Stat 511 HW#3 Spring 2009 (Not to be Collected, but Covered on Exam 1)

1. On the course Web page you will find the file `homes.TXT` We're going to do some statistical analysis on this set of home sale price data obtained from the Ames City Assessor's Office. Data on sales May 2002 through June 2003 of $1\frac{1}{2}$ and 2 story homes built 1945 and before, with (above grade) size of 2500 sq ft or less and lot size 20,000 sq ft or less, located in Low- and Medium-Density Residential zoning areas are given in this file. $n = 88$ different homes fitting this description were sold in Ames during this period. (2 were actually sold twice, but only the second sales prices of these were included in our data set.) For each home, the value of the response variable

Price = recorded sales price of the home

and the values of 14 potential explanatory variables were obtained. These variables are

- *Size*, the floor area of the home above grade in sq ft
- *Land*, the area of the lot the home occupies in sq ft
- *Bed Rooms*, a count of the number in the home
- *Central Air*, a dummy variable that is 1 if the home has central air conditioning and is 0 if it does not
- *Fireplace*, a count of the number in the home
- *Full Bath*, a count of the number of full bathrooms above grade
- *Half Bath*, a count of the number of half bathrooms above grade
- *Basement*, the floor area of the home's basement (including both finished and unfinished parts) in sq ft
- *Finished Bsmt*, the area of any finished part of the home's basement in sq ft
- *Bsmt Bath*, a dummy variable that is 1 if there is a bathroom of any sort (full or half) in the home's basement and is 0 otherwise
- *Garage*, a dummy variable that is 1 if the home has a garage of any sort and is 0 otherwise
- *Multiple Car*, a dummy variable that is 1 if the home has a garage that holds more than one vehicle and is 0 otherwise
- *Style* (*2 Story*), a dummy variable that is 1 if the home is a 2 story (or a $2\frac{1}{2}$ story) home and is 0 otherwise
- *Zone* (*Town Center*), a dummy variable that is 1 if the home is in an area zoned as "Urban Core Medium Density" and 0 otherwise

The first row of the file has the variable names in it. (You might open this file by double clicking on the link to have a look at it.)

While connected to the network, enter these data into `R` using the command

```
> homes<-
read.table("http://www.public.iastate.edu/~vardeman/stat511/homes.TX
T",header=T)
```

In theory, one should also be able to get this loaded by placing `homes.TXT` into an appropriate directory of the local machine (where `R` knows to search) and issuing the above command with `homes.TXT` only (instead of the URL). I have on occasion made this work and other times have not been able to do so.

Use the command

```
> homes
```

to view the data frame. It should have 15 columns and 88 rows. Now create two matrices that will be used to fit a regression model to some of these data. Type

```
> Y<-as.matrix(homes[,1])
> X<-as.matrix(homes[,c(2,5,10,11,13)])
```

Note the use of `[]`to select columns from the data frame. Here, the function `as.matrix` is used to create a matrix from one or more columns of the data frame. To add a column of ones to the model matrix, type

```
> X0<-rep(1,length(Y))
> X<-cbind(X0,X)
```

Make a scatterplot matrix for $y, x_1, x_2, \ldots, x_5$. To do this, first load the `lattice` package. (Look under the "Packages" heading on the `R` GUI, select "Load package" and then `lattice`.) Then type

```
> splom(~homes[,c(1,2,5,10,11,13)],aspect="fill")
```

If you had to guess based on this plot, which single predictor do you think is probably the best predictor of Price? Do you see any evidence of multicollinearity (correlation among the predictors) in this graphic?

Also compute a sample correlation matrix for y, x_1, x_2, x_3, x_4 and x_5. You may compute the matrix using the `cor()` function and round the printed values to four places using the `round()` function as

```
> round(cor(homes[,c(1,2,5,10,11,13)]),4)
```

Use the `qr()` function to find the rank of $\mathbf{X}$.

Use `R` matrix operations on the $\mathbf{X}$ matrix and $\mathbf{Y}$ vector to find the estimated regression coefficient vector $\mathbf{b}_{OLS}$, the estimated mean vector $\hat{\mathbf{Y}}$, and the vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.

Plot the residuals against the fitted means. After loading the `MASS` package, this can be done using the following code.

```
> b<-solve(t(X)%*%X)%*%t(X)%*%Y
```

```
> yhat<-X%*%b
> e<-Y-yhat
> par(fin=c(6.0,6.0),pch=18,cex=1.5,mar=c(5,5,4,2))
> plot(yhat,e,xlab="Predicted Y",ylab="Residual",main="Residual Plot")
```

Type `> help(par)` to see the list of parameters that may be set on a graphic. What does the first specification above do, i.e. what does `fin=c(6.0,6.0)` do?

Plot the residuals against home size. You may use the following code.

```
> plot(homes$Size,e,xlab="Size",ylab="Residual",main="Residual
Plot")
```

And you can add a smooth trend line to the plot by typing

```
> lines(loess.smooth(homes$Size,e,0.90))
```

What happens when you type

```
> lines(loess.smooth(homes$Size,e,0.50))
```

(The values 0.90 and 0.50 are values of a "smoothing parameter." You could have discovered this (and more) about the `loess.smooth` function by typing `> help(loess.smooth)`)

Now plot the residuals against each of x_2, x_3, x_4 and x_5.

Create a normal plot from the values in the residual vector. You can do so by typing

```
> qqnorm(e,main="Normal Probability Plot")
> qqline(e)
```

Now compute the sum of squared residuals and the corresponding estimate of σ^2, namely

$$\widehat{\sigma^2} = \frac{\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)'\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)}{n - \text{rank}\left(\mathbf{X}\right)}$$

Use this and compute an estimate of the covariance matrix for $\mathbf{b}_{\text{OLS}}$, namely

$$\widehat{\sigma^2}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Sometimes you may want to write a summary matrix out to a file. This can be done as follows. First prepare the row and columns labels and round all entries to 4 places using the code

```
> case<-1:88
> temp<-cbind(case,homes[,c(2,5,10,11,13)],Y,yhat,e)
> round(temp,4)
```

2. Continue with the homes data from Problem 1. Consider a regression of y on $x_1, x_2, \ldots x_5$. Use R matrix calculations to do the following in a full rank Gauss-Markov normal linear model.

a) Find 90% two-sided confidence limits for σ.

b) Find 90% two-sided confidence limits for the mean response under the conditions of data point #1.

c) Find 90% two-sided confidence limits for the difference in mean responses under the conditions of data points #1 and #2.

d) Find a *p*-value for testing the hypothesis that the conditions of data points #1 and #2 produce the same mean response.

e) Find 90% two-sided prediction limits for an additional response for the set of conditions $x_1 = 1500, x_2 = 3, x_3 = 1000, x_4 = 500, x_5 = 8000$.

f) Find 90% prediction limits for the difference in two additional responses under the two sets of conditions $x_1 = 1500, x_2 = 3, x_3 = 1000, x_4 = 500, x_5 = 8000$ and $x_1 = 1800, x_2 = 3, x_3 = 1000, x_4 = 1000, x_5 = 10000$.

g) Find a *p*-value for testing the hypothesis that a model including only x_1, x_3 and x_5 is adequate for "explaining" home price.

3. In the context of Problem 2, part g), suppose that in fact $\tau_1 = \tau_2, \tau_3 = \tau_4 = \tau_1 - d\sigma$. What is the distribution of the F statistic? Use R to plot the power of an $\alpha = .05$ level test as a function of d for $d \in [-5, 5]$ (that is, plot $P[F >$ the cut-off value$]$ against d). The `R` function `pf` will compute cumulative (non-central) F probabilities for you. The call `pf(q,df1,df2,ncp)` returns the cumulative non-central F probability corresponding to the value `q`, for degrees of freedom `df1` and `df2` when the non-centrality parameter is `ncp`.

5. Use the `R` function `dchisq(x,df,ncp)` and plot on the same set of axes the chi-square probability density functions for 3 degrees of freedom and non-centrality parameters $0, 1, 3, 5$.