1181506

MAST30025

Linear Statistical Models Assignment 2

Due: 29 Apr 2022 5:00 pm

1)

Q1

Let $C' = \begin{bmatrix} [(X^TX)^{-1}]_{ii} & [(X^TX)^{-1}]_{ij} \\ [(X^TX)^{-1}]_{ji} & [(X^TX)^{-1}]_{jj} \end{bmatrix}$ where C' is nonsingular

$\vec{b}'^T = [b_i, b_j]$

$\vec{\beta}'^T = [\beta_i, \beta_j]$

$$\frac{(\vec{b}' - \vec{\beta}')^T [C']^{-1} (\vec{b}' - \vec{\beta}')}{\sigma^2} \sim \chi^2_2 \qquad (p = 2)$$

by 3.10 as $\vec{b} \sim MVN(\vec{\beta}, C\sigma^2)$ where $C = (X^TX)^{-1}$ so $\vec{b}' \sim MVN(\vec{\beta}', C'\sigma^2)$ by $(z_1, \dots, z_k) \sim MVN \Rightarrow (z_1, z_2) \sim MVN$ and C' contains b_i, b_j's var and covar

$$\Rightarrow \frac{(\vec{b}' - \vec{\beta}')^T [C']^{-1} (\vec{b}' - \vec{\beta}')}{2\sigma^2} \Big/ \frac{SS_{Res}}{(n-p)\sigma^2} \sim F_{2, n-p}$$

as $\vec{b}'$ is indep of s^2 by 4.14 ($\vec{b}'$ is merely a part out of the full model whilst s^2 is from the full model so this analysis should still follow a full model's framework)

$$\Rightarrow \frac{(\vec{b}' - \vec{\beta}')^T [C']^{-1} (\vec{b}' - \vec{\beta}')}{2\sigma^2} \Big/ \frac{(n-p)s^2}{(n-p)\sigma^2} \sim F_{2, n-p}$$

$$\Rightarrow \frac{(\vec{b}' - \vec{\beta}')^T [C']^{-1} (\vec{b}' - \vec{\beta}')}{2s^2} \sim F_{2, n-p}$$

$$\Rightarrow \mathbb{P}\left[\frac{(\vec{b}' - \vec{\beta}') [C']^{-1} (b' - \beta')}{2s^2} \le f_\alpha \right] = 1 - \alpha$$

$$\Rightarrow (\vec{b}' - \vec{\beta}')^T [C']^{-1} (\vec{b}' - \vec{\beta}') \le 2s^2 f_\alpha$$

where f_α is $(1-\alpha)$th quantile of $F_{2, n-p}$

and s^2 is the sample variance of the full model $= \frac{SS_{Res}}{n-p}$

n = # samples and p = # param of full model

Q 2

Question 2

Setup

```
sold = c(5.5, 5.9, 6.5, 5.9, 8.0, 9.0, 10.0, 10.8)
cost = c(7.2, 10, 9, 5.5, 9, 9.8, 14.5, 8.0)
unemp = c(8.7, 9.4, 10, 9, 12, 11, 12, 13.7)
intRate = c(5.5, 4.4, 4, 7, 5, 6.2, 5.8, 3.9)

data = data.frame(carsSold = sold, cost = cost, unempRate = unemp, intRate =
intRate)



a)

x = cbind(rep(1, length(data)), data$cost, data$unempRate, data$intRate)
y = data$carsSold
n = dim(x)[1]
p = dim(x)[2]
(b = solve(t(x) %*% x, t(x) %*% y))

##            [,1]
## [1,] -7.4044796
## [2,]  0.1207646
## [3,]  1.1174846
## [4,]  0.3861206

e = y - x %*% b
SSRes = sum(e^2)
(s2 = SSRes/(n-p))

## [1] 0.3955368
```

Therefore the estimated parameters for intercept, cost, unemployment rate and interest rate are -7.404, 0.121, 1.117, 0.0386 respectively. The estimated error variance is 0.396.

b)

```
c = solve(t(x) %*% x)
alpha = 0.05
df = n-p
ta = qt(1-alpha/2, df = df)

(b0CI = b[1] + c(-1, 1) * ta * sqrt(s2*c[1,1]))

## [1] -13.8196491  -0.9893101
```

Q2

```
(b1CI = b[2] + c(-1, 1) * ta * sqrt(s2*c[2,2]))

## [1] -0.1525428  0.3940720

(b2CI = b[3] + c(-1, 1) * ta * sqrt(s2*c[3,3]))

## [1] 0.6817719 1.5531974

(b3CI = b[4] + c(-1, 1) * ta * sqrt(s2*c[4,4]))

## [1] -0.2563181  1.0285593
```

Therefore the 95% confidence intervals of the model parameters are

Beta0: (-13.8196491, -0.9893101)

Beta1: (-0.1525428, 0.3940720)

Beta2: (0.6817719, 1.5531974)

Beta3: (-0.2563181, 1.0285593)

c)

```
xst = c(1, 12, 8, 3.5)
(yst = t(xst) %*% b)

##          [,1]
## [1,] 4.335994

yst[1] + c(-1, 1) * qt(0.975, df) * sqrt(s2) * sqrt(1 + (t(xst) %*% c %*% xst
)[1])

## [1] 1.444686 7.227303
```

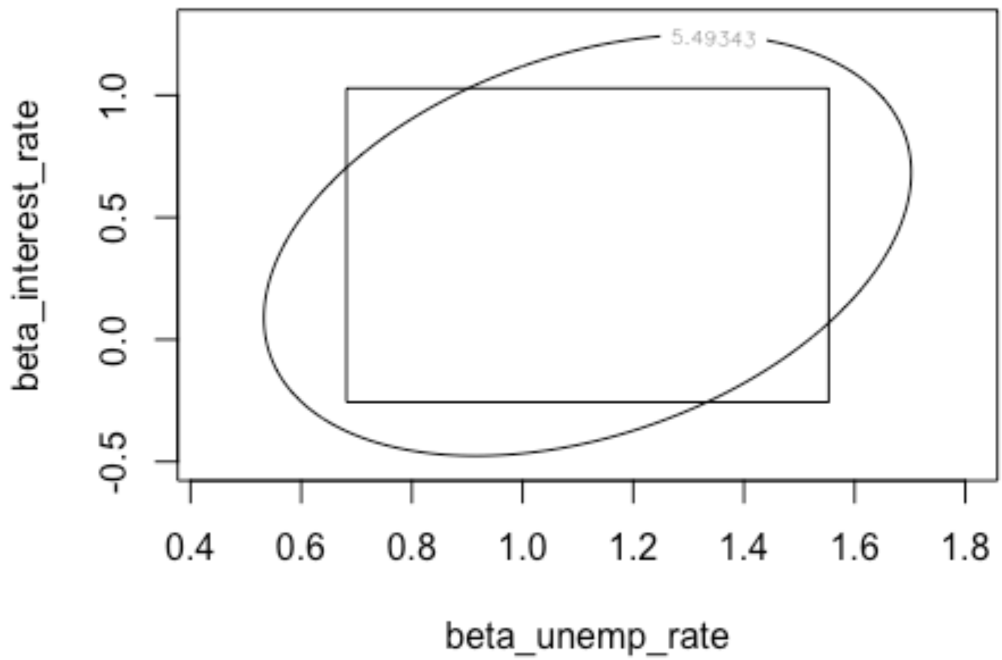
This is not an atypical year because 7 (k) is still within the 95% prediction interval of (1.44, 7.23) for xst = 1, 12, 8, 3.5.

d)

```
n <- dim(x)[1]
c_ = matrix(c(c[3,3], c[3,4], c[4,3], c[4,4]), c(2,2))

b0 <- seq(b2CI[1]-0.25, b2CI[2]+0.25,length=100)
b1 <- seq(b3CI[1]-0.25, b3CI[2]+0.25,length=100)
f <- function(beta0, beta1) {
  f.out <- rep(0, length(beta0))
  for (i in 1:length(beta0)) {
    beta <- matrix(c(beta0[i], beta1[i]), 2, 1)
    f.out[i] <- t(matrix(c(b[3], b[4]), 2, 1) - beta) %*% solve(c_) %*% (matr
```

Q2

```
ix(c(b[3], b[4]), 2, 1) - beta)
  }
  return(f.out)
}
z <- outer(b0, b1, f)
contour(b0, b1, z, levels=2*s2*qf(0.95, 2, n-p),
xlab='beta_unemp_rate', ylab='beta_interest_rate')

rect(b2CI[1], b3CI[1], b2CI[2], b3CI[2])
```



e)

I would expect the joint confidence region to be larger than the rectangle. Theoretically, the joint independent regions should be larger than any joint confidence regions where there exists correlation. However, that is only the case if we are looking at all parameters of the full model, whereas here we are only looking at the joint confidence intervals of two of the four parameters, so without the restrains caused by the correlations with the other two factors the confidence region has expanded to be larger than the joint confidence region.

f)

Q2

$$\frac{(b_1-\beta_1, b_2-\beta_2)^T \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} (b_1-\beta_1, b_2-\beta_2)}{ps^2} \sim F_{2, n-p=4}$$

Treat this as an observation from $F_{2,4}$

$$\Rightarrow x = \frac{(b_1-\beta_1, b_2-\beta_2)^T \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} (b_1-\beta_1, b_2-\beta_2)}{ps^2}$$

Now pdf of $F_{2,4}$ is

$$f(x: 2, 4) = \frac{1}{\beta(2,6)} \left(\frac{2}{6}\right)^{2/2} x^{2/2-1} \left(1+\frac{2}{6}x\right)^{-(2+6)/2}$$
$$= f(b_1, b_2 : 2, 4)$$

$$P((\beta_1,\beta_2) \in \text{rectangle}) = \int_{-0.2563}^{1.0285} \int_{0.6817}^{1.55719} f(b_1, b_2 : 2, 4)\, db_1\, db_2$$

3) Q3

$y_1 = (\vec{x}_1^*)^T \vec{b} + \varepsilon_1^*$, $y_2 = (\vec{x}_2^*)^T \vec{b} + \varepsilon_2^*$ where y_1 and y_2 are indep

$\Rightarrow \varepsilon_1^*$ and ε_2^* are independent

Since $\varepsilon_1, \varepsilon_2 \sim N(0, \sigma^2)$ iid:

$E(y_1+y_2) = (\vec{x}_1^*)^T\vec{\beta} + (\vec{x}_2^*)^T\vec{\beta} = (\vec{x}_1^* + \vec{x}_2^*)^T\vec{\beta}$

and the BLUE for $(\vec{x}_1^* + \vec{x}_2^*)^T\vec{\beta}$ is $(\vec{x}_1^* + \vec{x}_2^*)^T\vec{b}$ by 4.5

$$\text{Error of } y_1+y_2 = (y_1+y_2) - \left((\vec{x}_1^*)^T\vec{b} + (\vec{x}_2^*)^T\vec{b}\right)$$
$$= \left((\vec{x}_1^*)^T\vec{\beta} + \varepsilon_1^* + (\vec{x}_2^*)^T\vec{\beta} + \varepsilon_2^*\right) - \left((\vec{x}_1^* + \vec{x}_2^*)^T\vec{b}\right)$$
$$= \left((\vec{x}_1^* + \vec{x}_2^*)^T\vec{\beta} + \varepsilon_1^* + \varepsilon_2^*\right) - \left((\vec{x}_1^* + \vec{x}_2^*)^T\vec{b}\right)$$

of which ε_1^* ε_2^* are only associated with future obs y_1, y_2 whilst $\vec{b}$ is only dependent on $\vec{y}$, thus they are independent

$$Var\left((y_1+y_2) - ((\vec{x}_1^* + \vec{x}_2^*)^T\vec{b})\right) = Var(\varepsilon_1^* + \varepsilon_2^*) + Var\left((\vec{x}_1^* + \vec{x}_2^*)^T\vec{b}\right)$$
$$= 2\sigma^2 + (\vec{x}_1^* + \vec{x}_2^*)^T(X^TX)^{-1}\sigma^2(\vec{x}_1^* + \vec{x}_2^*)$$
$$= \sigma^2\left(2 + (\vec{x}_1^* + \vec{x}_2^*)^T(X^TX)^{-1}(\vec{x}_1^* + \vec{x}_2^*)\right)$$

$$\Rightarrow \frac{\underbrace{\dfrac{(y_1+y_2) - (\vec{x}_1^* + \vec{x}_2^*)^T\vec{b}}{\sigma\sqrt{2+(x_1^*+x_2^*)^T(X^TX)^{-1}(\vec{x}_1^*+\vec{x}_2^*)}}}_{\sim N(0,1)}}{\underbrace{\sqrt{\dfrac{SS_{Res}/\sigma^2}{(n-p)}}}_{\sim \chi^2_{n-p}}} = \frac{(y_1+y_2) - (\vec{x}_1^* + \vec{x}_2^*)^T\vec{b}}{s\sqrt{2 + (x_1^*+x_2^*)^T(X^TX)^{-1}(\vec{x}_1^*+\vec{x}_2^*)}} \sim t_{n-p}$$

Prediction Interval: $(\vec{x}_1^* + \vec{x}_2^*)^T\vec{b} \pm t_{\alpha/2}\, s\sqrt{2+(\vec{x}_1^*+\vec{x}_2^*)^T(X^TX)^{-1}(\vec{x}_1^*+\vec{x}_2^*)}$

where $t_{\frac{\alpha}{2}}$ is $(1-\frac{\alpha}{2})$th quantile of the t_{n-p} distribution.

Q4

Question 4

Setup

```
#setwd('Desktop/1. University/1. Undergraduate/17. Linear Statistical
Models/assignments/assignment 2')

file = read.csv('bike.csv')
```

a)

```
model = lm(count~temp + hum + wind + visi + dew + solar, data = file)
model$coefficients

##   (Intercept)          temp           hum          wind          visi
## 1247.61646416  -14.65762399  -13.07275492  -21.49446333   -0.02019842
##           dew         solar
##   33.65683041  141.23177238
```

So the full model is

count = 1247.61 - 14.6576 * temp - 13.072 * hum - 21.49 * wind - 0.02 * visi + 33.656 * dew + 141.2317 * solar

b)

```
summary(model)

##
## Call:
## lm(formula = count ~ temp + hum + wind + visi + dew + solar,
##     data = file)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -981.25 -180.39  -10.49  216.83  943.70
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1247.61646  365.74154   3.411 0.000721 ***
## temp         -14.65762   13.03306  -1.125 0.261491
## hum          -13.07275    4.22315  -3.095 0.002120 **
## wind         -21.49446   17.47352  -1.230 0.219461
## visi          -0.02020    0.03652  -0.553 0.580512
## dew           33.65683   13.91736   2.418 0.016090 *
## solar        141.23177   29.83307   4.734 3.18e-06 ***
```

Q4

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 312.4 on 358 degrees of freedom
## Multiple R-squared:  0.4828, Adjusted R-squared:  0.4741
## F-statistic: 55.69 on 6 and 358 DF,  p-value: < 2.2e-16
```

The p-value is <2.2e^-16, meaning that the model is relevant when tested against the corrected sum of squares (which is what R uses by default).

c)

```
basemodel = lm(count~1, data = file)
add1(basemodel, scope = ~ . + temp + hum + wind + visi + dew + solar, test =
"F")

## Single term additions
##
## Model:
## count ~ 1
##        Df Sum of Sq      RSS    AIC  F value    Pr(>F)
## <none>              67535096 4428.8
## temp    1  22345147 45189949 4284.2 179.4932 < 2.2e-16 ***
## hum     1   2388851 65146245 4417.7  13.3109 0.0003025 ***
## wind    1     51159 67483937 4430.5   0.2752 0.6001921
## visi    1   2400889 65134206 4417.6  13.3804 0.0002919 ***
## dew     1  11335989 56199107 4363.7  73.2212  3.31e-16 ***
## solar   1  24146010 43389086 4269.3 202.0094 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model1 = lm(count~solar, data = file)
add1(model1, scope = ~ . + temp + hum + wind + visi + dew, test = "F")

## Single term additions
##
## Model:
## count ~ solar
##        Df Sum of Sq      RSS    AIC F value    Pr(>F)
## <none>              43389086 4269.3
## temp    1   7079973 36309113 4206.3 70.5870 1.016e-15 ***
## hum     1   1063725 42325361 4262.3  9.0978   0.00274 **
## wind    1   1018262 42370824 4262.7  8.6996   0.00339 **
## visi    1      8203 43380883 4271.3  0.0685   0.79375
## dew     1   5687072 37702014 4220.0 54.6050 1.028e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q4

```r
model2 = lm(count~temp+solar, data = file)
add1(model2, scope = ~ . + hum + wind + visi + dew, test = "F")

## Single term additions
##
## Model:
## count ~ temp + solar
##        Df Sum of Sq      RSS    AIC F value  Pr(>F)
## <none>               36309113 4206.3
## hum     1    651690 35657424 4201.7  6.5978 0.01061 *
## wind    1    187042 36122072 4206.4  1.8693 0.17241
## visi    1     39892 36269221 4207.9  0.3971 0.52901
## dew     1    296784 36012329 4205.3  2.9751 0.08541 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model3 = lm(count~temp+solar+hum, data = file)
add1(model3, scope = ~ . + wind + visi + dew, test = "F")

## Single term additions
##
## Model:
## count ~ temp + solar + hum
##        Df Sum of Sq      RSS    AIC F value  Pr(>F)
## <none>               35657424 4201.7
## wind    1    136444 35520980 4202.3  1.3828 0.24040
## visi    1     40570 35616854 4203.3  0.4101 0.52234
## dew     1    520151 35137273 4198.3  5.3292 0.02154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model4 = lm(count~temp+solar+hum+dew, data = file)
add1(model4, scope = ~ . + wind + visi, test = "F")

## Single term additions
##
## Model:
## count ~ temp + solar + hum + dew
##        Df Sum of Sq      RSS    AIC F value Pr(>F)
## <none>               35137273 4198.3
## wind    1    176433 34960840 4198.5  1.8117 0.1792
## visi    1     58641 35078632 4199.7  0.6001 0.4390
```

Thus, the variables selected for the model using Forward Selection (F test) is temp, solar, hum and dew.

d)

```r
basemodel = lm(count ~ 1, data = file)
step(basemodel, scope = ~ . + temp + hum + wind + visi + dew + solar)
```

Q4

```
## Start:  AIC=4428.82
## count ~ 1
##
##          Df Sum of Sq      RSS    AIC
## + solar  1  24146010 43389086 4269.3
## + temp   1  22345147 45189949 4284.2
## + dew    1  11335989 56199107 4363.7
## + visi   1   2400889 65134206 4417.6
## + hum    1   2388851 65146245 4417.7
## <none>               67535096 4428.8
## + wind   1     51159 67483937 4430.5
##
## Step:  AIC=4269.32
## count ~ solar
##
##          Df Sum of Sq      RSS    AIC
## + temp   1   7079973 36309113 4206.3
## + dew    1   5687072 37702014 4220.0
## + hum    1   1063725 42325361 4262.3
## + wind   1   1018262 42370824 4262.7
## <none>               43389086 4269.3
## + visi   1      8203 43380883 4271.3
## - solar  1  24146010 67535096 4428.8
##
## Step:  AIC=4206.3
## count ~ solar + temp
##
##          Df Sum of Sq      RSS    AIC
## + hum    1    651690 35657424 4201.7
## + dew    1    296784 36012329 4205.3
## <none>               36309113 4206.3
## + wind   1    187042 36122072 4206.4
## + visi   1     39892 36269221 4207.9
## - temp   1   7079973 43389086 4269.3
## - solar  1   8880836 45189949 4284.2
##
## Step:  AIC=4201.69
## count ~ solar + temp + hum
##
##          Df Sum of Sq      RSS    AIC
## + dew    1    520151 35137273 4198.3
## <none>               35657424 4201.7
## + wind   1    136444 35520980 4202.3
## + visi   1     40570 35616854 4203.3
## - hum    1    651690 36309113 4206.3
## - solar  1   2237558 37894982 4221.9
## - temp   1   6667937 42325361 4262.3
##
## Step:  AIC=4198.33
## count ~ solar + temp + hum + dew
```

Q4

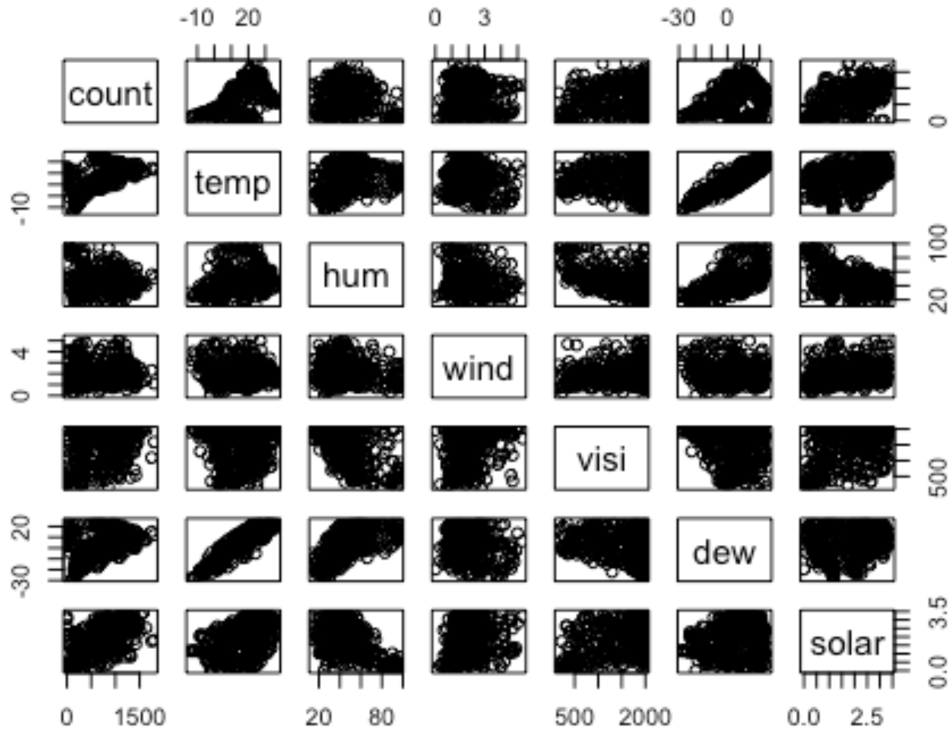
```
##
##         Df Sum of Sq      RSS    AIC
## - temp   1     95758 35233031 4197.3
## <none>               35137273 4198.3
## + wind   1    176433 34960840 4198.5
## + visi   1     58641 35078632 4199.7
## - dew    1    520151 35657424 4201.7
## - hum    1    875056 36012329 4205.3
## - solar  1   2126739 37264012 4217.8
##
## Step:  AIC=4197.32
## count ~ solar + hum + dew
##
##         Df Sum of Sq      RSS    AIC
## <none>               35233031 4197.3
## + wind   1    148902 35084129 4197.8
## + temp   1     95758 35137273 4198.3
## + visi   1     55841 35177190 4198.7
## - solar  1   2078814 37311844 4216.2
## - hum    1   2468983 37702014 4220.0
## - dew    1   7092330 42325361 4262.3

##
## Call:
## lm(formula = count ~ solar + hum + dew, data = file)
##
## Coefficients:
## (Intercept)        solar          hum          dew
##     803.994      132.713       -8.641       18.453
```

The variables included are solar, hum and dew. Of which, solar's coefficient is 132.713 (positive effect on count), humidity coefficient is -8.641 (negative) and dew's coefficient is 18.453 (positive). The intercept term is 803.994. Stepwise has included one less variable (temp) compared to forward selection; this is likely because forward selection uses F-test whilst stepwise uses AIC.

The absolute value of all 3 coefficients and the intercept have reduced in this final model compared to the full model. Particularly for dew, which reduced from 33.65 to 18.45, this could be because of the removal of temp, which from the pairs plot below can be seen to have a high positive correlation with dew. This is supported by the fact that in the full model temp has a coefficient of -14.6576, which is approximately the difference between 33.65 and 18.45. The other variables do not show such significant pattern.

```
pairs(file)
```

Q4




```
e)

finalModel = lm(count ~ solar + hum + dew, data = file)

library(car)

## Loading required package: carData

dst = c(0)
c = matrix(c(0, 1, 0, 0, 0, -1, 0), c(1, 7))
linearHypothesis(model, c, dst)

## Linear hypothesis test
##
## Hypothesis:
## temp - dew = 0
##
## Model 1: restricted model
## Model 2: count ~ temp + hum + wind + visi + dew + solar
##
##   Res.Df      RSS Df Sum of Sq     F  Pr(>F)
## 1    359 35246728
## 2    358 34930987  1    315741 3.236 0.07288 .
```

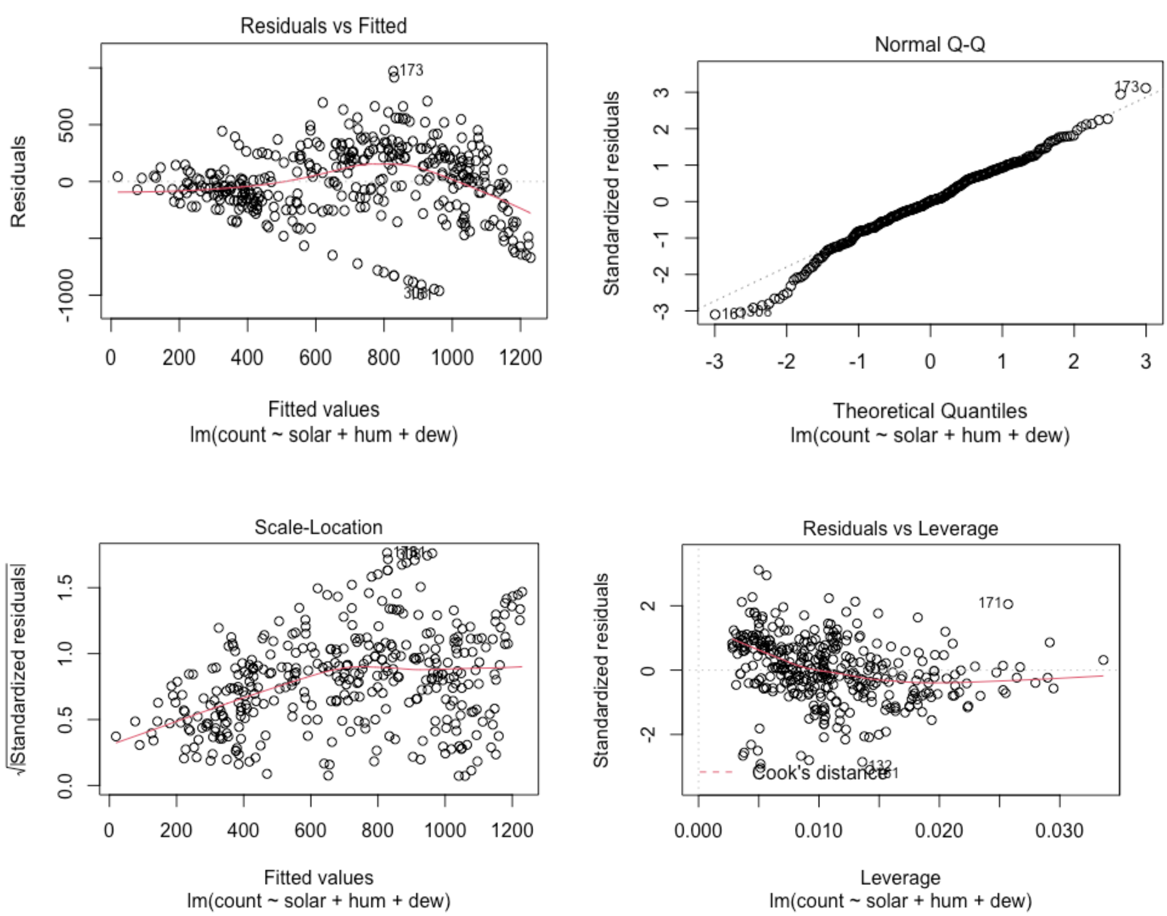
Q4

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The General Linear Hypothesis for this test returned a p value of 0.07288 > 0.05 which means there is not enough evidence to reject the null hypothesis that temp and dew have the same effect on the number of bikes rented.

f)

```
plot(finalModel)
```



The residual vs fitted values plot show an initially increasing but later decreasing trend. However this doesn't seem to be a big problem. However, instances 173, 308 and 161 seem to have a very large residual and thus their removal should be considered.

The QQ plot of standardised residuals has a problem in the lower values whereby the standardised residuals deviate from Normal Distribution (heavy left tail). Once again instances 161 and 308 seems to be causing the greatest problem.

Q4

The Sqrt of abs(standardised residuals) suggests a problem as there seems to be an increasing trend in the sqrt of abs(standardised residuals) as fitted values increase.

The Standardised residuals vs leverage plot looks fine as none of the points exceed 0.5 in terms of cook's distance. However there seems to be an initial decreasing trend in the residuals which whilst is not overly significant should be monitored and assessed.

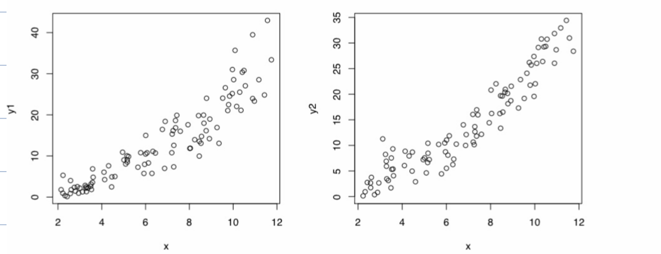
Overall though, there are no significant issues with this model and it should be suitable.

5) Q5

The difference between transformation $y \sim x + x^2$ and $\sqrt{y} \sim x$ is:

1) for $\sqrt{y} \sim x$ y must be strictly positive (i.e. $y > 0$) or else the transformation will not work.
Meanwhile, $y \sim x^2 + x$ does not carry this precondition.

2) Assuming all y values are strictly positive,
$$\sqrt{y} = \beta_1 x + \beta_0 + \varepsilon \equiv y = \beta_0^2 + 2\beta_0\beta_1 x + \beta_1^2 x^2 + 2\beta_0 \varepsilon + 2\beta_1 x \varepsilon + \varepsilon^2$$
so $\sqrt{y} \sim x$ and $y \sim x^2 + x$ $(\equiv y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon)$
differs in that:
1) the errors on the former will better fit data displaying heteroskedacity due to multiplicity of errors.
(if $|\varepsilon_i| > 1$ then $\varepsilon_i^2 > |\varepsilon_i|$ if ; $|\varepsilon_i| < 1$ then $\varepsilon_i^2 < |\varepsilon_i|$)
and 2) the latter has one more parameter available to be fitted, so it can better fit more complexly distributed quadratic data but is more prone to noise (more chance of overfit)



Observing Diagram 1 and 2, all y values in both diagrams are strictly positive so both can use $\sqrt{y} \sim x$
However: Diagram 1's points seem sparser as $x \to \infty$ and thus any fitted line will likely have increasing absolute val of residuals (display heteroskedacity) - so best use $\sqrt{y} \sim x$.

Meanwhile Diag 2's does not seem to display heteroskedacity, so it should use $y \sim x^2 + x$