Department of Statistics

COURSE STATS 330

Model answers for Assignment 3, 2007

.

Question 1

(a) Read the data into R, checking the data as usual. The data are in the file ***cows.txt*** *is on the course web page. [5 marks]*
There was one typo in the data: the weight for observation 20 was recorded as 1145 rather than 11.45. This should be corrected in the data file cows.txt and the data read in again.

(b) You are required to fit a suitable model to these data, taking note of the following points:

- *Are all the variables required in the regression? Use variable selection techniques to choose a suitable subset if not.*

First, lets fit the full model:

```
> cows.reg = lm(Price ~ ., data=cows.df)
> summary(cows.reg)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  881.638    149.227   5.908 4.25e-08 ***
Age          -38.244      6.713  -5.697 1.10e-07 ***
Bred          24.013     22.167   1.083    0.281
Angus         13.288     20.152   0.659    0.511
Frame         29.135     20.517   1.420    0.159
Weight        24.117     14.565   1.656    0.101
Cond          33.299     24.951   1.335    0.185
Reg           28.776     21.374   1.346    0.181
---
Residual standard error: 101.9 on 106 degrees of freedom
Multiple R-Squared: 0.3899,     Adjusted R-squared: 0.3497
F-statistic: 9.679 on 7 and 106 DF,  p-value: 2.871e-09
```

This doesn't look too promising – the R^2 is low and it looks as though age might be the only significant variable. Lets do a stepwise regression:

```
cows.reg = lm(Price ~ ., data=cows.df)
null.reg = lm(Price~1, cows.df)
step(null.reg , scope=Price ~ Age + Bred+Angus+Frame+Weight+Cond+Reg,
direction = "both")
```

This results in the model `Price ~ Age + Weight + Frame + Cond.`

An all possible regressions analysis gives

```
> all.poss.regs(cows.reg)
     rssp   sigma2 adjRsq     Cp     AIC     BIC       CV Age Bred Angus Frame Weight Cond Reg
1 1309079 11688.20  0.268 16.039 130.039 135.511 128093.9   1    0     0     0      0    0   0
2 1194839 10764.31  0.326  7.040 121.040 129.249 119844.8   1    0     0     0      1    0   0
3 1163097 10573.61  0.338  5.984 119.984 130.929 119452.1   1    0     0     1      0    1   0
4 1135896 10421.06  0.347  5.365 119.365 133.046 118944.3   1    0     0     1      1    1   0
5 1117793 10349.93  0.352  5.622 119.622 136.039 120041.1   1    0     0     1      1    1   1
6 1105463 10331.43  0.353  6.435 120.435 139.588 119274.7   1    1     0     1      1    1   1
7 1100947 10386.30  0.350  8.000 122.000 143.890 121613.1   1    1     1     1      1    1   1
```

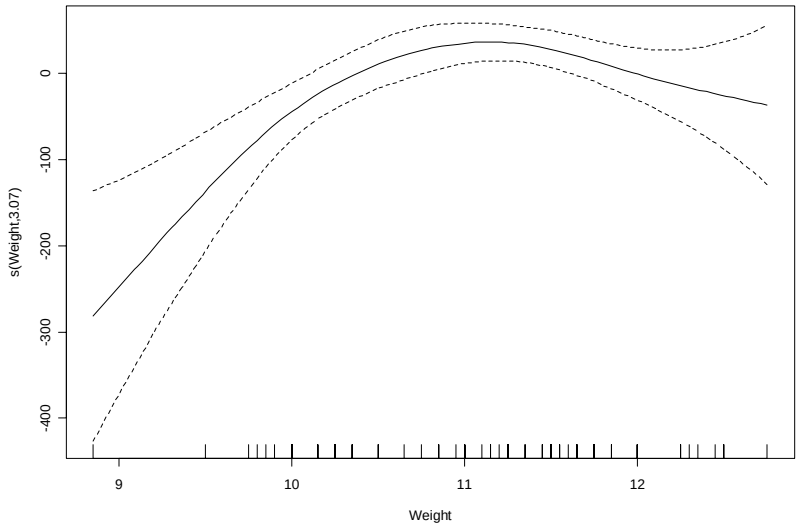
The model suggested by stepwise regressions is also a good model on the CV and AIC criteria. We will go with this one.

Is the relationship between the response and the explanatory variables linear? If not, can a suitable transformation be made?

If we refit the model and look at residual plots we don't get a very strong impression of any curvature in the regression surface. There don't seem to be any high leverage points and the residuals are of reasonable size. The normal plot is a bit curved.

A gam plot suggests that transforming Weight might be a good idea: note that we can't do gam plots for Age, Cond or Frame since they have too few values. We can do one for Weight:

```
> plot(gam(Price~s(Weight) + Age + Frame + Cond, data=cows.df))
```



Lets try a quadratic in weight:

```
> quad.reg =lm(formula = Price ~ Age + Frame + Weight + I(Weight^2)+
Cond, data = cows.df)
> summary(quad.reg)
Call:
lm(formula = Price ~ Age + Frame + Weight + I(Weight^2) + Cond,
```

```
    data = cows.df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6085.688   1589.676  -3.828 0.000217 ***
Age           -39.442      5.877  -6.711 9.27e-10 ***
Frame          17.505     19.079   0.918 0.360913
Weight       1304.925    290.004   4.500 1.72e-05 ***
I(Weight^2)   -58.135     13.140  -4.424 2.32e-05 ***
Cond           23.800     22.366   1.064 0.289655

Residual standard error: 94.36 on 108 degrees of freedom
Multiple R-Squared: 0.4672,     Adjusted R-squared: 0.4425
F-statistic: 18.94 on 5 and 108 DF,  p-value: 1.754e-13
```

This is clearly better, the coefficient of Weight^2 is significant and the R^2 is improved to nearly 47%.

- *Are there any outliers or influential points in the subset?*

However, influence plots suggest that points 17 and 92 might be having an undue influence on the regression. We refit with these points deleted, but the results are similar.

- *Is the normality assumption satisfied?*

The normal plot is not very straight. A boxcox plot suggeststransfroming with a cube, but this does not improve the fit. We decided to ignore the lack of normality.

- *Is the equal variance assumption satisfied?*

From the residuals versus fitted value plot, seems OK.

We also repeated the variable selection including Weight^2 as a variable. This time the stepwise regression came up with the model

```
Price ~ Age + Weight + I(Weight^2) + Angus
```

Again, points 17 and 92 where influential.

It seems that there are 4 models under consideration

```
A: Price ~ Age + Frame + Weight + I(Weight^2) + Cond, 17, 92 included
B: Price ~ Age + Frame + Weight + I(Weight^2) + Cond, 17, 92 excluded
C: Price ~ Age + Weight + I(Weight^2) + Angus, 17, 92 included
D: Price ~ Age + Weight + I(Weight^2) + Angus, 17, 92 excluded
```

The R2's are 46.7%, 46.2%, 46.9% and 46.1% respectively. There is not much between these.

(c) *Use your final model to predict the price of a 3 year old cow weighing 1100 pounds with the following characteristics: Bred =Yes, Angus=Yes, Frame=No, Cond=Yes, Reg=No. Express your prediction in the form of a 90% prediction interval.*

Predictions based on models A,B,C,D above give the following results

```
predict.at = data.frame(Age=3, Weight=11, Cond=1, Frame=0, Angus=1)
p1=predict(quad.reg, newdata=predict.at, interval="p")
p2=predict(quad2.reg, newdata=predict.at, interval="p")
p3=predict(angus.reg, newdata=predict.at, interval="p")
p4=predict(angus2.reg, newdata=predict.at, interval="p")
result=rbind(p1,p2,p3,p4)

row.names(result)=LETTERS[1:4]
result
```

```
       fit        lwr      upr
A   1139.672   949.4546 1329.889
B   1143.482   954.9154 1332.049
C   1155.830   967.7321 1343.928
D   1159.156   972.3404 1345.973
```

In actual fact, the cow's weight was 1185 pounds, so all these methods under-predicted.

Question 2.

a) Repeat the process 10000 times, saving the result each time. See Tutorial 4 for some ideas on how to do this. [5 marks]

The following code calculates 10000 values of the maximum absolute residual:

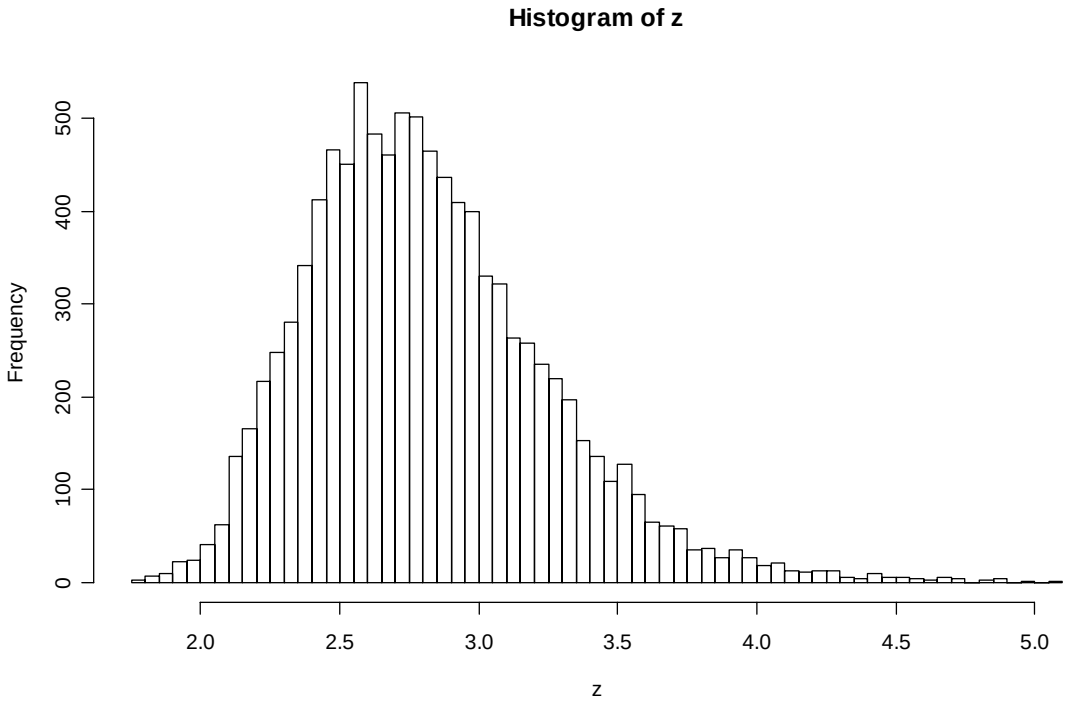
```
n=100
k=5
N = 10000
z=numeric(N)
for(i in 1:N)z[i] = max(abs(rt(n,n-k-1)))
# rt draws random samples from a t distribution
```

b) Summarize the results. What do you conclude? What magnitude would you expect the largest residual to have? [5marks]

Let's draw a histogram:

```
hist(z, nclass=50) # 50 bins in histogram
```

From the histogram overleaf, we expect that the largest absolute externally studentised residual in samples of 100 is between about 2 and 4. The 2.5 % and 97.5% quantiles are about 2.1 and 3.8 (`sort(z)[250]= 2.130345`, `sort(z)[9750]=3.810094`) so with about 95% probability the maximum absolute residual will be between 2.1 and 3.8.



c) *Would you be surprised if you got a maximum externally studentized residual of 3.0 in a regression with 100 observations? Would this indicate any problems with the regression? [5marks]*

```
> mean(z>3.0) # proportion of results less than 3.0
[1] 0.2918
```

Thus, about 29% of the time, we would get a result bigger than 3.0. Clearly, this is not an extreme result, and doesn't indicate anything wrong. We shouldn't be surprised.

Question for bonus marks: we have ignored the correlation between the residuals in our simulation approach above. Given a specific regression, can you suggest how we could modify our approach to take these correlations into account?

We could fit the regression. Then, assuming that the fitted model was the correct one, we could generate data from the model and calculate the studentized residuals instead of merely drawing from a t-distribution. [5 marks]