Q2 The following data record the amount of electricity output, y, generated by a windmill over 25 separate fifteen minute periods, during each of which the average wind speed (in miles per hour) is recorded. The `R` data frame `windmill` contains the measured output (`output`) and average wind speed (`velocity`). Some output has been deleted.

Two predictors of output are considered:

- x_1 – velocity;
- x_2 – the reciprocal of velocity, that is, $x_2 = 1/x_1$,

and a simple linear regression model is fit separately using each predictor in turn.

```
 1 > x1<-windmill$velocity
 2 > x2<-1/windmill$velocity
 3 > y<-windmill$output
 4 >
 5 > fit1<-lm(y~x1);summary(fit1)
 6 Coefficients:
 7             Estimate Std. Error t value Pr(>|t|)
 8 (Intercept)  0.13796    0.16287   0.847    0.406
 9 x1           0.23764    0.02421
10 ---
11 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
12
13 Residual standard error: 0.2985 on 23 degrees of freedom
14 Multiple R-squared:  0.8072,    Adjusted R-squared:  0.7989
15 F-statistic:      on 1 and 23 DF,  p-value: 1.087e-09
16
17 > fit2<-lm(y~x2);summary(fit2)
18 Coefficients:
19             Estimate Std. Error t value Pr(>|t|)
20 (Intercept)  3.04008    0.09597   31.68   < 2e-16 ***
21 x2          -7.41975    0.46032
22 ---
23 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
24
25 Residual standard error: 0.1939 on 23 degrees of freedom
26 Multiple R-squared:  0.9187,    Adjusted R-squared:  0.9151
27 F-statistic:     on 1 and 23 DF,  p-value: 5.022e-14
```

(a) What conclusions about the predictive capability of the two predictors can be made on the basis of this output ? Make specific reference to line numbers when citing evidence to support your conclusions.

4 Marks

(b) Predict, using each model in turn, the electricity output produced if the average wind velocity in a given period is 6.5 miles per hour.

4 Marks

Question continued on the next page.

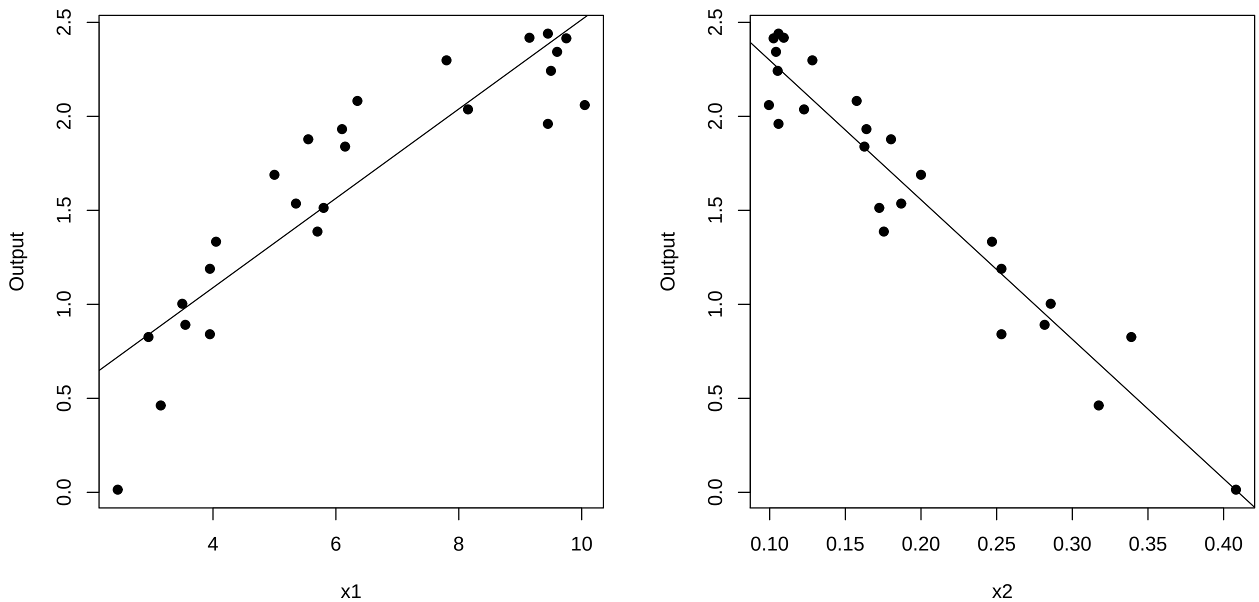
(c) For any linear regression analysis, the sums of squares decomposition

$$\mathrm{SS_T} = \mathrm{SS_{Res}} + \mathrm{SS_R}$$

holds. Compute $\mathrm{SS_T}$ for the model based on x_1, and the model based on x_2.

4 Marks

(d) The following plot depicts the two straight line fits



By considering the residuals implied by these plots, comment on the adequacy of the two straight line models.

2 Marks

(e) The following code computes the hat matrix, $\mathbf{H}$, corresponding to the analysis based on x_1, and extracts the diagonal elements, here rounded to four decimal places.

```
> X<-cbind(1,x1)
> H<-X %*% solve(t(X) %*% X) %*% t(X)
> round(diag(H),4)
 [1]  0.0414 0.0750 0.0950 0.0882 0.0433 0.0636 0.0504 0.0900 0.1135 0.1070
 [11] 0.0454 0.0401 0.1070 0.0420 0.1202 0.0402 0.0556 0.1120 0.1354 0.0401
 [21] 0.1091 0.1346 0.0721 0.0750 0.1036
```

What is the sum of the diagonal elements of $\mathbf{H}$?

2 Marks

(f) The $\mathrm{SS_T}$ quantity from part (c) can be written

$$\mathrm{SS_T} = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}$$

Define the matrix $\mathbf{H}_1$, and write down its trace.

2 Marks

(g) Show that $\mathbf{H}_1$ is symmetric and idempotent.

2 Marks

Answer to Q2:

(a) In both models, there is good predictive ability for the response. The first analysis using x_{i1} indicates a significantly positive slope (line (9: 0.23764/0.02421 = 9.815) and an R^2 statistic of 0.8072 (line 14), which is reasonably high. The second analysis using $1/x_{i1}$ indicates a significantly negative slope (line (21: -7.41975/0.46032 = -16.118) and an even higher R^2 statistic of 0.98 (line 26), which is very high. The models indicate that generation increases with increasing wind velocity, as might be predicted.

2 Marks

(b) We have

$$\text{First model}: \widehat{y} = 0.13796 + 0.23764 \times 6.5 = 1.68262$$

$$\text{Second model}: \widehat{y} = 3.04008 - 7.41975/6.5 = 1.89858$$

```
> predict(fit1,newdata=data.frame(x1=6.5))
       1
1.682629
> predict(fit2,newdata=data.frame(x2=1/6.5))
       1
1.898581
```

4 Marks

(c) We know that

$$R^2 = \frac{\text{SS}_\text{R}}{\text{SS}_\text{T}} = 1 - \frac{\text{SS}_\text{Res}}{\text{SS}_\text{T}} = 1 - \frac{(n-2)\widehat{\sigma}^2}{\text{SS}_\text{T}}$$

Therefore for the first analysis

$$\text{SS}_\text{Res} = 23 \times 0.2985^2 = 2.049352$$

$$\text{SS}_\text{T} = \frac{\text{SS}_\text{Res}}{(1-R^2)} = \frac{2.049352}{1-0.8072} = 10.62942$$

$$\text{SS}_\text{R} = 10.62942 - 2.049352 = 8.580068$$

and for the second analysis

$$\text{SS}_\text{Res} = 23 \times 0.1939^2 = 0.8647883$$

$$\text{SS}_\text{T} = \frac{\text{SS}_\text{Res}}{(1-R^2)} = \frac{0.8647883}{1-0.9187} = 10.63356$$

$$\text{SS}_\text{R} = 10.62942 - 0.8647883 = 9.7688$$

Of course, SS_T should be identical in the two models, so any difference must be due to rounding.

To confirm in R:

```
> anova(fit1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 8.5839  8.5839  96.324 1.087e-09 ***
Residuals 23 2.0496  0.0891
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

> anova(fit2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x2         1 9.7688  9.7688  259.81 5.022e-14 ***
Residuals 23 0.8648  0.0376
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
```

4 Marks

(d) The left panel implies a quadratic pattern (residuals greater than zero near the mean x as the points are above the line of best fit, below near the ends of the range of x): the model fit is deficient, indicating an incorrect conditional mean model $\mathbb{E}_{Y_i|X_i}[Y_i|\mathbf{x}_i]$ specification. In the right panel, no such pattern would be evident, so the model fit is adequate 4 Marks

(e) The sum of the diagonal elements of $\mathbf{H}$ is always equal to $p = 2$ for the simple linear regression. There is no need to add the 20 numbers !

2 Marks

(f) The matrix $\mathbf{H}_1$ is the hat matrix from the linear regression model with only the intercept included. It is an $(n \times n)$ matrix with all elements equal to $1/n$.

2 Marks

(g) $\mathbf{H}_1$ is clearly symmetric as all elements are identical. It is idempotent as, in the calculation of $\mathbf{H}_1^2$, the $(i,j)th$ entry is the inner product of the ith row and jth column of $\mathbf{H}_1$ – this equals

$$[1/n\ 1/n \cdots 1/n][1/n\ 1/n \cdots 1/n]^\top = \sum_{l=1}^{n} \frac{1}{n^2} = \frac{1}{n}$$

so again each element is equal to $1/n$.

2 Marks