Course EPIB-621 - Data Analysis for the Health Sciences

Assignment 2

1. Consider the data below (available as dosedat.txt on the course web site):

```
dose <- c(4, 4, 2, 8, 5, 5, 5, 6, 7, 5, 5, 5, 3, 4, 7, 5, 6, 4,
9, 7, 5, 5, 1, 8, 3, 5, 3, 2, 5, 6, 6, 9, 6, 2, 5, 3, 7, 4, 6,
3, 5, 4, 5, 2, 3, 6, 8, 6, 5, 5, 4, 1, 5, 6, 3, 6, 3, 4, 3, 4,
9, 2, 8, 4, 7, 9, 1, 5, 3, 5, 7, 7, 6, 5, 3, 8, 7, 5, 4, 8)

weight.gain<- c(8.1, 13.7, 1.9, 20.9, 11.7, 17.5, 19.9, 17.9, 23.6, 6.7,
20.2, 15.5, 5.2, 11.5, 16.9, 15.5, 9.9, 15.1, 21.6, 19.7, 16.8, 16.2,
3.7, 25.1, 11.7, 14.7, 18.6, 13.2, 18.8, 20, 17.8, 25.9, 13.9, 8.6,
11.8, 10.7, 21.3, 16.7, 18.8, 7.7, 16.6, 12.2, 15.5, 6, 9, 15.2, 26.5,
23, 10.7, 7.6, 13.7, 4, 16.7, 17.2, 12, 20, 14.1, 7.2, 7, 7.5, 25.7,
8.3, 23.7, 6.6, 18.7, 24.8, 8.2, 21.2, 6.9, 20.5, 14, 18.9, 16.6, 18,
12.3, 17.5, 20.7, 15.9, 9.3, 20.5)
```

The data come from an experiment including 80 subjects, each taking a drug that is supposed to increase weight. We will analyze the effects of the different dosages on the weights. The weight gains are in pounds, while the dosages are in milligrams. The subjects each took the drug for a period of one year.

Answer the following questions using R:

(a) Draw a scatter plot to visually examine the association between the dosage (x-axis) and weight gain (y-axis). Does there (visually) appear to be a relationship?

(b) State the regression line for these data, that is, provide the best values for the intercept (α) and slope (β) of the least squares (also maximum likelihood) line.

(c) State the estimate of the residual standard deviation, σ.

(d) Provide the 95% confidence intervals for the intercept and slope values you calculated in part (b).

(e) Suppose the next subject that enters the study is given a dosage of 5 mg. What

is your prediction for the weight gain for this (individual) subject? Provide the 95% confidence interval around this individual estimate.

(f) What is your prediction for the mean weight gain for a large group of subjects, all given a dosage of 5 mg? Provide the 95% confidence interval around this mean estimate.

(g) Suppose the exact dosage values are not available, but all we know are whether the dose was high ($> 5mg$) or low ($\leq 5mg$). Create a new variable based on dose, called dose.dichot that is equal to 0 for low dose subjects, and is equal to 1 for high dose subjects. Run a linear regression of weight.gain on this newly created variable. How do the point estimates of the slopes from the two different models compare? Can you explain any differences in the two sets of parameter estimates?

2. There is a data set called satisfaction.txt on the course web site. There are four variables in this data set, defined as follows:

(Y) satisfaction:	patient satisfaction with hospital services higher numbers indicate greater satisfaction
(X_1) age:	patient's age at hospital admission
(X_2) severity:	severity index, higher numbers are more severe cases
(X_3) anxiety:	anxiety index, higher numbers indicate more anxiety

(a) Create histograms of all four variables. Note the general features of each variable.

(b) Use the pairs function to look at scatter plots of all possible pairs of variables. Summarize your findings.

(c) Fit a linear regression for each variable separately. Report all parameter estimates with confidence intervals.

(d) Fit a multiple linear regression for all three variables. Report all parameter estimates with confidence intervals.

(e) Plot a histogram of the residuals from the model with all three variables included. Does it look like any assumptions are being violated?

(f) Create a scatter plot of the residuals against each of the X variables (so three plots). Comment on what these plots indicate.

3. There is a data set called assign2num3.txt on the course web site. There are

five variables in this data set, simply called `x1, x2, x3, x4`, and `y`. The `x`'s are all independent variables, while `y` is the outcome.

(a) Create histograms of all five variables. Note the general features of each variable.

(b) Create a correlation matrix of all five variables. Summarize your findings.

(c) Fit a linear regression for each variable separately. Report all parameter estimates with confidence intervals.

(d) Fit a multiple linear regression for all four independent variables. Report all parameter estimates with confidence intervals.

(e) Compare the simple linear regression (univariate) results from part (c) to the multivariate results in part (d). Summarize your findings.

4. Consider the kidney data set on the course web site. The variables in that data set are defined as follows:

(Y) creatinine clearance:	a measure of kidney fuinction
(X_1) creatinine concentration:	more easily measured than clearance
(X_2) age:	patient's age in years
(X_3) weight:	weight in Kg

Creatinine clearance is an important measure of kidney function that is difficult to measure, as it requires 24 hour urine collection. We would like to see if creatinine clearance can be predicted from creatinine concentration, age and weight.

(a) Create histograms of all four variables. Note the general features of each variable.

(b) Use the pairs function to examine the univariate relationships between all pairs of variables. Note the main feature(s) for each pair.

(c) Fit a linear regression for each variable separately. Report all parameter estimates with confidence intervals.

(d) Fit a multiple linear regression for all three independent variables. Report all parameter estimates with confidence intervals.

(e) Examine the residuals from the model, both via a histogram, and via scatter

plots of each independent variable against the residual.

(f) Make a prediction of creatinine clearance for an individual aged 50 years old, with weight 80 Kg, and with creatinine concentration of 1.00. Report both the prediction and the confidence interval for the prediction.

(g) Overall, what can you conclude about the ability of the three independent variables to predict the outcome?

5. Consider the data set plasma.txt on the course web site. The data consist of plasma levels of a polyamine (plasma variable Y), against age in children (X variable, age = 0 indicate a new born).

(a) Create histograms for both variables. Note the general features of each variable.

(b) Create a scatter plot of age versus plasma. Does the relationship seem linear?

(c) Transform the Y variable with a log transform. In other words, rather than Y, create a $\log(Y) = \log(\text{plasma})$. The logarithm should be to the base e. Re-create the scatter plot, but now plotting age versus log(plasma). Does the relationship now seem more linear?

(d) Fit a linear regression for age versus log(plasma). Report all parameter estimates with confidence intervals.

(e) Provide an interpretation of the β coefficient calculated in (d).

(f) As a child ages from 3 to 4 years old, on average, by how much does their plasma (not log(plasma)) change?

6. Consider the data set called assign2num6.txt. This data set consists of a y outcome variable and a matrix x with 5 columns, each column representing a potential predictor variable for y. Read the data into R (either by using the source command, or just cut and paste from the web site directly into R).

(a) Create a scatter plot for each, using commands such as

```
plot(x[,1], y)
```

Do you see any relationships between y and any of the columns of x?

(b) Now run a linear regression with outcome y, using all five variables in the matrix x, i.e., using a command such as

```
a<-lm(y ~ x[,1] + x[,2] + x[,3] + x[,4] +x[,5] )
```

No need to report the results here, this will be used in part (c).

(c) Now plot the residuals of this model against the fitted values, using your model in (b). Do you see any pattern? You can use a command such as:

```
plot(a$fitted.values, a$residuals)
```