Department of Statistics

COURSE STATS 330

Model answers for Assignment 3, 2006

Instructions: Hand in your completed assignment to the Student Resource Centre by 4pm on Thursday 14th September.

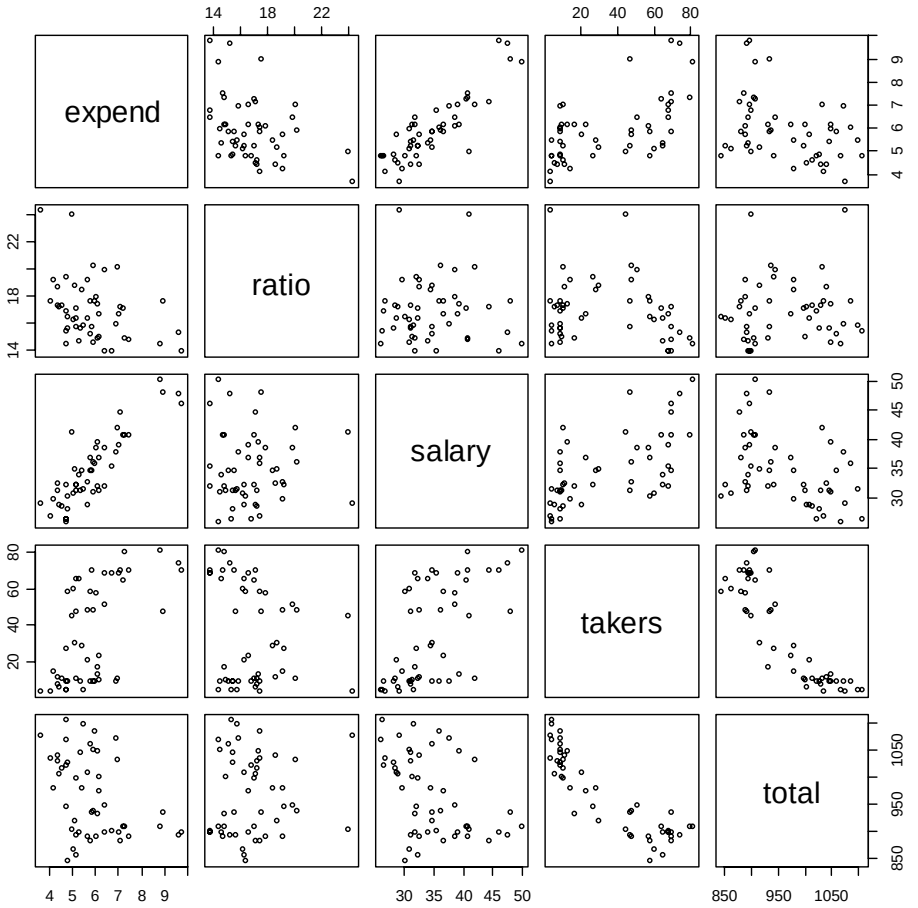
Question 1

(a) Read the data into R, checking the data as usual. The data are in the file ***sat.csv*** *on the course web page. [5 marks]*

```
sat.df = read.csv(file.choose(), header=T)
```

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

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



Pairs plots indicate some non-linearity between total and takers. Quite a strong relationship between response (**`total`**) and **`takers`**. **`salary`** and **`expend`** are highly correlated, as we might expect:

```
round(cor(sat.df),2)
```

```
        expend ratio salary takers total
expend    1.00 -0.37   0.87   0.59 -0.38
ratio    -0.37  1.00   0.00  -0.21  0.08
salary    0.87  0.00   1.00   0.62 -0.44
takers    0.59 -0.21   0.62   1.00 -0.89
total    -0.38  0.08  -0.44  -0.89  1.00
```

Let's fit the regression and examine the residual plots:

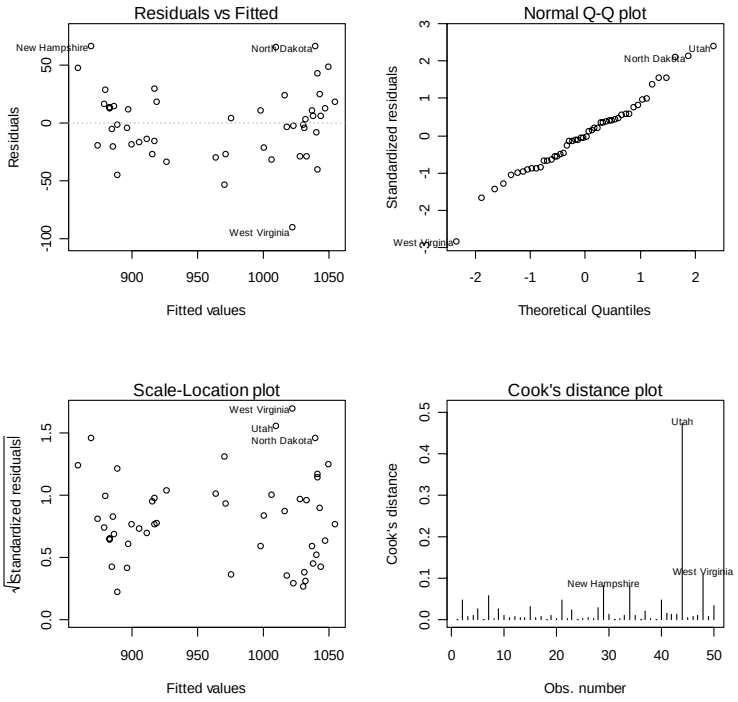
```
sat.lm = lm(total ~ ., data = sat.df)
```

```
Call:
lm(formula = total ~ ., data = sat.df)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674
ratio         -3.6242     3.2154  -1.127    0.266
salary         1.6379     2.3872   0.686    0.496
takers        -2.9045     0.2313 -12.559 2.61e-16 ***

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-Squared: 0.8246,     Adjusted R-squared: 0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```



Seems to be some curvature in the residuals/fitted value plot.

- *Are all the variables required in the regression? Use variable selection techniques to choose a suitable subset if not. [3 marks]*

Not all variables are significant. Try APR and stepwise regression:

```
library(leaps)
all.poss.regs(sat.lm)
```

```
      rssp   sigma2 adjRsq    Cp    AIC    BIC       CV expend ratio salary takers
1 58433.15 1217.357  0.783 8.640 58.640 62.464 6295.526      0     0      0      1
2 49520.06 1053.618  0.812 2.306 52.306 58.042 5608.548      1     0      0      1
3 48315.37 1050.334  0.812 3.179 53.179 60.827 5963.290      0     1      1      1
4 48123.90 1069.420  0.809 5.000 55.000 64.560 6076.456      1     1      1      1
```

```
step(lm(total~1, data=sat.df), formula(sat.lm),
direction="both")
… stuff omitted…
Call:
lm(formula = total ~ takers + expend, data = sat.df)

Coefficients:
(Intercept)       takers       expend
    993.832       -2.851       12.287
```

Model indicated is either one with **expend** and **takers** (based on stepwise, CV BIC, adjusted R^2, Cp) or one with **ratio, salary** and **takers** (based on AIC)

Go with simpler model, which has $R^2 = 0.8195$ (full model has $R^2 = 0.8246$)

```
sub.lm=lm(formula = total ~ takers + expend, data = sat.df)
summary(sub.lm)

Call:
lm(formula = total ~ takers + expend, data = sat.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 993.8317    21.8332  45.519  < 2e-16 ***
takers       -2.8509     0.2151 -13.253  < 2e-16 ***
expend       12.2865     4.2243   2.909  0.00553 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.46 on 47 degrees of freedom
Multiple R-Squared: 0.8195,     Adjusted R-squared: 0.8118
F-statistic: 106.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

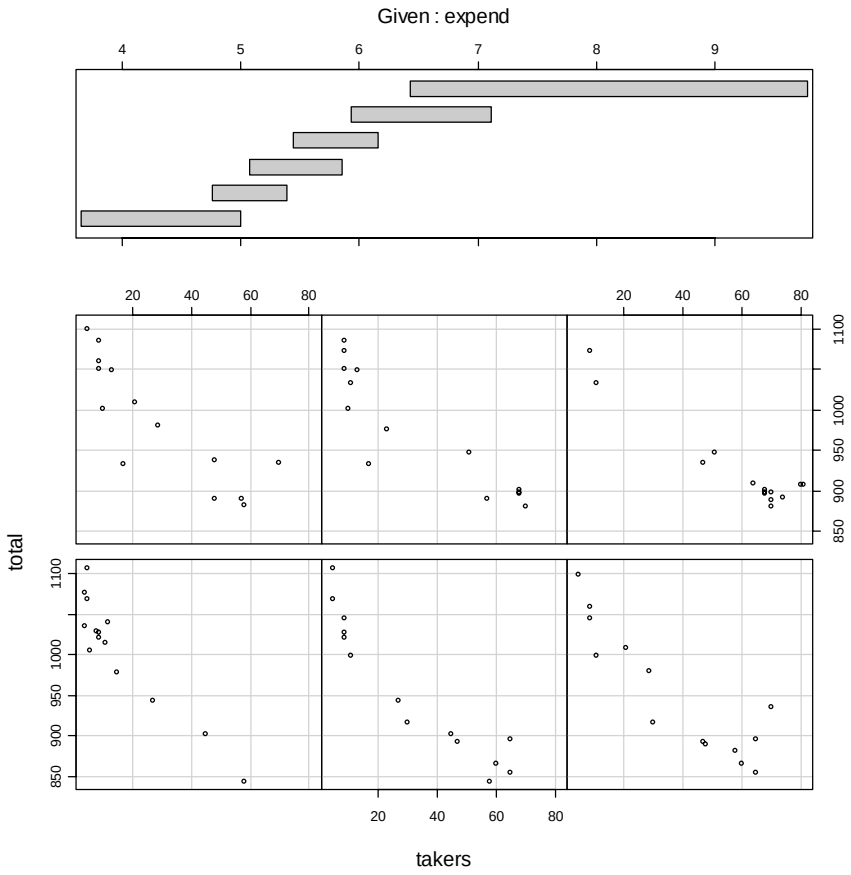
- *Is the relationship between the response and the explanatory variables linear? If not, can a suitable transformation be made? [3 marks]*

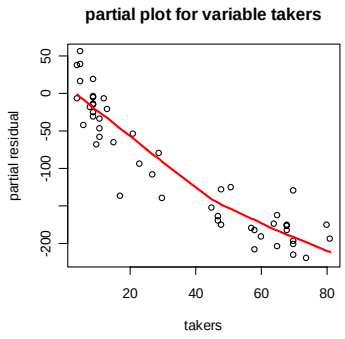
Recheck for planarity – coplot is good since only 2 independent vars. Coplot (shown overleaf) shows evidence of non-planar relationship. (Panels don't have same slope)

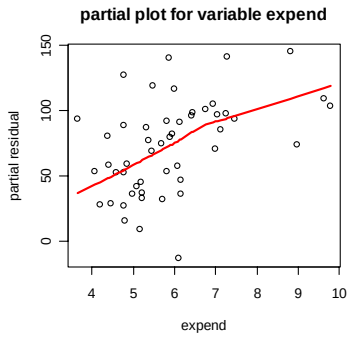
Let's investigate transformations of independent variables. See plots overleaf.

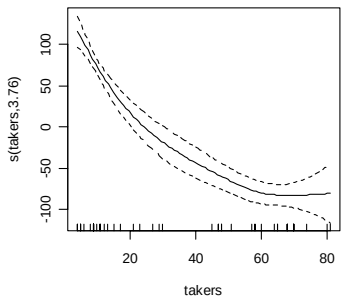
Plots suggest transforming both. A transformation of the response might be good, but in this case the Box-Cox technique indicates a transformation that doest work very well. We tried transforming with quadratics, but only the quadratic in takers was significant. This suggests the model

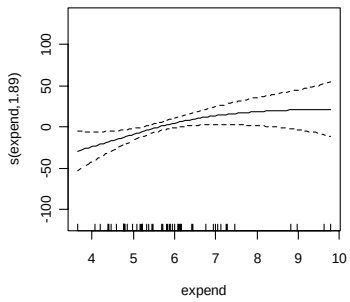
total ~ takers + I(takers^2) + expend
with $R^2 = 0.8859$.


Given : expend
4
5
6
7
8
9
20
40
60
80
1100
1000
950
900
850
total
takers


partial plot for variable takers
partial residual
50
0
-50
-100
-200
20
40
60
80
takers


partial plot for variable expend
partial residual
150
100
50
0
4
5
6
7
8
9
10
expend


s(takers,3.76)
100
50
0
-50
-100
20
40
60
80
takers


s(expend,1.89)
100
50
0
-50
-100
4
5
6
7
8
9
10
expend

Plots suggest transforming both. A transformation of the response might be good, but in this case the Box-Cox technique indicates a transformation that doest work very well. We tried transforming with quadratics, but only the quadratic in takers was significant. This suggests the model

total ~ takers + I(takers^2) + expend

with $R^2 = 0.8859$.

```
sub.lm=lm(formula = total ~ takers + I(takers^2) + expend,
data = sat.df)
summary(sub.lm)

Call:
lm(formula = total ~ takers + I(takers^2) + expend, data = sat.df)

Residuals:
    Min      1Q  Median      3Q     Max
-73.446 -14.631  -2.651  16.208  51.080

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.052e+03  2.082e+01  50.511  < 2e-16 ***
takers      -6.381e+00  7.036e-01  -9.068 8.30e-12 ***
I(takers^2)  4.741e-02  9.161e-03   5.175 4.87e-06 ***
expend       7.914e+00  3.498e+00   2.262   0.0285 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.08 on 46 degrees of freedom
Multiple R-Squared: 0.8859,     Adjusted R-squared: 0.8785
F-statistic: 119.1 on 3 and 46 DF,  p-value: < 2.2e-16
```

Plot of residuals etc (next page) shows not evidence of non-planarity.

- *Are there any outliers or influential points? [3 marks]*

```
> influence.measures(sub.lm)
Influence measures of
         lm(formula = total ~ takers + I(takers^2) + expend, data =
sat.df) :

                dfb.1_ dfb.tkrs dfb.I..2  dfb.expn   dffit cov.r   cook.d    hat inf
Alaska         -0.139248  0.08460 -0.10005  0.133134  0.1583 1.438 6.39e-03 0.2459   *
Connecticut     0.004483  0.11706 -0.12851 -0.044062 -0.1858 1.401 8.80e-03 0.2288   *
Massachusetts   0.032638 -0.06196  0.07413 -0.017076  0.0946 1.360 2.29e-03 0.1996   *
New Jersey      0.243316  0.01641 -0.00322 -0.282270 -0.3457 1.287 3.02e-02 0.1891   *
West Virginia   0.288634 -0.23969  0.35949 -0.395101 -0.7353 0.516 1.13e-01 0.0513   *
```

Several points have large COV RATIOS.

LR plot shows that pt 2 (Alaska) has big HMD but small residual, and pt 48 (West Virginia) has a big residual

In fact, if point 2 is removed, then **expend** is no longer significant, so 2 seems influential.

- *Is the normality assumption satisfied? [3 marks]*

WB is 0.992, p=0.65 so even with the large outlier, no problem with the normality. Normal plot is straight.

- *Is the equal variance assumption satisfied? [3marks]*

No relationship between residuals and fitted values, so OK.

[15 marks altogether for Question 1b]

(c) Use your model to carefully explain the relationship between the explanatory variables and the response.

There is weak evidence that increased expenditure makes a very small difference in the SAT score (about points on the SAT for each extra $1000 spent per pupil.)

In particular, comment on the following:

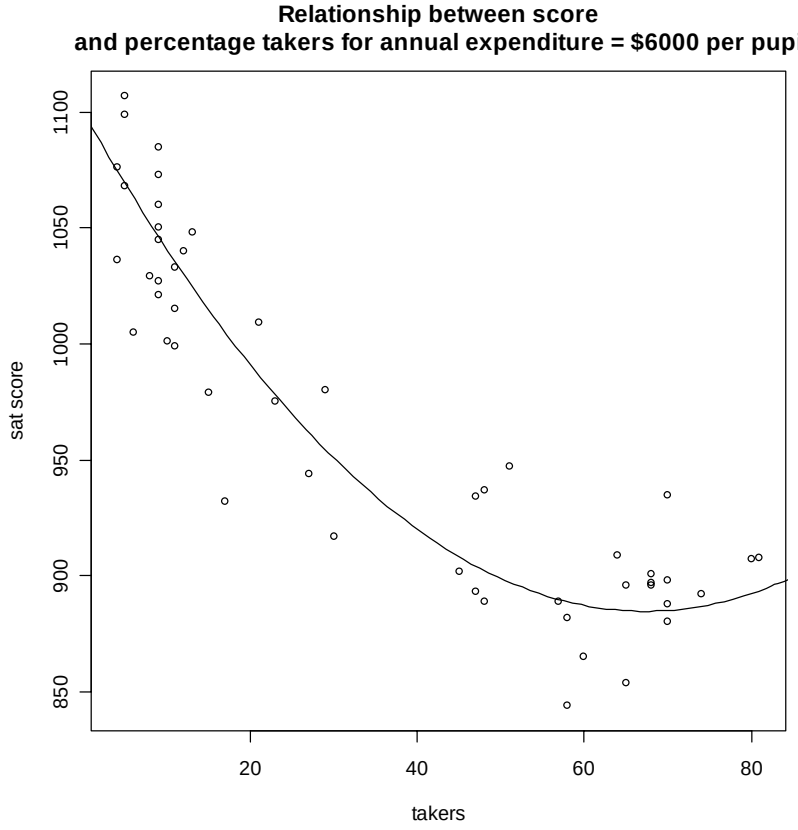
- *Do these data provide any evidence that, other things being equal, higher teacher salaries are associated with better student performance?*

There is no strong relationship between teacher salaries and student performance, given the other variables. Correlation is -0.44, Salary is insignificant in the regression, p=0.496. However, expenditure is important, and salaries are strongly related to expenditure. We can't tell if it is the salary or non-salary part of expenditure that is associated with higher scores. [2 marks]

- *Other things being equal, is student achievement associated with the percentage of students taking the SAT? If so, can you explain why this might be so?*

Strong (quadratic) relationship between these. Can graph the quadratic to see the relationship:

```
x = seq(0,100, length=100)
y =  1.052e+03  - 6.381e+00*x + 4.741e-02*x^2 + 7.914*6.000
plot(sat.df$takers,sat.df$total, xlab="takers",
ylab="sat score",
main = "Relationship between score\n
and percentage takers for annual expenditure = $6000 per
pupil")
lines(x,y)
```



Possible explanation: Bright students will take the test in every state. In some states a bigger group (necessarily including some not so bright students) sits the test, bringing the average down. [3 marks]

Question 2.

(a) In Question 1 in Assignment 2, we treated the variables **`exposure, danger`** *and* **`predation`** *as numeric variables. Would the model have been improved if they had been treated as factors? [5 marks]*

Fitting all the variables, we get

```
mammal.factor.lm = lm(sleep ~ lbrain +lbody + loglifespan +
gestation + factor(danger) + factor(exposure) +
factor(predation),  data=my.mammal.df)
summary(mammal.factor.lm)
```

```
Call:
lm(formula = sleep ~ lbrain + lbody + loglifespan + gestation +
    factor(danger) + factor(exposure) + factor(predation), data =
my.mammal.df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-5.26537 -1.01138 -0.05975  0.99197  6.04915

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         10.024955   2.057566   4.872 2.51e-05 ***
lbrain              -0.660506   0.627996  -1.052 0.300325
lbody               -0.383605   0.471393  -0.814 0.421440
loglifespan          1.974603   0.753645   2.620 0.013045 *
gestation           -0.009958   0.005757  -1.730 0.092733 .
factor(danger)2     -7.972284   1.923293  -4.145 0.000213 ***
factor(danger)3    -13.170985   2.726749  -4.830 2.85e-05 ***
factor(danger)4    -15.292253   3.421387  -4.470 8.27e-05 ***
factor(danger)5    -20.592646   4.484226  -4.592 5.77e-05 ***
factor(exposure)2   -0.881185   1.153132  -0.764 0.450039
factor(exposure)3   -0.542910   1.711592  -0.317 0.753035
factor(exposure)4    2.868010   1.885700   1.521 0.137525
factor(exposure)5    3.240505   3.147954   1.029 0.310559
factor(predation)2   7.795531   1.812640   4.301 0.000136 ***
factor(predation)3  11.092403   2.912312   3.809 0.000558 ***
factor(predation)4  12.402748   3.420297   3.626 0.000932 ***
factor(predation)5  12.015182   3.709072   3.239 0.002678 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.316 on 34 degrees of freedom
Multiple R-Squared: 0.8334,     Adjusted R-squared: 0.755
F-statistic: 10.63 on 16 and 34 DF,  p-value: 6.134e-09
```

R^2 has gone up to 0.833. The improvement is significant:

```
mammal.lm = lm(sleep ~ lbody + lbrain + loglifespan +
gestation + predation + exposure + danger, data=my.mammal.df)

anova(mammal.lm, mammal.factor.lm)
Analysis of Variance Table

Model 1: sleep ~ lbody + lbrain + loglifespan + gestation + predation +
    exposure + danger
Model 2: sleep ~ lbrain + lbody + loglifespan + gestation + factor(danger)
+ factor(exposure) + factor(predation)
  Res.Df    RSS Df Sum of Sq     F  Pr(>F)
1     43 313.90
2     34 182.39  9    131.51 2.724 0.01656 *
```

There seems to be a significant improvement.

(b) Use stepwise model selection to select a model, treating ***exposure, danger*** *and* ***predation*** *as factors. Do you wind up with the same model you chose in Assignment 2? [5 marks]*

Output from step:

```
Call:
lm(formula = sleep ~ lbrain + factor(danger) + factor(predation) +
loglifespan + gestation, data = my.mammal.df)

Coefficients:
       (Intercept)              lbrain     factor(danger)2     factor(danger)3
         10.114874           -1.004687           -8.306688          -13.561982
   factor(danger)4     factor(danger)5  factor(predation)2  factor(predation)3
        -14.790109          -18.772589            7.650430           11.680245
factor(predation)4  factor(predation)5         loglifespan           gestation
         13.609115           12.710682            2.125674           -0.008383
```

The model now has extra variables: **`predation`**, **`loglifespan`** and **`gestation`**. All except **`gestation`** are significant at the 5% level.

```
anova(lm(formula = sleep ~ lbrain + factor(danger) + factor(predation) +
loglifespan + gestation, data = my.mammal.df))
Analysis of Variance Table

Response: sleep
                   Df Sum Sq Mean Sq F value    Pr(>F)
lbrain              1 416.10  416.10 76.1571 1.044e-10 ***
factor(danger)      4 322.26   80.56 14.7454 2.012e-07 ***
factor(predation)   4  87.80   21.95  4.0176  0.008004 **
loglifespan         1  35.04   35.04  6.4135  0.015462 *
gestation           1  20.58   20.58  3.7669  0.059529 .
Residuals          39 213.08    5.46
```

(c) Repeat the prediction you made in assignment 2, using the model fitted in (b). Do you get a more convincing prediction?. [5 marks]

```
step.lm = lm(sleep ~ lbrain + factor(danger) + factor(predation) +
loglifespan + gestation, data = my.mammal.df)

newdata=data.frame(lbrain=log(490), danger = 5, predation=5,
loglifespan=log(23.6), gestation=440)

predict(step.lm, newdata, se=T, interval="prediction")
$fit
           fit      lwr     upr
[1,] 0.8606802 -4.52243 6.24379

$se.fit
[1] 1.272457

$df
[1] 39

$residual.scale
[1] 2.337454
```

Prediction previously was 3.191147 . Current prediction seems a bit low. [5 marks]