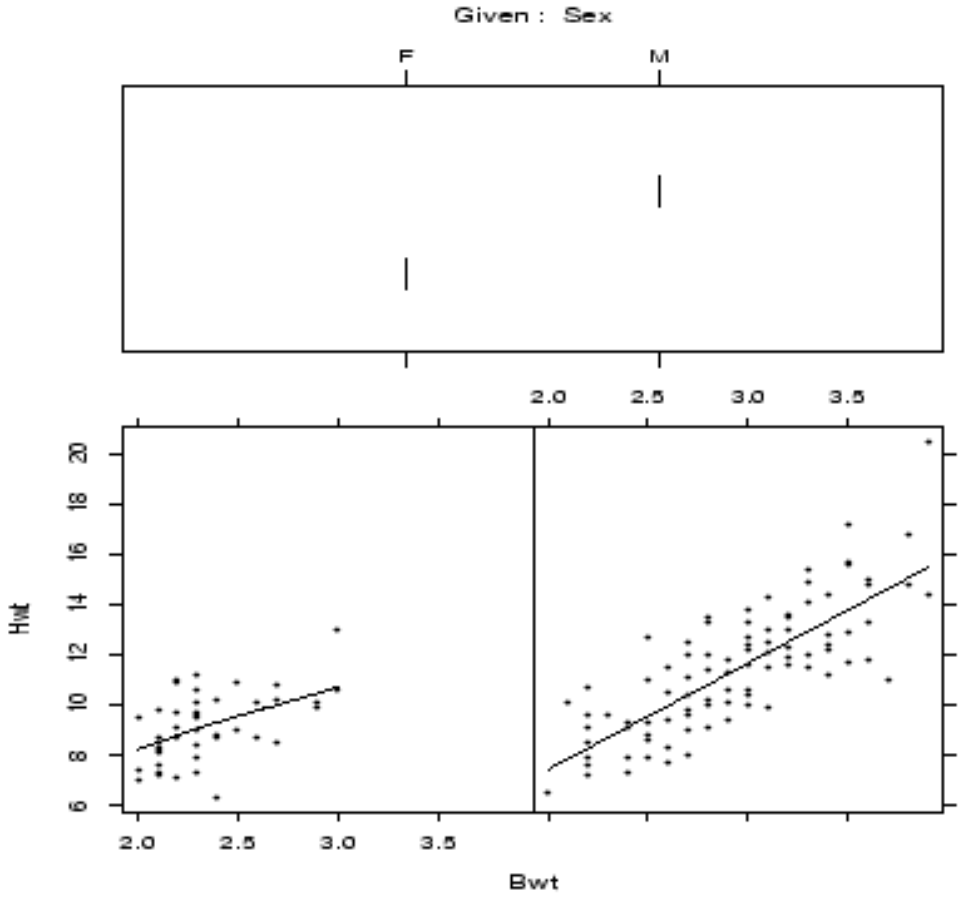
1. A study on the variation of heart weight, Hwt (g), with body weight, Bwt (kg), was carried out on a sample of 47 female and 97 male adult cats. All animals had a body weight of at least 2 kilograms. The question of interest is whether the relationship between heart weight and body weight is the same for both sexes. A trellis plot of the data is given below with smoothed lines for each sex superimposed.



(a) Several models were fitted to the data in the attempt to relate variation in heart weight to the continuous variable body weight, and the factor Sex. For each model the table below briefly describe the form of the model . [**Hint:** A sketch of the fitted model may be useful.]

Model	Formula	Residual SS	Degrees of freedom
1	Hwt~1	847.62	143
2	Hwt~Sex	705.26	142
3	Hwt~Bwt	299.53	142
4	Hwt~Bwt+Sex	299.37	141
5	Hwt~Sex*Bwt	291.05	140

(b) Using the trellis plot and your answers to (a) which of the 5 models describes the data adequately. Give reasons for your conclusions.
The following p-values of the F-distribution may be useful:

$$F_{1,142}(0.95) \cong F_{1,141}(0.95) \cong F_{1,140}(0.95) \cong 3.91$$

(c) Model 5 can be specified in two ways:

```
> model.5<-lm(Hwt~Sex*Bwt,data=cats.df)
> model.5a<-lm(Hwt~Bwt*Sex,data=cats.df)
```

The analysis of variance tables for these two models, given below, show that the Sum of Squares for Sex differs between the two models. Explain why this is so.

```
> anova(model.5)
Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
Sex        1  142.3657 142.3657  68.4811 0.00000000
Bwt        1  405.8815 405.8815 195.2381 0.00000000
Sex:Bwt    1    8.3317   8.3317   4.0077 0.04722465
Residuals 140  291.0467   2.0789

> anova(model.5a)
Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Bwt        1  548.0924 548.0924 263.6448 0.0000000
Sex        1    0.1548   0.1548   0.0745 0.7853491
Bwt:Sex    1    8.3317   8.3317   4.0077 0.0472246
Residuals 140  291.0467   2.0789
```

(d) Use the coefficients for Model 5 to give the fitted model relating heart weight to body weight for male cats. Explain what the coefficients labelled `MBwt` is estimating. Using the fitted model predict the heart weight of a male cat weighing 2.75 kg.

```
> dummy.coef(model.5)
$"(Intercept)":
 (Intercept)
    2.981312

$Sex:
 F       M
 0 -4.1654

$Bwt:
      Bwt
 2.636414

$"Sex:Bwt":
 FBwt     MBwt
    0 1.676265
```

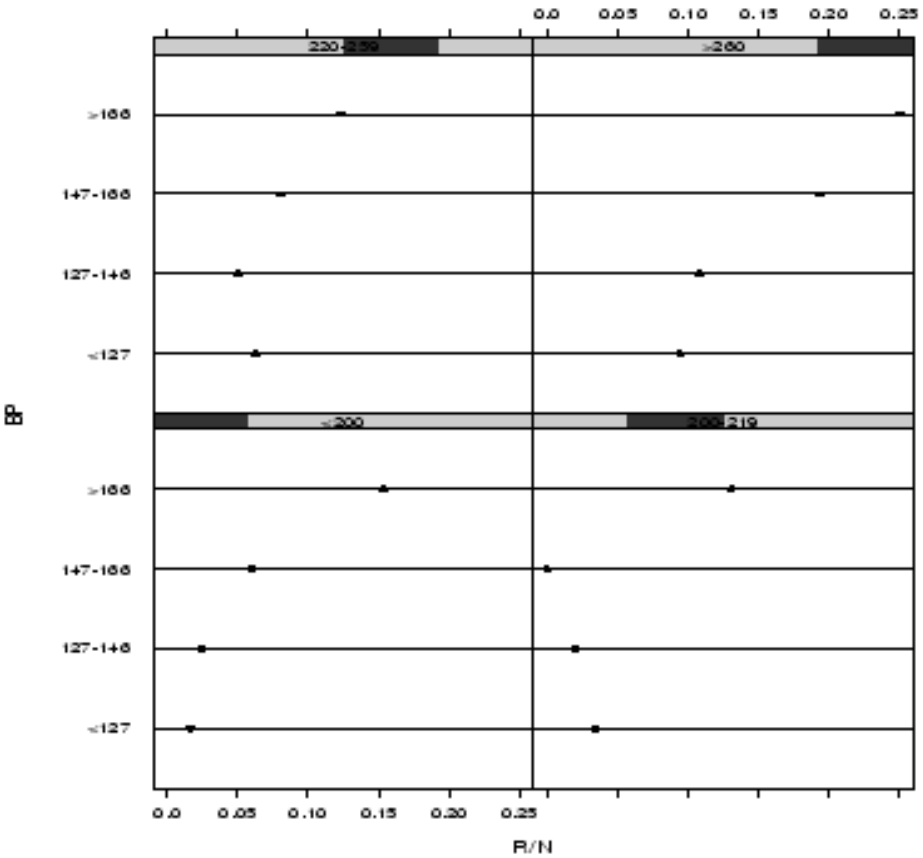
2. A 1974 study of Coronary heart Disease, CHD, classified 1329 men by their Blood pressure, (mm Hg) and Serum Cholesterol, (mg/100cc). The number of men, R, with CHD and the total number of men, N, at each level of Blood Pressure and Serum Cholesterol are given in the table below.

	Blood Pressure, BP			
Serum Cholesterol, Chol	<127	127-146	147-166	>166
<200	2/119	3/124	3/50	4/26
200-219	3/88	2/100	0/43	3/23
220-259	8/127	11/220	6/74	6/49
>260	7/74	12/111	11/57	11/44

The data are stored in the Splus data frame chol.df containing the factors Chol, and BP, and the variables, R and N.

(a) When modelling this data, the response is the proportion of men with coronary heart disease, R/N. The question of interest is how this proportion is affected by the two factors BP and Chol. Explain why a generalised linear model should be used to model this data. Briefly discuss differences between linear models and generalised linear models.

(b) Using the trellis plot below describe the relationship between the proportion R/N and the factors BP and Chol. Suggest an initial model to fit to the data. Make sure you specify both the type of model you would fit and the terms you would include in the model formula.



(c) Some output from fitting the model R/N∼BP*Chol is given below. Does the interaction term BP:Chol need to be retained in the model? Give a reason to justify your conclusions

```
> chol.m1<-glm(R/N~BP*Chol,family=binomial,weights=N,data=chol.df,maxit=25)
> anova(chol.m1,test="Chi")
Analysis of Deviance Table
Binomial model
Response: R/N
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev   Pr(Chi)
   NULL                    15   58.72622
     BP  3 23.56317         12   35.16305 0.0000308
   Chol  3 27.08690          9    8.07615 0.0000056
BP:Chol  9  8.07615          0    0.00000 0.5264905
```

(d) Suppose that the interaction term BP:Chol is dropped and the model R/B∼BP+Chol is fitted. Using the trellis plot identify which data point is most likely to be an outlier from the fitted model.

(e) Use the coefficients of the fitted model, R/N∼BP+Chol, predict the proportion of men with a blood pressure of 147-166(mm), and a serum cholesterol level >260(mg/100cc) who have CHD. Is there any combination of factors for which the predicted incidence is greater than 25%?

```
> dummy.coef(chol.m2)
$"(Intercept)":
 (Intercept)
   -3.481939

$BP:
 <127     127-146   147-166     >166
    0 -0.04146085 0.5323561 1.200422

$Chol:
 <200    200-219   220-259     >260
    0 -0.2079774 0.5622288 1.344121
```

3. The rate at which red blood cells settle out of blood plasma is known as the erythrocyte sedimentation rate (ESR). ESR increases if the concentration of certain proteins in the blood rises. This happens in many diseases and the determination of ESR is one of the most common blood tests. One study wished to determine whether ESR is related to two blood plasma protein, fibrinogen and γ-globulin. The concentration, (units g/ml), of these two proteins was measured in 32 individuals. In a "healthy" person ESR is <20 mm/hour. The response variable takes the value 0 for a healthy individual and 1 if the level of ESR is ≥ 20 mm/h.

Ind	Fibrinogen	γ-globulin	Resp	Ind	Fibrinogen	γ-globulin	Resp
1	2.52	38	0	17	3.53	46	1
2	2.56	31	0	18	2.68	34	0
3	2.19	33	0	19	2.60	38	0
4	2.18	31	0	20	2.23	37	0
5	3.41	37	0	21	2.88	30	0
6	2.46	36	0	22	2.65	46	0
7	3.22	38	0	23	2.09	44	1
8	2.21	37	0	24	2.28	36	0
9	3.15	39	0	25	2.67	39	0
10	2.60	41	0	26	2.29	31	0
11	2.29	36	0	27	2.15	31	0
12	2.35	29	0	28	2.54	28	0
13	5.06	37	1	29	3.93	32	1
14	3.34	32	1	30	3.34	30	1
15	2.38	37	1	31	2.99	36	0
16	3.15	36	0	32	3.32	35	0

The main aim of the analysis is to determine the strength of any relationship between the probability that ESR is > 20mm/h and the concentrations of the two proteins. Specifically in inflammatory diseases levels of fibrinogen and γ-globulin are elevated. If there is no relationship between their levels and ESR, then determining ESR has no diagnostic value.

Let π_i be the probability that individual i has Response = 1. The data is stored in the data frame esr.df containing the variables Fib (Fibrogen), Glob (γ-globulin) and Resp (Response).

(a) A generalised linear model was fitted to this data using the following command:

```
esr.m1<-glm(Resp~Fib+Glob,family=binomial,data=esr.df)
```

What type of regression model is being fitted? Write out the theoretical form of the model.

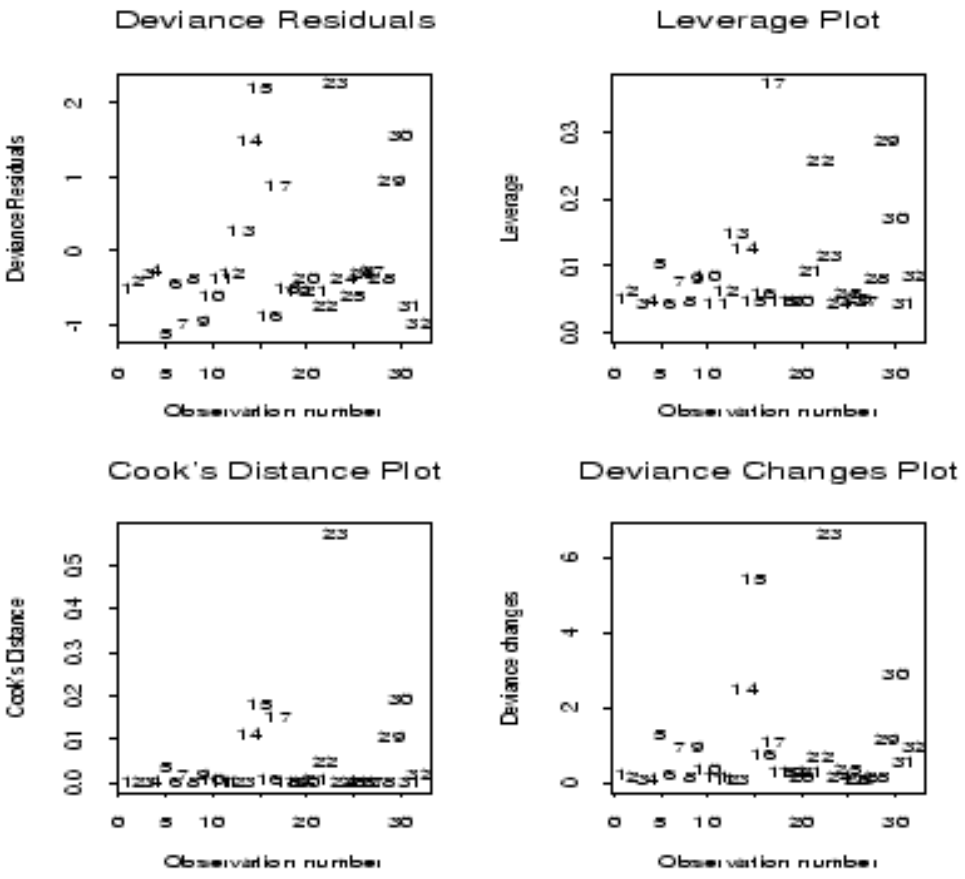
(b) An edited analysis of the variance tables for the model given above and for the model and the summary of one of them are given below.

```
> esr.m2<-glm(Resp~Glob+Fib,family=binomial,data=esr.df)
> anova(esr.m1,test="Chi")
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev   Pr(Chi)
NULL                    31   33.62056
 Fib  1 7.643327         30   25.97724 0.0056983
Glob  1 0.459807         29   25.51743 0.4977141
> anova(esr.m2,test="Chi")
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev   Pr(Chi)
NULL                    31   33.62056
Glob  1 0.622428         30   32.99814 0.4301464
 Fib  1 7.480707         29   25.51743 0.0062364
```

Explain why the deviance of the term Globulin is 0.46 in the first model esr.m1 but is 0.62 in the second model esr.m2. Why are these two deviances not equal?

(c) Determine whether either of both of the proteins affect Response and give brief reasons to justify your conclusions.

(d) Diagnostic plots of the fitted model are displayed below. Which individuals, if any, have undue influence on the fitted model? Which individuals, if any, are badly fitted by the model? Describe what action you would take.



(e) Summaries of two models fitting the two proteins separately are given below:

```
> esr.m3<-glm(Resp~Fib,family=binomial,data=esr.df)
> summary(esr.m3,correlation=F)
Coefficients:
               Value Std. Error   t value
(Intercept) -7.465688  2.9125926 -2.563245
        Fib  2.117927  0.9600345  2.206094

    Null Deviance: 33.62056 on 31 degrees of freedom
Residual Deviance: 25.97724 on 30 degrees of freedom

> esr.m4<-glm(Resp~Glob,family=binomial,data=esr.df)
> summary(esr.m4,correlation=F)
Coefficients:
                 Value Std. Error    t value
(Intercept) -3.93179906 3.42759034 -1.1471030
       Glob  0.07371592 0.09331333  0.7899827

    Null Deviance: 33.62056 on 31 degrees of freedom
Residual Deviance: 32.99814 on 30 degrees of freedom
```

Based on which of the two models esr.m3 and esr.m4 you think fits the data best, estimate the expected concentration of protein, i.e. Fibrinogen or Globulin, required so that the probability Response = 1 is > 0.80.

4. (a) Suppose that we have fitted the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \qquad (i = 1, \ldots, n)$$

In the process we have computed estimates $\hat{\beta}_0, \ldots, \hat{\beta}_k$, fitted values $\hat{y}_i$ and residuals r_i. Explain the reasons for producing the following plots:

i. the residuals against the fitted values, i.e. r_i vs $\hat{y}_i$ for $i = 1, \ldots, n$;

ii. the residuals against each independent variable, i.e. r_i vs x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, k$;

iii. the residuals against their indices, i.e. r_i vs i for $i = 1, \ldots, n$;

iv. the residuals against the lagged residuals, i.e. r_i vs r_{i-1} for $i = 2, \ldots, n$;

(b) Explain the various ways in which a high leverage observation can affect the results of a regression analysis. When are high leverage observations likely to create problems in an analysis?

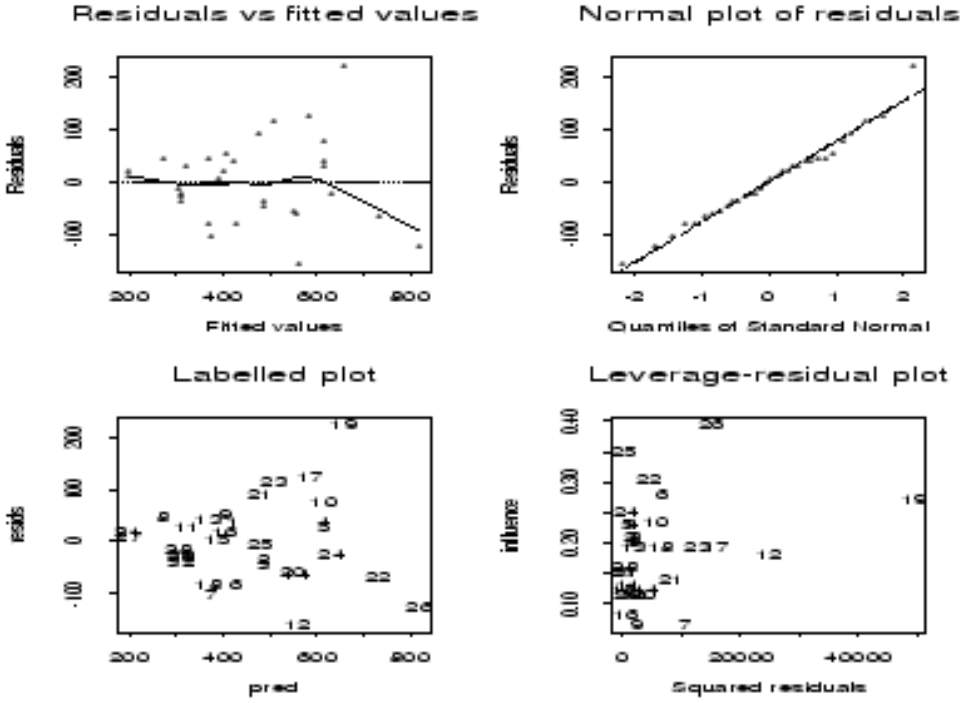
5. In a Rand Corporation study, data were collected on 32 light water reactor nuclear power plants in the US. The investigators wanted to identify factors impacting on construction costs, and wanted to develop a formula that would help predict the cost of future plants. The variables were:

C	Cost, in constant millions of dollars (i.e. 490=$490 million)
D	Date, expressed as a decimal (ie 68.58 = 58% of the way through 1968)
CT	Takes the value 1 if plant has a cooling tower, 0 otherwise.
N	Number or previous plants designed by the design team
S	Power plant capacity (MWe)
NE	Takes value 1 if plant is in North East United States, 0 otherwise

Suppose that the data are in a data frame **power.df** containing the variables above. Two models are fitted, Model 1 and Model 2. Study the output and then answer the questions.

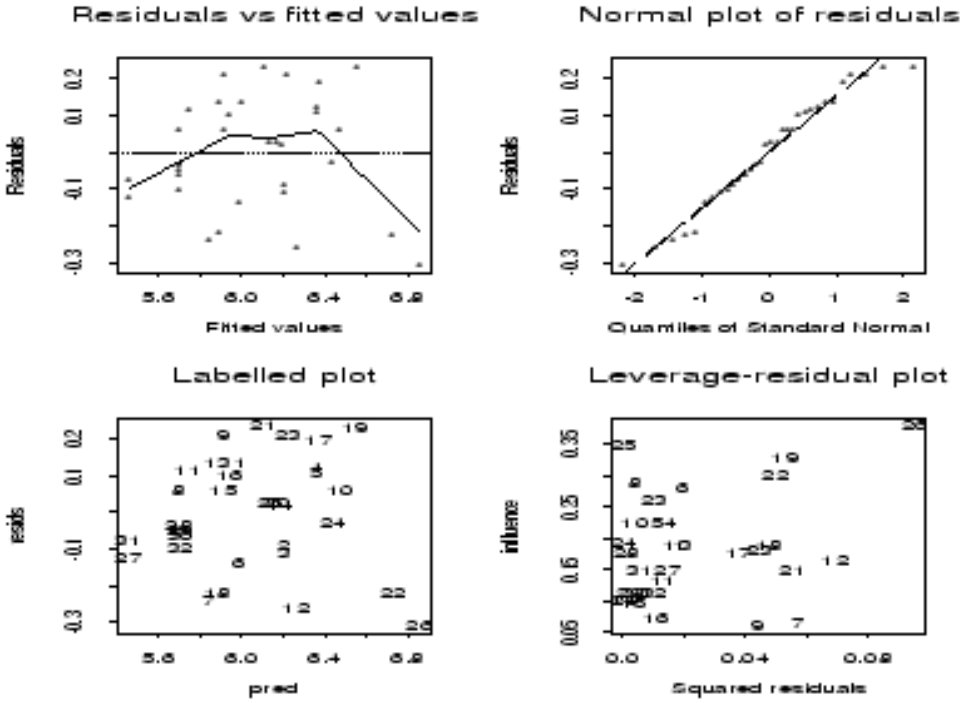
Model 1 output:

```
> m1<-lm(C~D+CT+N+S+NE,data=reactor.df)
> summary(m1,correlation=F)
                Value Std. Error    t value   Pr(>|t|)
(Intercept) -8533.9104  1234.4769    -6.9130     0.0000
          D   125.7379    18.0577     6.9631     0.0000
         CT    72.6537    30.7217     2.3649     0.0258
          N    -8.5374     3.0004    -2.8454     0.0085
          S     0.4624     0.0823     5.6184     0.0000
         NE   135.6183    35.5037     3.8198     0.0007

Residual standard error: 84.53 on 26 degrees of freedom
Multiple R-Squared: 0.7929
F-statistic: 19.91 on 5 and 26 degrees of freedom, the p-value is 3.776e-08
```



Model 2 output:

```
> m2<-lm(log(C)~D+CT+log(N)+log(S)+NE,data=reactor.df)
> summary(m2,correlation=F)
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -17.5713    2.3998  -7.3221   0.0000
          D   0.2734    0.0322   8.4867   0.0000
         CT   0.1739    0.0613   2.8379   0.0087
     log(N)  -0.1373    0.0349  -3.9303   0.0006
     log(S)   0.7446    0.1249   5.9612   0.0000
         NE   0.3114    0.0709   4.3895   0.0002

Residual standard error: 0.168 on 26 degrees of freedom
Multiple R-Squared: 0.8342
F-statistic: 26.16 on 5 and 26 degrees of freedom, the p-value is 2.255e-09
```



(a) The second model uses the logarithms, rather than the raw values of some of the variables. Use the output above to decide which of the two models is better. Give reasons for your answer.

(b) The regression model makes certain assumptions about the data. Are these justified for the model you have chosen in (a)? Quote specific details of the output above in support of your answer.

(c) Does the analysis above indicate that a cooling tower significantly increases costs? What about building a plant in the North East? If there is an increase, what is it?

(d) Estimate the cost of a plant built at the beginning of 1971 (ie with D=71) in the North East, with a cooling tower and power plant net capacity of 800 MWe. Assume the designers of the plant have designed 5 previous plants.