a3-q2, q3

06/11/2019

2.(a)

```r
options(scipen=999)
car = read.table("car_consumption.txt", header = TRUE)
price = car$price
engine = car$engine
hp = car$hp
weight = car$weight
consumption = car$consumption
fit = lm(consumption ~ price + engine + hp + weight)
summary(fit)
```

```
##
## Call:
## lm(formula = consumption ~ price + engine + hp + weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99443 -0.45646 -0.04083  0.40251  1.06211
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  1.83800628  0.79336708   2.317  0.03022 *
## price        0.00003394  0.00004508   0.753  0.45959
## engine       0.00120783  0.00072210   1.673  0.10856
## hp          -0.00374192  0.01503044  -0.249  0.80570
## weight       0.00372829  0.00129971   2.869  0.00893 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6512 on 22 degrees of freedom
## Multiple R-squared:  0.9295, Adjusted R-squared:  0.9167
## F-statistic: 72.54 on 4 and 22 DF,  p-value: 0.000000000002393
```

Thus the fitted equation is $y = 1.83800628 + 0.00003394x_1 + 0.00120783x_2 - 0.00374192x_3 + 0.00372829x_4$

(b)

```r
qf(0.95, 4, 22)
```

```
## [1] 2.816708
```

Form summary in part(a), $|F| > F_{0.05,4,22}$. Hence we reject H_0 at 0.05 significance level, at least one variable is significant. 92.95% of the total variation in responses can be explianed by the linear model.

(c) we test explaplanatory variables at 0.05 significance level

```r
qt(0.975, 22)
```

```
## [1] 2.073873
```

From summary(fit) in part(a), $t_{\beta_4} > t_{0.025,22}$. We can conclude that weight is important in determining the consumption of the car.

(d)

```r
newdata = data.frame(price = 40000, engine = 2000, hp = 100, weight = 1500)
predict(fit, newdata, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 10.82934 9.37299 12.28569
```

The 95% prediction interval is $[9.37299, 12.28569]$

(e)

```r
max(car$price)
```

```
## [1] 50900
```

It is not appropriate to predict the consumption for another new carwith the same engine size, weight and horse power as the one in (d), but is much moreexpensive with a price tag of 60000.Since the maximum value of price is 50900 which is smaller than 60000.

(f) From summary(fit), horsepower is insignificant

```r
fit2 = lm(car$consumption~car$price+car$engine+car$weight)
summary(fit2)
```

```
##
## Call:
## lm(formula = car$consumption ~ car$price + car$engine + car$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98306 -0.47722  0.01806  0.39881  1.05185
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.82416707 0.77511083   2.353   0.0275 *
## car$price   0.00002902 0.00003970   0.731   0.4721
## car$engine  0.00108060 0.00049962   2.163   0.0412 *
## car$weight  0.00380332 0.00123824   3.072   0.0054 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6378 on 23 degrees of freedom
## Multiple R-squared:  0.9293, Adjusted R-squared:  0.9201
## F-statistic: 100.8 on 3 and 23 DF,  p-value: 0.0000000000002232
```

Car price is the most insignificant predictor, sow we remove it.

```r
fit3 = lm(consumption~engine+weight)
summary(fit3)
```

```
##
## Call:
## lm(formula = consumption ~ engine + weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0166 -0.4245  0.1030  0.3238  1.2116
##
```

```
## Coefficients:
##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 1.3922757  0.4968842   2.802    0.00988 **
## engine      0.0013110  0.0003839   3.415    0.00227 **
## weight      0.0045047  0.0007751   5.812 0.00000542 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6315 on 24 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9217
## F-statistic: 153.9 on 2 and 24 DF,  p-value: 0.00000000000002047
```

Multiple R-squared: 0.9295, Adjusted R-squared: 0.9167 R^2 from the first model is slightly greater than the R^2 from this model. Adjusted R^2 from the first model is slightly smaller than the adjusted R^2 from this model.

(g)

```r
newdata = data.frame(engine = 2000, weight = 1500)
predict(fit3, newdata, interval = "prediction")
```

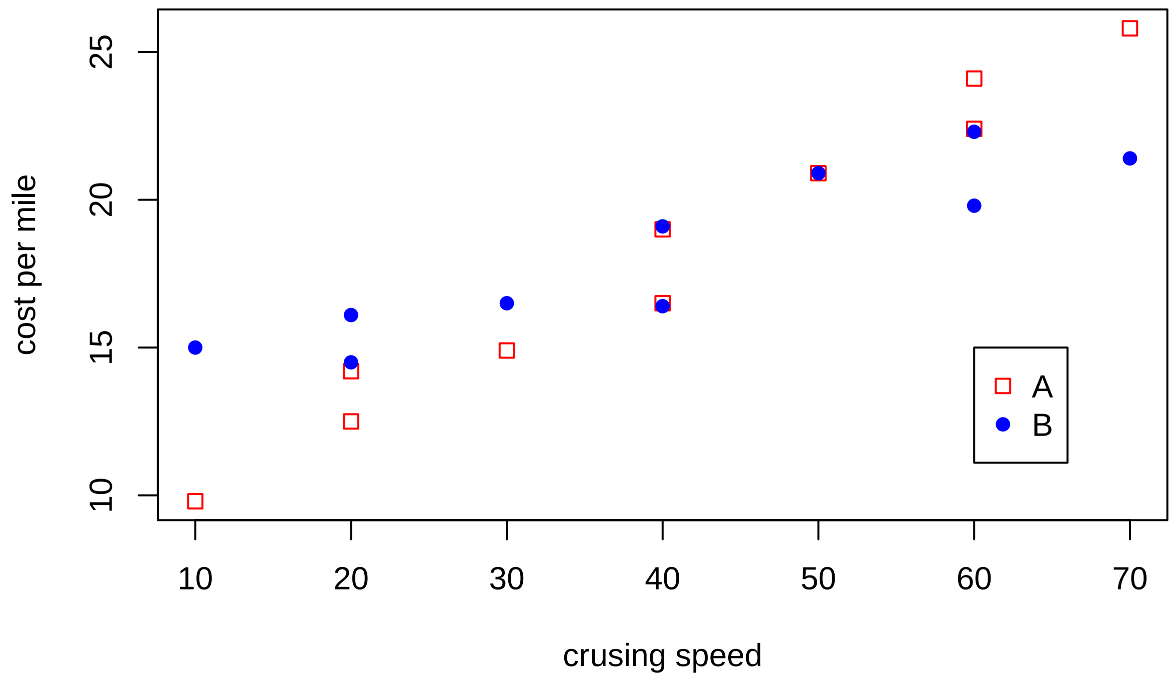
```
##       fit      lwr      upr
## 1 10.7714 9.394866 12.14793
```

The 95% prediction interval is $[9.394866, 12.14793]$ $[9.37299, 12.28569]$ The length of two intervals are really close, but for the first model we have two insignificant variables and the last mode has narrower length ,I would prefer the last model.

3.

(a)

```r
options(scipen=999)
tire = read.table("tire.txt", header = TRUE,)
plot(tire[tire$x2 == 'A', ]$x1, tire[tire$x2 == 'A', ]$y, xlab = 'crusing speed', ylab = 'cost per mile
par(new = TRUE)
points(tire[tire$x2 == 'B', ]$x1, tire[tire$x2 == 'B', ]$y, xlab = 'crusing speed', ylab = 'cost per mil
legend(60, 15, legend=c('A', 'B'), col=c('red', 'blue'), pch=c(0,16))
```



The relationship appears to be the same in the middle, but different at small or large speed.

(b)

```
type = factor(tire$x2)
tirefit = lm(tire$y~type + tire$x1 + tire$x1 * type)
summary(tirefit)
```

```
##
## Call:
## lm(formula = tire$y ~ type + tire$x1 + tire$x1 * type)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8000 -0.7150 -0.1600  0.8925  1.5111
##
## Coefficients:
##                  Estimate Std. Error t value      Pr(>|t|)
## (Intercept)       7.61000    0.80635   9.438 0.00000006108 ***
## typeB             5.41222    1.14035   4.746      0.000219 ***
## tire$x1           0.26000    0.01821  14.275 0.00000000016 ***
## typeB:tire$x1    -0.13056    0.02576  -5.069      0.000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.093 on 16 degrees of freedom
## Multiple R-squared:  0.9408, Adjusted R-squared:  0.9297
## F-statistic: 84.81 on 3 and 16 DF,  p-value: 0.0000000004878
```

We want to test $H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$

```r
nrow(tire)
```

```
## [1] 20
```

```r
qt(0.975, nrow(tire) - 4)
```

```
## [1] 2.119905
```

using t-test
$|t| = 5.069$(from summary) and $t_{0.025,16} = 2.119905$. Hence, $|t| > t_{0.025,16}$
And, p-value$= 0.000114 < 0.05$.
We reject H_0. The makes of tires is significant to the slop whcih is the additional operation cost per mile if curusing speed increased by 1 unit.

(c)

```r
tirefit2 = lm(tire$y ~ tire$x1)
anova(tirefit2, tirefit)
```

```
## Analysis of Variance Table
##
## Model 1: tire$y ~ tire$x1
## Model 2: tire$y ~ type + tire$x1 + tire$x1 * type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     18 49.969
## 2     16 19.108  2    30.861 12.921 0.0004572 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We want to test $H_0 : \beta_1 = \beta_3 = 0$ vs $H_a : \beta_1 \neq \beta_3 \neq 0$
using f-test

```r
((49.969 - 19.108) / 2) / (19.108 / 16)
```

```
## [1] 12.92066
```

```r
qf(0.95, 2, 16)
```

```
## [1] 3.633723
```

$F = \frac{(49.969-19.108)/2}{19.108/16} = 12.92066 > F_{2,16}$ so we reject H_0, the makes of tires is significant to operation cost per mile.