Math 484   HW7 Fall 2021, due Friday, Oct. 21.

Quiz 7 on Oct. 19 covers predictor and response transformations, scatterplot matrices, variable selection and models with interactions, factors and powers (HW6 and HW7).

5 pages, problems A)–G)

A) 3.6 Download *cbrainx*, *cbrainy* and *lregpack.txt* into *R*: copy and paste the two source commands at the top of (http://parker.ad.siu.edu/Olive/lreghw.txt) into *R*. Copy and paste the *R* commands for this problem from this URL, too.

The data is the brain weight data from Gladstone (1905). The response Y is *brain weight* while the predictors are *age, breadth, cephalic, circum, headht, height, len, sex* and a constant. The *step* function can be used to perform forward selection and backward elimination in *R*.

a) Copy and paste the commands for this problem into *R*. The commands fit the full model, display the LS output and perform backward elimination using the AIC criterion. Copy and paste the output for backward elimination into *Word* (one page of output). Some commands are shown below.

```
zx <- cbrainx[,c(1,3,5,6,7,8,9,10)]
zbrain <- as.data.frame(cbind(cbrainy,zx))
zfull <- lm(cbrainy~.,data=zbrain)
summary(zfull)
back <- step(zfull)
```

b) Want low AIC and as few predictors as possible. Backward elimination starts with the full model then deletes one nontrivial predictor at a time. The term <None> corresponds to the current model that does not eliminate any terms. The terms listed above <None> correspond to models that have smaller AIC than the current model. *R* stops when eliminating terms makes the AIC higher than the current model. Which terms, including a constant, were in this minimum AIC model?

c) Copy and paste the commands for this problem into *R*. The commands fit the null model that only contains a constant. Forward selection starts at the null model (corresponding to lower) and considers 8 nontrivial predictors (given by upper).

Copy and paste the output for forward selection into *Word* (two pages of output). Some commands are shown below.

```
zint <- lm(cbrainy~1,data=zbrain)
forw <- step(zint,scope=list(lower=~1,
upper=~age+breadth+cephalic+circum+headht+height+len+sex),
direction="forward")
```

d) Forward selection in *R* starts with the null model and then adds a predictor *circum* to the model. Forward selection in *R* allows you to consider models with fewer predictors than the minimum AIC model (unlike backward elimination). Which terms, including a constant, were in the minimum AIC model?

B) 3.1

```
Output for problem 3.1. Current terms:
(finger to ground nasal height sternal height)
                          df  RSS       |  k    C_I
Delete: nasal height      73  35567.2   |  3    1.617
Delete: finger to ground  73  36878.8   |  3    4.258
Delete: sternal height    73  186259.   |  3  305.047
```

3.1. From the above output from backward elimination, what terms should be used in the MLR model to predict Y? (You can tell that the nontrivial variables are finger to ground, nasal height, and sternal height from the "delete lines." DON'T FORGET THE CONSTANT!)

C) 3.2 The table below gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The response plot and residual plot for the full model L1 was good. Model L3 was the minimum C_p model found. Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Output for Problem 3.2.

	L1	L2	L3	L4
# of predictors	10	6	4	3
# with $0.01 \leq$ p-value ≤ 0.05	0	0	0	0
# with p-value > 0.05	6	2	0	0
$R^2(I)$	0.774	0.768	0.747	0.615
$\text{corr}(\hat{Y}, \hat{Y}_I)$	1.0	0.996	0.982	0.891
$C_p(I)$	10.0	3.00	2.43	22.037
$\sqrt{MSE}$	63.430	61.064	62.261	75.921
p-value for partial F test	1.0	0.902	0.622	0.004

D) 3.4 b) The output below is from software that does all subsets variable selection. The data is from Ashworth (1842). The predictors were A = log(1692 property value), B = log(1841 property value), and C = log(percent increase in value), while the response variable is Y = log(1841 population).

b) The (bottom) output corresponds to the data with the 2 outliers removed. From this output, what is the best model? Explain briefly.

```
(bottom) Output for Problem 3.4.

              ADJ  97 cases after deleting 2 outliers
 k      CP    R SQ   R SQ    RESID SS    VARIABLES
--   -----    ----   ------  --------   -------------
 1  903.5  0.0000   0.0000  183.102   INTERCEPT ONLY
 2    0.7  0.9052   0.9062  17.1785   B
 2  406.6  0.4944   0.4996  91.6174   A
 2  426.0  0.4748   0.4802  95.1708   C
 3    2.1  0.9048   0.9068  17.0741   A C
 3    2.6  0.9043   0.9063  17.1654   B C
 3    2.6  0.9042   0.9062  17.1678   A B
 4    4.0  0.9039   0.9069  17.0539   A B C
```

E) 3.13 Activate the *cyp.lsp* data set. Choosing no more than 3 nonconstant terms, try to predict *height* with multiple linear regression. Include a plot with the fitted values on the horizontal axis and height on the vertical axis. Is your model linear? Also include a plot with the fitted values on the horizontal axis and the residuals on the vertical axis. Does the residual plot suggest that the linear model may be inappropriate? (There may be outliers in the plot. These could be due to typos or because the error distribution has heavier tails than the normal distribution.) **State which model you use.**

F) 3.11 a) In this problem we want to build a MLR model to predict $Y = t(BigMac)$ where t is some power transformation. In *Arc* enter the menu commands “File>Load>Data” and open the file *big-mac.lsp.* Make a scatterplot matrix of the variables, except “City”, and include the plot in *Word.*

For part a) include the variables in order BigMac, Bread, ..., WorkHrs. Do not include City. This will put BigMac on the bottom of the scatterplot matrix.

b) The log rule makes sense for the BigMac data. From the scatterplot matrix, use the “Transformations” menu and select “Transform to logs”. Include the resulting scatterplot matrix in *Word.*

c) From the “Mac” menu, select “Transform”. Then select all 10 variables and click on the “Log transformations” button. Then click on “OK”. From the “Graph&Fit” menu, select “Fit linear LS.” Use log[BigMac] as the response and the other 9 “log variables” as the Terms. This model is the full model. Include the output in *Word.*

(As a shortcut to c), there is a *Transformations* menu on the scatterplot matrix. Select *Add transformed variables to data set.*)

d) Make a response plot (L1:Fit-Values in H and log(BigMac) in V) and residual plot (L1:Fit-Values in H and L1:Residuals in V), and include both plots in *Word*.

e) Using the "L1" menu, select "Examine submodels" and try forward selection and backward elimination. Using the $C_p \leq \min(2k, p)$ rule suggests that the submodel using log[service], log[TeachSal], and log[TeachTax] may be good. From the "Graph&Fit" menu, select "Fit linear LS", fit the submodel and include the output in *Word*.

f) Make a response plot (L2:Fit-Values in H and log(BigMac) in V) and residual plot (L2:Fit-Values in H and L2:Residuals in V) for the submodel, and include the plots in *Word*.

g) Make an RR plot (L2:Residuals in H and L1:Residuals in V) and FF plot (L2:Fit-Values in H and L1:Fit-Values in V) for the submodel, and include the plots in *Word*. Move the OLS slider bar to 1 in each plot to add the identity line. For the RR plot, click on the *Options menu* then type $y = x$ in the long horizontal box near the bottom of the window and click on OK to add the identity line.

h) Do the plots and output suggest that the submodel is good? Explain.

Comment that could be ignored:

One of the main goals of this course is for you to be able to find good final submodel for the data. You will be asked to do this on HW9 A and HW10 A. Problem F) illustrates the procedure.

i) Make a scatterplot matrix of the predictors and response with the response on either the top or bottom row. If there are more than 9 nontrivial predictors, use the first few (up to 9) predictors and the response to make a scatterplot matrix, then the next few predictors and the response, and continue until all predictors have been used. Each scatterplot matrix should include the response on the top or bottom row.

ii) Use the log rule and other power transformations to remove strong nonlinearities from the predictors and response. It is also often useful to use the log rule to transform highly skewed predictors.

iii) The row of the scatterplot matrix with the response should primarily have linear marginal plots.

iv) Use the transformed variables from parts ii) and iii) to build an initial full model. **Check that the full model is reasonable by making the response and residual plots.** *In general, the full model need not use every predictor (possibly transformed).*

v) Assuming that all predictor variables are equally important, use forward selection and backward elimination and rules like $C_p(I) \leq \min(2k, p)$ to find a good submodel that contains no more predictors than the minimum C_p submodel I_{min}.

vi) Check the submodel with response, residual, RR and FF plots. From the output for t tests, see if any predictors have pval > 0.05, if so, try deleting the predictor with the largest pval. Repeat until no more predictors can be deleted.

vii) Check that the submodel I has $R^2(I)$ and MSE(I) close to that of the full model. The partial F test using I as the reduced model should have pval ≥ 0.01 (not the usual 0.05). If not, add or delete predictors until I seems to be good.

viii) **Check that the final submodel I is good by making response, residual, RR and FF plots.**

G) Consider the following data set: votes for preseason 1A basketball poll Nov. 22, 2011, WSIL News.

```
111     89     778     78     76
```

Find shorth(3).

Hint: sort the data from smallest to largest and follow example done in class.