on traffic characteristics at 10 intersections in Madras, India. The file `HW7-6.csv` provides data on road width in m ($x_1$), traffic volume in vehicles per lane per hour ($x_2$), and median speed in km/h ($y$).

   (a) Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Find the P-values for testing that the coefficients are equal to 0. Use the `anova` command to obtain the analysis of variance table.

   (b) Fit the model $y = \beta_0 + \beta_1 x_1 + \epsilon$. Find the P-values for testing that the coefficients are equal to 0. Use the `anova` command to obtain the analysis of variance table.

   (c) Fit the model $y = \beta_0 + \beta_2 x_2 + \epsilon$. Find the P-values for testing that the coefficients are equal to 0. Use the `anova` command to obtain the analysis of variance table.

   (d) Which of the models (a) through (c) do you think is best? Why?

7. The article "Modeling Resilient Modulus and Temperature Correction for Saudi Roads" (H. Wahhab, I. Asi, and R. Ramadhan, Journal of Materials in Civil Engineering, 2001:298–305) describes a study designed to predict the resilient modulus of pavement from physical properties. The file `HW7-7.csv` presents data for the resilient modulus at 40 °C in $10^6$ kPa ($y$), the surface area of the aggregate in $\mathrm{m}^2/\mathrm{kg}$ ($x_1$), and the softening point of the asphalt in °C ($x_2$). The full quadratic model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$. Which submodel of this full model do you believe is most appropriate? Justify your answer by fitting two or more models and comparing the results. For each model that you fit, you should also examine the ANOVA table. For this problem, do not use stepwise regression.

8. The article "Seismic Hazard in Greece Based on Different Strong Ground Motion Parameters" (S. Koutrakis, G. Karakaisis, et al., Journal of Earthquake Engineering, 2002:75–109) presents a study of seismic events in Greece during the period 1978–1997. Of interest is the duration of "strong ground motion." For each event, measurements of the duration of strong ground motion were made at one or more locations. The file `HW7-8.csv` presents, for each of 121 such measurements, the data for the duration of time $y$ (in seconds) that the ground acceleration exceeded twice the acceleration due to gravity, the magnitude $m$ of the earthquake, the distance $d$ (in km) of the measurement from the epicenter, and two indicators of the soil type $s_1$ and $s_2$, defined as follows: $s_1 = 1$ if the soil consists of soft alluvial deposits, $s_1 = 0$ otherwise, and $s_2 = 1$ if the soil consists of tertiary or older rock, $s_2 = 0$ otherwise. Cases where both $s_1 = 0$ and $s_2 = 0$ correspond to intermediate soil conditions.

   Use these data to construct a linear model to predict duration $y$ from some or all of the variables $m$, $d$, $s_1$, and $s_2$. Be sure to consider transformations of the variables, as well as powers of and interactions between the independent variables. Describe the steps taken to construct your model. Plot the residuals versus the fitted values to verify that your model satisfies the necessary assumptions. In addition, note that the data are presented in chronological order, reading down the columns. Make a plot to determine whether time should be included as an independent variable.

   Earthquake magnitude on the Richter scale is already on a log scale. Ground shaking is known to decline with distance more than linearly. There are only three soil types, not four, because $s_1$ and $s_2$ cannot both be 1, so you may consider defining a new soil-type variable from these, but make sure that you create a factor with three levels, not a numerical variable.

   This is a long, difficult exercise, with more than one possible correct answer, and many incorrect ones. Describe your reasoning for each step and decision. After considering transformations and interactions, you may use stepwise regression.