Department of Statistics

COURSE STATS 330

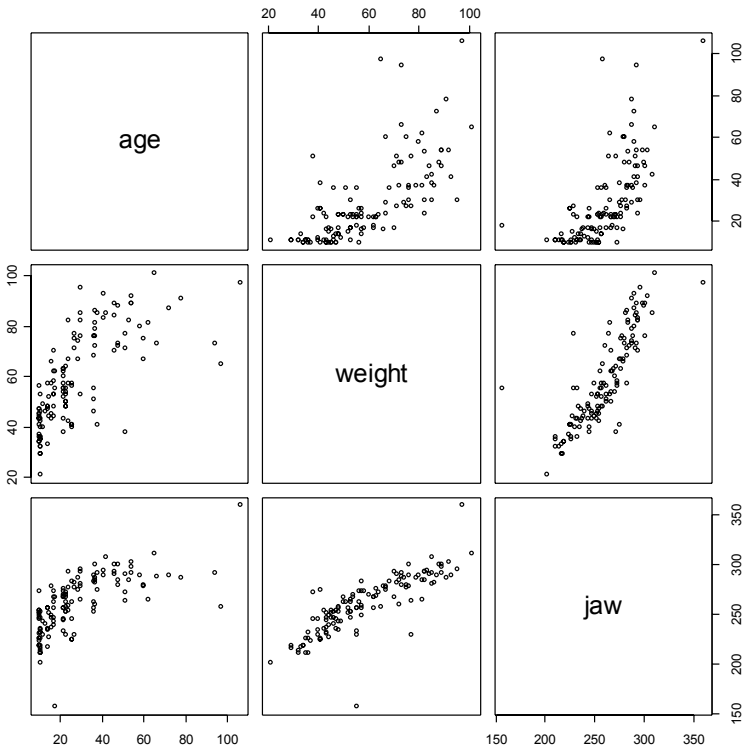
Assignment 2, 2003 Model Answer

1. *Fit a regression model to the data, using age as the response. Then, having fitted the model, examine the fit for*
   - *Non-planar regression*
   - *Non-constant variance*
   - *Outliers and high-leverage points*
   - *Lack of normality*

   Make a list of the defects in the fit that you have found. Show any plots used, together with the code used to produce them.

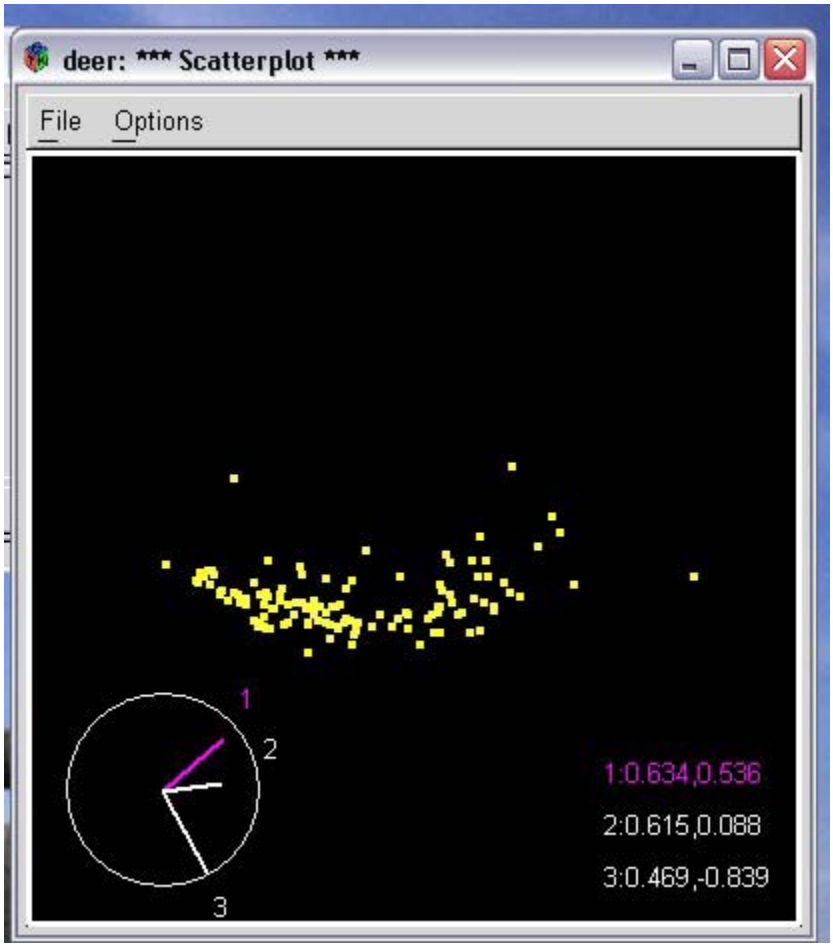
   The data were read in and a preliminary look taken using pairs plots, and ggobi (particularly useful since there are only 2 explanatory variables).

```
> deer.df <- read.table(
"http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/deer.txt",
header=T)

> pairs(deer.df)
```



From the pairs plot, there seems to be a reasonably strong relationship between both explanatory variables and the response. The relationship between age and weight seems reasonably linear, but the relationship with jaw is curved, indicating that jaw will possibly require transformation.

Spinning the data gives the same impression:



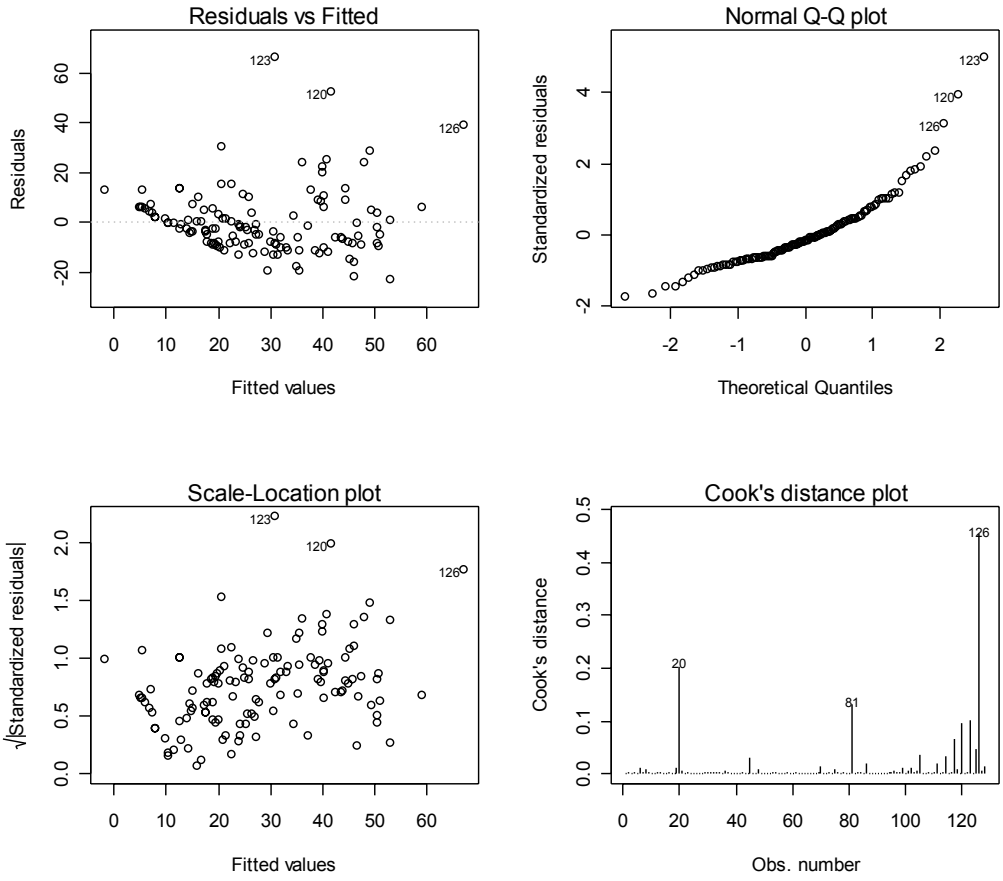
Now we fit the model and move on to the analysis of residuals:

```
> deer.lm<-lm(age~ weight + jaw, data=deer.df)

> summary(deer.lm)

Call:
lm(formula = age ~ weight + jaw, data = deer.df)

Residuals:
    Min      1Q  Median      3Q     Max
-23.082  -8.870  -2.549   5.206  66.151
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.1324    14.5380  -3.655 0.000378 ***
weight        0.4829     0.1181   4.088 7.73e-05 ***
jaw           0.2039     0.0760   2.682 0.008301 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 13.39 on 125 degrees of freedom
Multiple R-Squared: 0.5273,     Adjusted R-squared: 0.5197
F-statistic: 69.72 on 2 and 125 DF,  p-value: < 2.2e-16
```

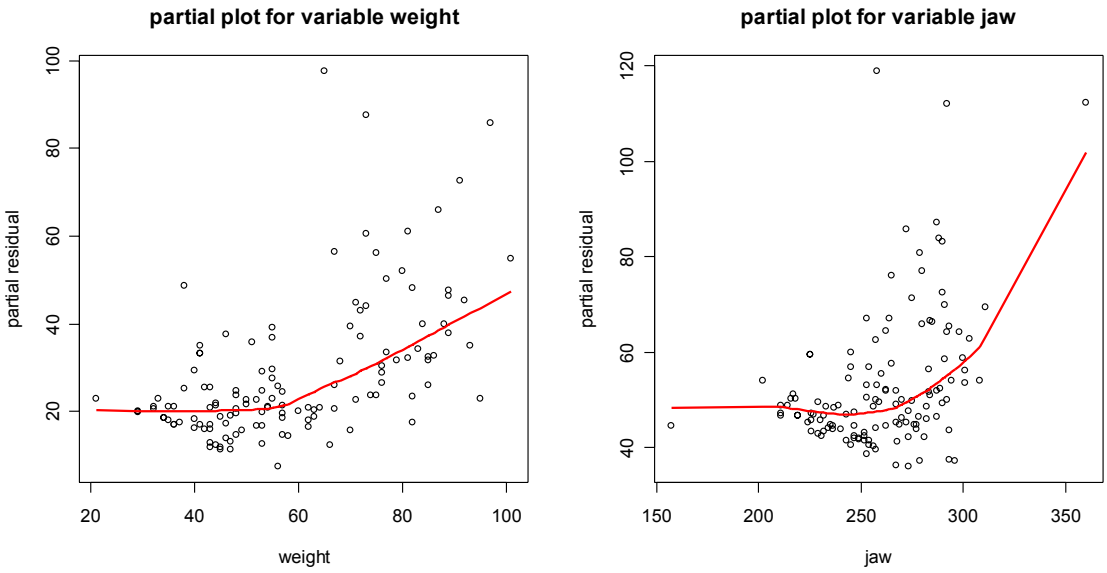
The R^2 is not great, and the error variance is quite large compared to the range of the y-data, but at least both explanatory variables are significant and should be retained. Lets do some plots to check the non-linearity. First do the basic plot:

> plot(deer.lm)



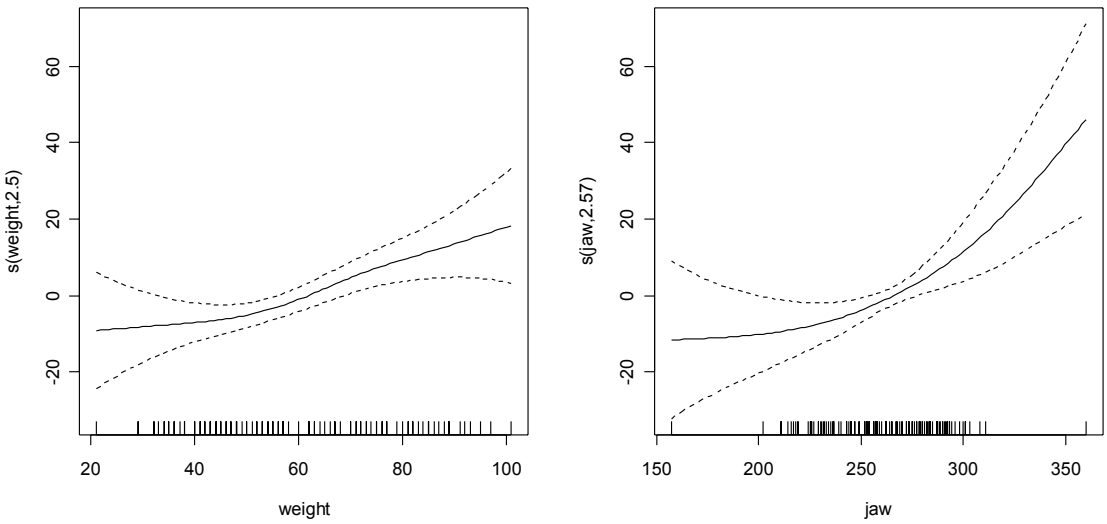
Next some partial residual plots

> partial.res.plot(deer.lm)



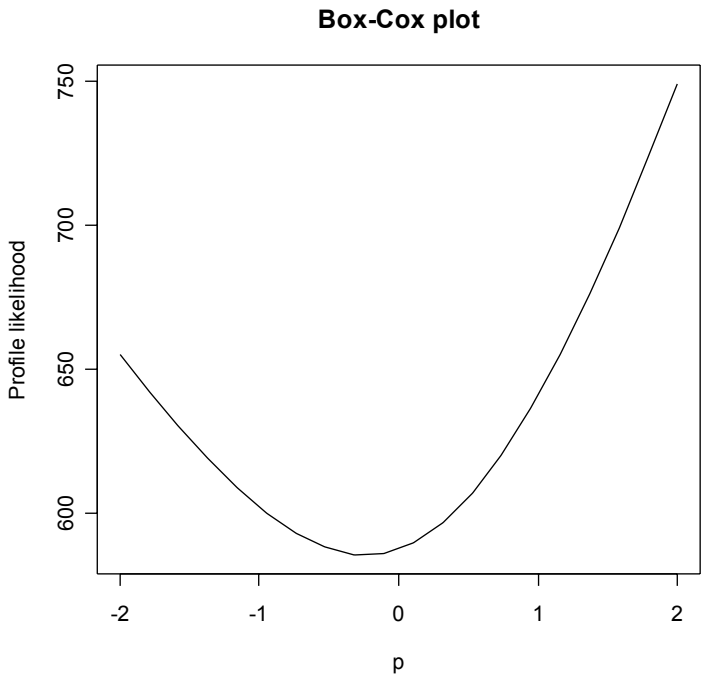
Next some gam plots

```
> library(mgcv)
> plot(gam(age~s(weight) + s(jaw),data=deer.df))
```



Some indication that both variables need transforming. May be worth transforming the response instead: a Box-Cox plot suggests a reciprocal fourth root:

```
> boxcoxplot(deer.df)
```



Thus, a strong indication that the data are not planar.

From the residuals/fitted values, there is also a hint of a funnel effect.

The normal plot of the residuals suggests non-normality as well. The Weisberg-Bingham test confirms this:

```
> WB.test(deer.lm)

WB test statistic =  0.931
p =  0
```

Finally, the Cook's distance plot suggests that points 20, 81 and 125 might be influential.

2. Find a suitable transformation that will cure (or at least partially cure) the defects you listed in 1. Document the reasoning that led you to your transformation.

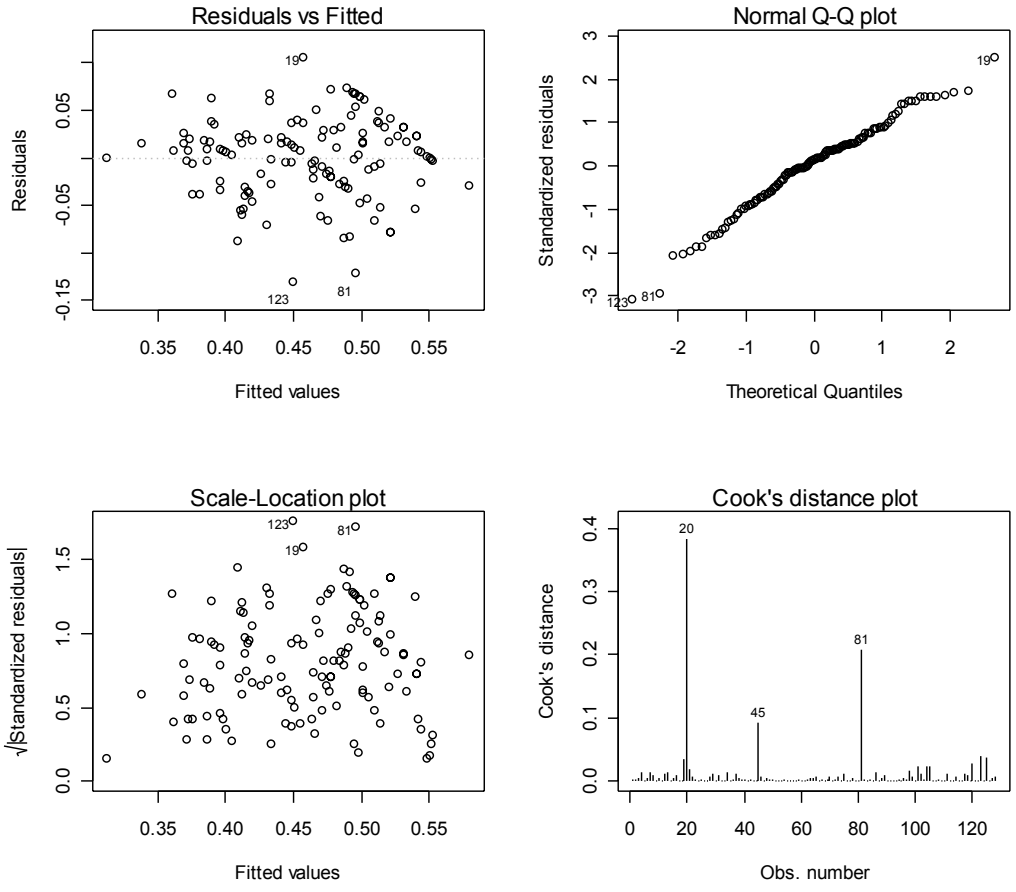
The Box-Cox plot suggested a reciprocal fourth root transformation. Let's begin with that:

```
> recipfourthroot.lm<-lm(I(1/age^0.25)~weight+jaw,data=deer.df)
> summary(recipfourthroot.lm)

Call:
lm(formula = I(1/age^0.25) ~ weight + jaw, data = deer.df)
```

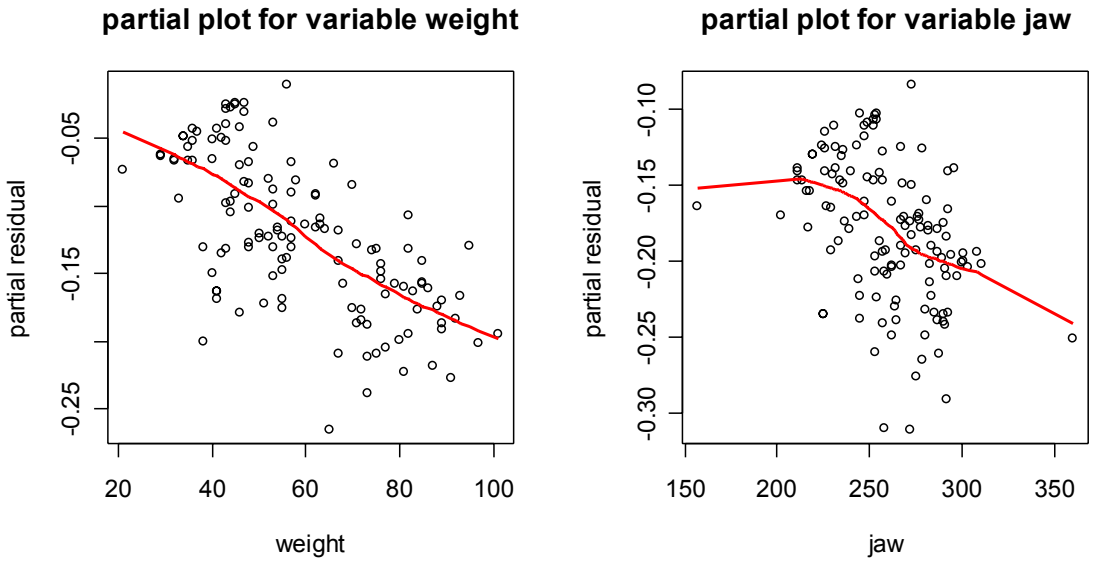
```
Residuals:
      Min        1Q    Median        3Q       Max
-0.130800 -0.027289  0.005627  0.025530  0.104709

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7628911  0.0461202  16.541  < 2e-16 ***
weight      -0.0020667  0.0003748  -5.515 1.92e-07 ***
jaw         -0.0006942  0.0002411  -2.879  0.00469 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.04246 on 125 degrees of freedom
Multiple R-Squared: 0.6328,     Adjusted R-squared: 0.6269
F-statistic: 107.7 on 2 and 125 DF,  p-value: < 2.2e-16
```

Lets check this model with further plots

> plot(recipfourthroot.lm)



```
> partial.res.plot(recipfourthroot.lm)
```



Since these plots look OK, we will go with the model using age$^{-1/4}$ as the response, and weight and jaw untransformed. There are 2 outliers, points 20 and 81, which we will remove.

Out final model is

```
> final.lm<-lm(I(1/age^0.25)~weight+jaw, subset=
                    (1:128)[-c(20,81)],data=deer.df))
> summary(final.lm)

Call:
lm(formula = I(1/age^0.25) ~ weight + jaw, data = deer.df, subset =
(1:128)[-c(20,
    81)])

Residuals:
      Min        1Q    Median        3Q       Max
-0.132201 -0.027619  0.005092  0.025110  0.103703

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7749681  0.0549893  14.093  < 2e-16 ***
weight      -0.0020844  0.0004308  -4.839 3.83e-06 ***
jaw         -0.0007312  0.0002915  -2.508   0.0134 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.04104 on 123 degrees of freedom
Multiple R-Squared: 0.6579,     Adjusted R-squared: 0.6523
F-statistic: 118.3 on 2 and 123 DF,  p-value: < 2.2e-16
```

3. Use the model you have developed in Questions 1 and 2 to predict the age of a deer that weighs 60 kg and has a jaw length of 200 mm.

```
> newdata.df<-data.frame(weight=60,jaw=200)
> predict(final.lm,newdata.df,interval="p")
           fit       lwr       upr
[1,] 0.5036751 0.4145429 0.5928074
```

We need to convert this back to the original units, by undoing the effect of the -1/4 power transformation.

```
> result<-predict(final.lm,newdata.df,interval="p")
> 1/result^4
          fit      lwr    upr
[1,] 15.53810 33.86275 8.0974
```

Our predicted age is between 8.1 and 33.8 months, with a best guess of 15 months.

Mark scheme:

Q1: 20 marks. 5 marks each for plots and comments on (i) non-planar regression (ii) non-constant variance (iii) outliers (iv) lack of normality

Q2: 10 marks. Allocate marks proportional to the success of the transformation. Deduct 3 marks if no outliers are deleted, unless a good argument is made for their inclusion. The final R^2 should be close to the one I got (66%) or hopefully even better.

Q3: 10 marks. Deduct 3 marks if the response has been transformed in Question 2 but not back transformed. Deduct 3 marks if no interval is calculated. Deduct 6 marks if the confidence interval for the mean response is given instead of the prediction interval.