MATH 4780 - Homework 5

Liam Fruzyna

December 3rd, 2018

#7.2 A solid-fuel rocket propellant loses weight after it is produced. The following data are available:

Dataset setup

```
library(MPV)
solid <- p7.2
```

a. Fit a second-order polynomial that expresses weight loss as a function of the number of months since production.

```
fit_solid <- lm(y ~ x + I(x^2), data=solid)
```

b. Test for significance of regression.

```
summary(fit_solid)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2), data = solid)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.005364 -0.002727  0.001045  0.002409  0.003273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.633000   0.004196   389.2  < 2e-16 ***
## x           -1.232182   0.007010  -175.8 5.09e-14 ***
## I(x^2)       1.494545   0.002484   601.6  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003568 on 7 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.859e+06 on 2 and 7 DF,  p-value: < 2.2e-16
```

The model has very low p-values and F-statistic so it is significant.

c. Test the hypothesis $H_0 : \beta_2 = 0$. Comment on the need for the quadratic term in this model.

```
anova(fit_solid)
```

```
## Analysis of Variance Table
##
## Response: y
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 42.703  42.703 3355253 < 2.2e-16 ***
## I(x^2)     1  4.607   4.607  361974 < 2.2e-16 ***
## Residuals  7  0.000   0.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is significant as $F = \frac{4.607}{0} = \infty$

d. Are there any potential hazards in extrapolating with this model?

There could be some risk in extrapolating due to the model being quadratic.

#8.3 Consider the delivery time data in Example 3.1. In Section 4.2.5 noted that these observations were collected in four cities.

Dataset setup

```
delivery <- p8.3
delivery$city <- c(replicate(7, "San Diego"), replicate(10, "Boston"), replicate(6, "Austin"), replicate
```

a. Develop a model that relates delivery time y to cases x_1, distance x_2, and the city in which the delivery was made. Estimate the parameters of the model.

```
fit_delivery <- lm(y ~ x1 + x2 + city, data=delivery)
```

b. Is there an indication that delivery site is an important variable?

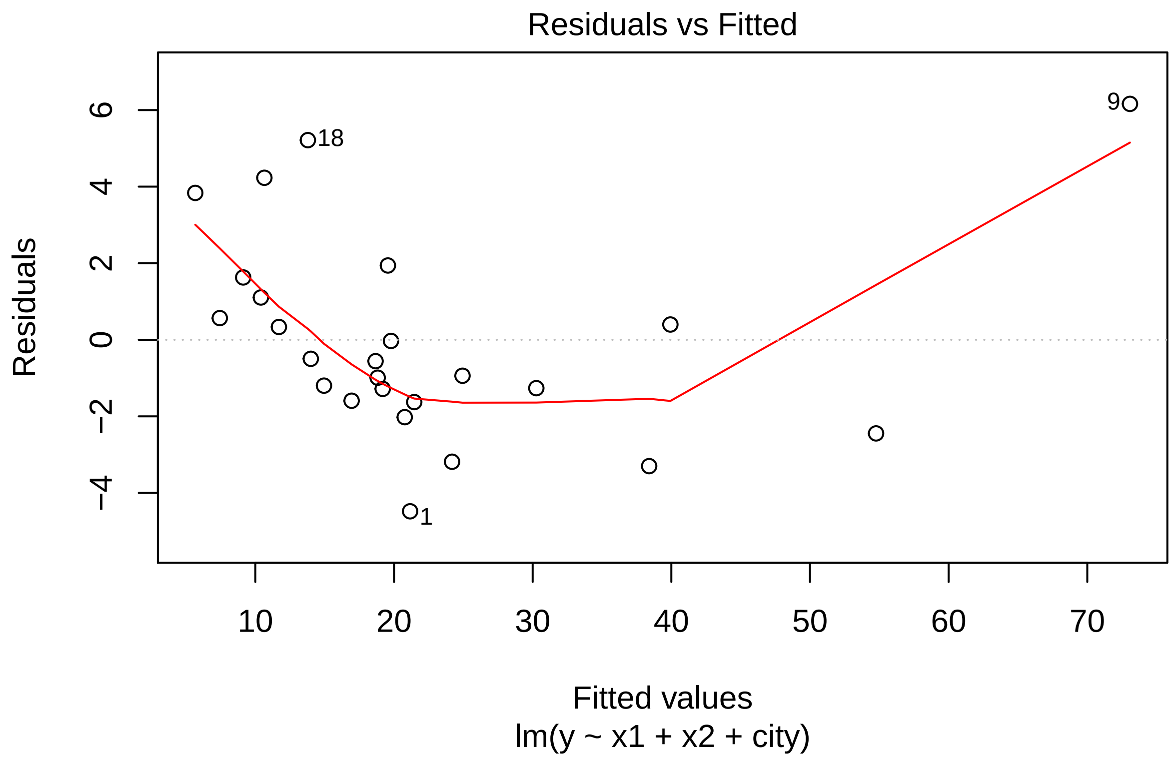
```
summary(fit_delivery)
```

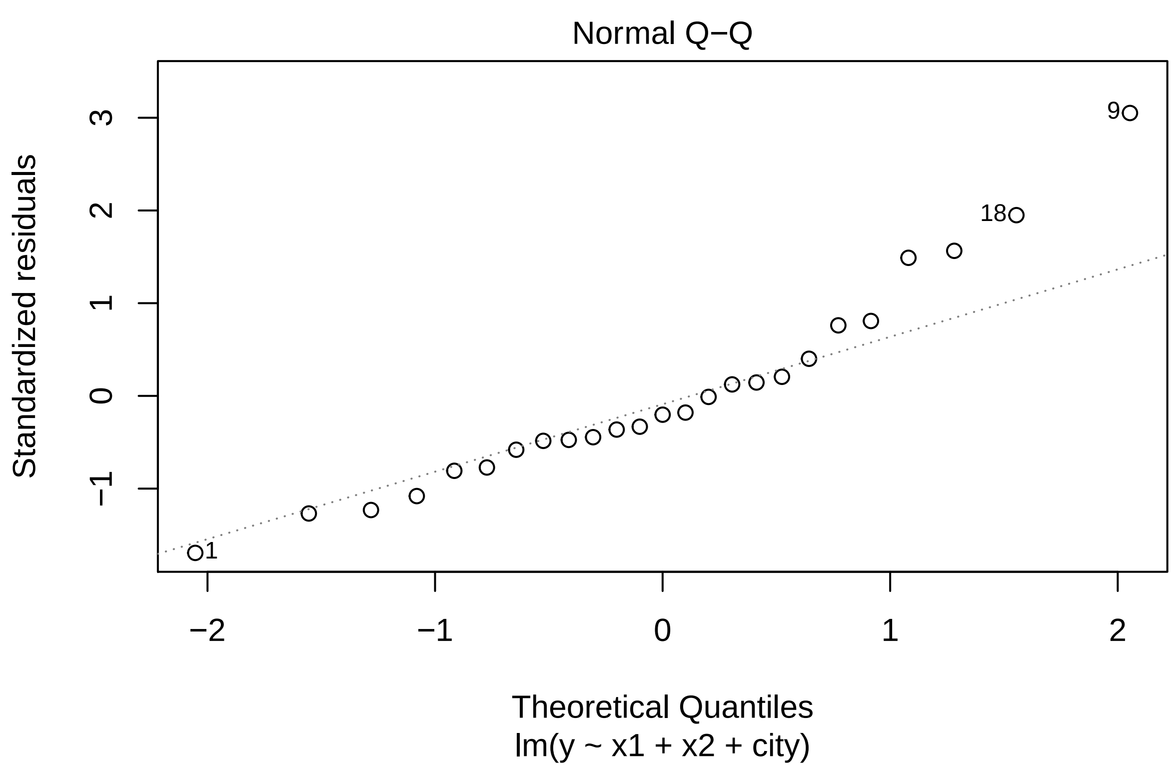
```
##
## Call:
## lm(formula = y ~ x1 + x2 + city, data = delivery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4800 -1.5922 -0.5583  1.1045  6.1611
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.036389   1.754025  -0.021  0.98366
## x1                1.770277   0.186790   9.477 1.24e-08 ***
## x2                0.010833   0.003786   2.862  0.00999 **
## cityBoston        4.190275   1.749048   2.396  0.02704 *
## cityMinneapolis   0.452636   2.687420   0.168  0.86803
## citySan Diego     2.737737   1.936269   1.414  0.17356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.986 on 19 degrees of freedom
## Multiple R-squared:  0.9707, Adjusted R-squared:  0.963
## F-statistic: 125.9 on 5 and 19 DF,  p-value: 6.919e-14
```

No, the delivery site is insignificant.

c. Analyze the residuals from this model. What conclusions can you draw regarding model adequacy?

```r
plot(fit_delivery, which=c(1,2))
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1 + x2 + city)
18
9
1
6
4
2
0
−2
−4
10
20
30
40
50
60
70


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + x2 + city)
9
18
1
3
2
1
0
−1
−2
−1
0
1
2

There is definitely a problem with normality.

#8.4 Consider the automobile gasoline mileage data in Table B.3

Dataset setup

```
gas <- table.b3
```

a. Build a linear regression model relating gasoline mileage y to engine displacement x_1 and the type of transmission $x_1 1$. Does the type of transmission significantly affect the mileage performance?

```
fit_gas <- lm(y ~ x1 + x11, data=gas)
summary(fit_gas)
```

```
##
## Call:
## lm(formula = y ~ x1 + x11, data = gas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9153 -1.8882  0.1106  1.7706  6.7829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.618408   1.539505  21.837  < 2e-16 ***
## x1          -0.045736   0.008682  -5.268  1.2e-05 ***
## x11         -0.498689   2.228198  -0.224    0.824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.115 on 29 degrees of freedom
## Multiple R-squared:  0.7727, Adjusted R-squared:  0.757
## F-statistic: 49.28 on 2 and 29 DF,  p-value: 4.696e-10
```

There does not seem to be a significant impact from type of transmission ($x_1 1$).

b. Modify the model developed in part a to include an interaction between engine displacement and the type of transmission. What conclusions can you draw about the effect of the type of transmission on gasoline mileage? Interpret the parameters in this model.

```
fit_gas_2 <- lm(y ~ x1 + x11 + x1 * x11, data=gas)
summary(fit_gas_2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x11 + x1 * x11, data = gas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2712 -1.2660  0.1342  1.5181  4.6599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  42.91963    2.73493  15.693 2.10e-15 ***
## x1           -0.11677    0.01984  -5.886 2.49e-06 ***
## x11         -13.46371    3.84413  -3.502 0.001567 **
## x1:x11        0.08165    0.02127   3.839 0.000647 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.566 on 28 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.8351
## F-statistic: 53.33 on 3 and 28 DF,  p-value: 1.064e-11
```

There does appear to be a increased influence of transmission when it is combined as an interaction. When using manual transmission there is a 0.117 decrease in MPG for every cubic inch of displacement. With an automatic transmission it is 0.035 ($x_1 + x_1 : x_{11} = -0.117 - 0.082$) lower MPGs.

#9.7 Consider the automobile gasoline mileage data in Table B.3

a. Does the correlation matrix give any indication of multicollinearity?

```r
cor(table.b3)
```

```
##              y         x1         x2 x3          x4         x5          x6
## y    1.0000000 -0.8787896 -0.8068797 NA  0.35654835  0.5947864 -0.48699716
## x1  -0.8787896  1.0000000  0.9452080 NA -0.33015370 -0.6315968  0.65906008
## x2  -0.8068797  0.9452080  1.0000000 NA -0.29205832 -0.5170425  0.77190992
## x3          NA         NA         NA  1          NA         NA          NA
## x4   0.3565483 -0.3301537 -0.2920583 NA  1.00000000  0.3737462 -0.04933889
## x5   0.5947864 -0.6315968 -0.5170425 NA  0.37374620  1.0000000 -0.20535194
## x6  -0.4869972  0.6590601  0.7719099 NA -0.04933889 -0.2053519  1.00000000
## x7   0.7220060 -0.7814778 -0.6431558 NA  0.49381043  0.8428620 -0.30057509
## x8  -0.7546216  0.8551981  0.7973892 NA -0.25810785 -0.5481227  0.42518809
## x9  -0.7731105  0.8013975  0.7176056 NA -0.31876434 -0.4343576  0.31567268
## x10 -0.8629267  0.9456621  0.8834004 NA -0.27721850 -0.5424247  0.52064243
## x11 -0.7450527  0.8354239  0.7266835 NA -0.36836123 -0.7032485  0.41733783
##             x7         x8         x9        x10        x11
## y    0.7220060 -0.7546216 -0.7731105 -0.8629267 -0.7450527
## x1  -0.7814778  0.8551981  0.8013975  0.9456621  0.8354239
## x2  -0.6431558  0.7973892  0.7176056  0.8834004  0.7266835
## x3          NA         NA         NA         NA         NA
## x4   0.4938104 -0.2581079 -0.3187643 -0.2772185 -0.3683612
## x5   0.8428620 -0.5481227 -0.4343576 -0.5424247 -0.7032485
## x6  -0.3005751  0.4251881  0.3156727  0.5206424  0.4173378
## x7   1.0000000 -0.6630802 -0.6682373 -0.7178265 -0.8549981
## x8  -0.6630802  1.0000000  0.8849771  0.9475859  0.6863079
## x9  -0.6682373  0.8849771  1.0000000  0.9015431  0.6507213
## x10 -0.7178265  0.9475859  0.9015431  1.0000000  0.7722283
## x11 -0.8549981  0.6863079  0.6507213  0.7722283  1.0000000
```

There appears to be a problem with multicollinearity due to some of the high values between independent variables.

b. Calculate the variance inflation factors and the condition number of $X'X$. Is there any evidence of multicollinearity?

```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11, data=table.b3))
```

```
##         x1         x2         x3         x4         x5         x6
## 119.487804  42.800811 149.234409   2.060036   7.729187   5.324730
##         x7         x8         x9        x10        x11
##  11.761341  20.917632   9.397108  85.744344   5.145052
```

There is evidence of multicollinearity as some of the VIFs are very high.