Statistics 206

Homework 6 Solution

Due : Nov. 16, 2015, In Class

1. *(Homework 4, Problem 4 Continued)* **Partial coefficients and added-variable plots.**

   A commercial real estate company evaluates age (X_1), operating expenses (X_2, in thousand dollar), vacancy rate (X_3), total square footage (X_4) and rental rates (Y, in thousand dollar) for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 *suburban commercial properties.* (The data is on smartsite under Resources/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

   (a) Perform regression of the rental rates Y on the four predictors X_1, X_2, X_3, X_4 (Model 1). (Hint: To help answer the subsequent questions, the predictors should enter the model in the order X_1, X_2, X_4, X_3.)

```
Call:
lm(formula = Y ~ X1 + X2 + X4 + X3, data = d)

Residuals:
Min      1Q  Median      3Q     Max
-3.1872 -0.5911 -0.0910  0.5579  2.9441

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
X3           6.193e-01  1.087e+00   0.570     0.57
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared:  0.5847,  Adjusted R-squared:  0.5629
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

   (b) Based on the R output of Model 1, obtain the fitted regression coefficient of X_3 and calculate the coefficient of partial determination $R^2_{Y3|124}$ and partial correlation $r_{Y3|124}$. Explain what $R^2_{Y3|124}$ measures and interpret the result.

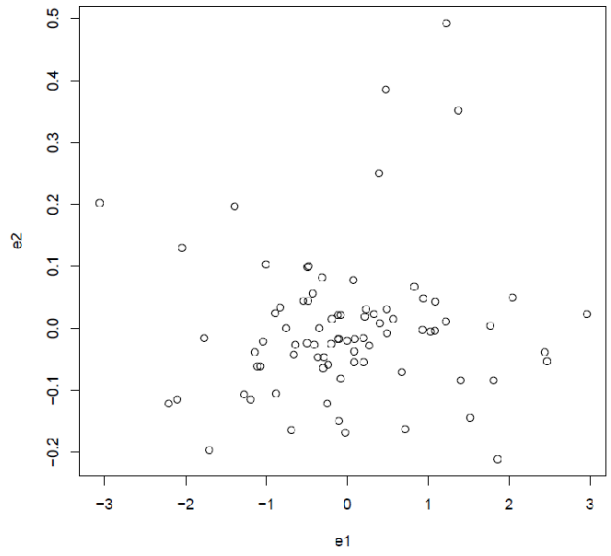
   From the R output in part (a), $\hat{\beta}_3 = 0.619$.
   The ANOVA is as follows:

```
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 14.819  14.819 11.4649  0.001125 **
X2         1 72.802  72.802 56.3262 9.699e-11 ***
X4         1 50.287  50.287 38.9062 2.306e-08 ***
X3         1  0.420   0.420  0.3248  0.570446
Residuals 76 98.231   1.293
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$$R^2_{Y3|124} = \frac{SSR(X_3|X_1, X_2, X_4)}{SSE(X_1, X_2, X_4)} = \frac{0.42}{98.231 + 0.42} = 0.42\%$$

This means adding X_3 to the model when X_1, X_2, X_4 is already in the model only reduces the SSE by 0.42%. The partial correlation is $r_{Y3|124} = \sqrt{0.0042} = 0.065$. The sign is positive because the estimated coefficient for X_3 is positive.

(c) Draw the added-variable plot for X_3 and make comments based on this plot.



From the plot we can conclude that after removing the effects of X_1, X_2 and X_4 from both Y and $X3$ they do not show any significant linear association. Hence adding $X3$ to the model will not be useful in helping explain Y.

(d) Regressing the residuals $e(Y|X_1, X_2, X_4)$ to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient

of X_3 from part (b). What do you find?

```
> r1=lm(Y~X1+X2+X4,data=data)   #regress Y on X1,X2,X4
> r2=lm(X3~X1+X2+X4,data=data)   #regress X3 on X1,X2,X4
> e1=residuals(r1)
> e2=residuals(r2)
> data=data.frame(e1=e1,e2=e2)
> fit2=lm(e1~e2,data=data)
> fit2$coefficients
```

The coefficients turn out to be:

```
> fit2$coefficients
(Intercept)            e2
-1.039391e-16  6.193435e-01
```

The fitted regression slope turns out to be same as the fitted coefficient from part (b).

(e) Obtain the regression sum of squares from part (d) and compare it with the extra sum of squares $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1. What do you find?

The ANOVA for part (d):

```
> anova(fit2)
Analysis of Variance Table

Response: e1
Df Sum Sq Mean Sq F value Pr(>F)
e2         1  0.420 0.41975  0.3376 0.5629
Residuals 79 98.231 1.24343
```

Hence the regression sum of squares from part (d) is 0.42 which is same as $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1.

(f) Calculate the correlation coefficient r between the two sets of residuals $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$. Compare it with $r_{Y3|124}$. What do you find? What is r^2?

```
> cor(e1,e2)
[1] 0.06522951
```

The correlation between $e1$ and $e2$ is the same as $r_{Y3|124}$.
The squared correlation between $e1$ and $e2$ is 0.004254889 which is same as $R^2_{Y3|124}$.

(g) Regressing Y to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find? Can you provide an explanation?

```
Call:
lm(formula = Y ~ e2, data = data)
```

```
Residuals:
Min      1Q  Median      3Q     Max
-4.7641 -1.1392 -0.1056  1.1221  4.1630

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1389     0.1921  78.807   <2e-16 ***
e2            0.6193     1.6528   0.375    0.709
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.729 on 79 degrees of freedom
Multiple R-squared:  0.001774,  Adjusted R-squared:  -0.01086
F-statistic: 0.1404 on 1 and 79 DF,  p-value: 0.7089
```

The fitted regression slope from this is same as that from part(b).
Let $\hat{Y}(X_1, X_2, X_4)$ be the fitted values from regressing Y on X_1, X_2, X_4. Then $Y = \hat{Y}(X_1, X_2, X_4) + e1$ where $e1 = e(Y|X_1, X_2, X_4)$. The fitted regression slope in this case is $\frac{Cov(Y,e2)}{Var(e2)}$ where $e2 = e(X_3|X_1, X_2, X_4)$.
However $Cov(Y, e2) = Cov(e1, e2)$ as $\hat{Y}(X_1, X_2, X_4)$ belongs to the span of (X_1, X_2, X_4) but $e2$ belongs to the orthogonal complement of this space. And so $\frac{Cov(Y,e2)}{Var(e2)} = \frac{Cov(e1,e2)}{Var(e2)}$.
The regression slope in part (d) is $\frac{Cov(e1,e2)}{Var(e2)}$ which is same as in part (b) as it captures the partial effect of X_3 on Y after adjusting for (X_1, X_2, X_4). Hence the two slopes are the same as the coefficient in part(b).

2. **Bias-variance trade-off**. Consider the following simulation study. You can modify the codes in bias-variance-trade-off-simulation.R under the smartsite/Resources/Homework. You should read the codes carefully and use R help whenever necessary to understand the codes.

   - The true regression function is
   $$f(x) = sin(x) + sin(2x).$$
   - The sample size is $n = 30$ and the design points X_i are equally spaced on $[-3, 3]$.
   - The models to be considered are polynomial regression models with order $l = 1, 2, 3, 5, 7, 9$.
   - The observed data are generated according to:
   $$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \cdots, n, \quad \epsilon_i \sim_{i.i.d.} N(0, \sigma^2).$$
   - Consider three different noise levels with $\sigma = 0.5, \quad 2, \quad 5$.

- Generate 1000 independent sets (replicates) of observations under each noise level.

Answer the following questions and include relevant plots along with your answers.

(a) Is there a correct model among the models being considered?Explain your answer.

ANS. There is no correct model in this case since $f(x)$ can not be represented by any finite order polynomials and all the models considered are finite order polynomials.

(b) What is the (in-sample) model variance for each of these models? Does the model variance change with the error variance?

ANS. The overall (in-sample) model variance Var_{in} for each of these models is $p\sigma^2$ where $p = 2, 3, 4, 6, 8, 9$ ($p = l + 1$). The model variance increases with error variance σ^2 (as well as p).

(c) Comment on the (in-sample) model bias for each of these models. Does the model bias change with the error variance?

ANS. The model bias does not change with error variance. However it does depend on l, the order of the polynomial. The overall in sample bias is determined by how well the column space of the design matrix approximates the mean response vector, so the models with higher polynomial orders have smaller model biases (as the column space of their respective design matrix is larger). In this case, for $l = 1, 2, 3$ we have large bias, for $l = 5$ the bias is small and for $l = 7, 9$ there is little bias.

(d) Which one, the model variance or the model bias, is the dominant component in the (in-sample) mean-squared-estimation-error? Does the answer depend on the error variance and why?

ANS. Recall the overall in sample $msee_{in} = Var_{in} + ||bias_{in}||_2^2$, where Var_{in} is the overall in sample variance which equals to $p\sigma^2$ and $||bias_{in}||_2^2$ is the overall in sample squared bias.

For $\sigma = 0.5$:, the squared-bias is more dominant than the model variance.
For $\sigma = 2$: the squared bias and the variance are on similar scale, and so it can be said that none of them is dominant.
For $\sigma = 5$: model variance is more dominant than squared bias.
The answer depends on error variance as model-variance increases with error variance. Larger error variance tends to make the model variance more dominant in msee.

(e) Which model is the best model according to the mean-squared-estimation-error? Does the answer depend on the error variance and why?
ANS. For $\sigma = 0.5 : l = 7$ is the best model in terms of overall in sample msee. Since for $\sigma = 0.5$ the squared bias is the dominant component in msee, so models with small bias (i.e., complex models) are preferred.
For $\sigma = 2 : l = 5$ is the best model in terms of msee. It achieves the optimal bias-variance trade-off.
For $\sigma = 5 : l = 1$ is the best model in terms of msee. Since for $\sigma = 5$ the model variance is the dominant component in msee , so models with small variance (i.e.,

simple models) are preferred.
The answer depends on error variance since the model variance increases with error variance and thus influences bias-variance trade-off.

(f) Comment on $E(SSE)$. Do you observe different patterns under different noise levels? Given an explanation.
Recall $E(SSE) = (n-p)\sigma^2 + ||bias_{in}||_2^2$.
ANS. For $\sigma = 0.5$: since squared bias is much larger than $(n-p)\sigma^2$ in underfitted models ($l = 1, 2, 3$), for these models $E(SSE)$ is much larger (in terms of relative magnitude) than $(n-p)\sigma^2$.
For $\sigma = 2$: since squared bias and $(n-p)\sigma^2$ are on comparable scales in underfitted models, for these models $E(SSE)$ is still considerably larger (in terms of relative magnitude) than $(n-p)\sigma^2$.
For $\sigma = 5$: since $(n-p)\sigma^2$ is much larger than the squared bias for all models, $E(SSE)$ is only slightly larger (in terms of relative magnitude) than $(n-p)\sigma^2$ for all models.
For a given noise level, $E(SSE)$ decreases with the increase of polynomial orders l as both $(n-p)$ decreases (note $p = l + 1$) and $||bias_{in}||_2^2$ decreases. On the other hand, for a given order l, $E(SSE)$ increases with the increase of noise level.

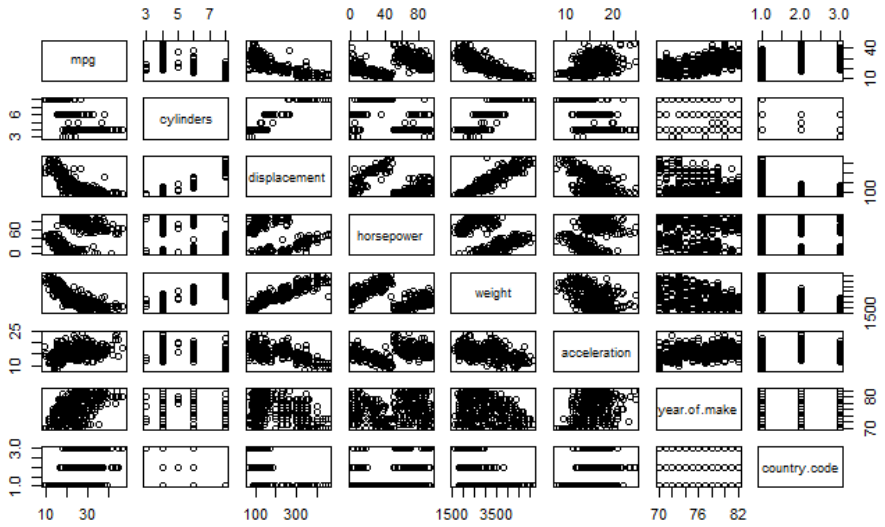
3. **Exploratory data analysis and preliminary investigation**. Read the data in "Car.csv" on smartsite/Resources/Homework into R:

```
cars = read.csv('Cars.csv', header=TRUE)
```

Consider building a model for "mpg" .

(a) Draw the scatterplot matrix of this data. Do you observe something unusual?

Figure 1: Scatter plot matrix of the data



One unusual fact is that scatter plots involving horsepower look weird with two clusters.

(b) Check the variable type for each variable. Do you observe something unusual? Which variables do you think should be treated as quantitative and which ones should be treated as qualitative/categorical?

```
mpg           cylinders    displacement    horsepower
"numeric"     "integer"     "numeric"      "factor"
weight      acceleration  year.of.make country.code
"integer"     "numeric"     "integer"     "integer"
```

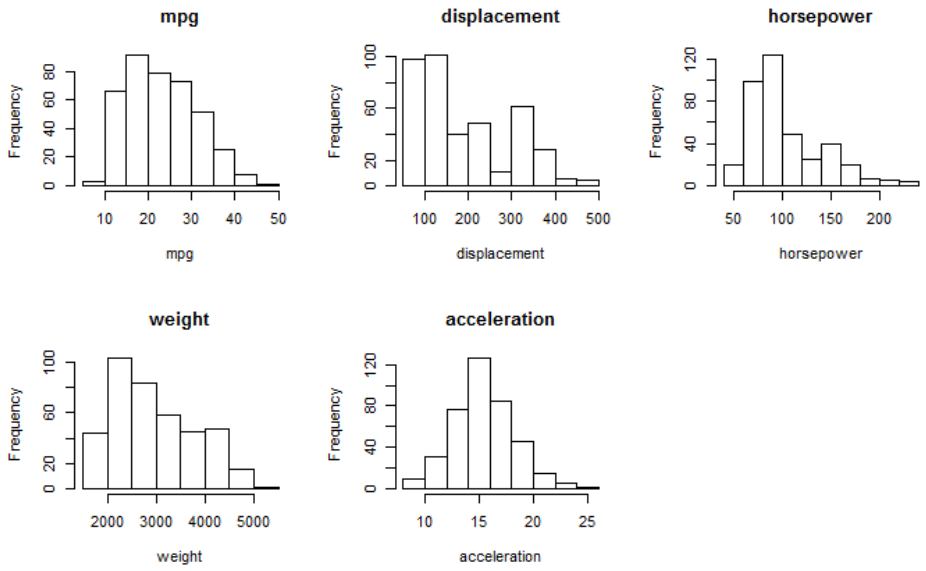
It is unusual that horsepower is a factor, it should have been numeric. (Probably because of the presence of "?" for some of the values).
Year of make is an ordinal variable.
Country code and cylinders should be treated as categorical variables and the rest as quantitave variables.

(c) Fix the problems that you have identified (if any) before proceeding to the next question. (Hint: Check out Lab5 handout)

All those observations which had horsepower="?" are removed. Country code and cylinders are converted to factors. Year of make is also not included in the model.

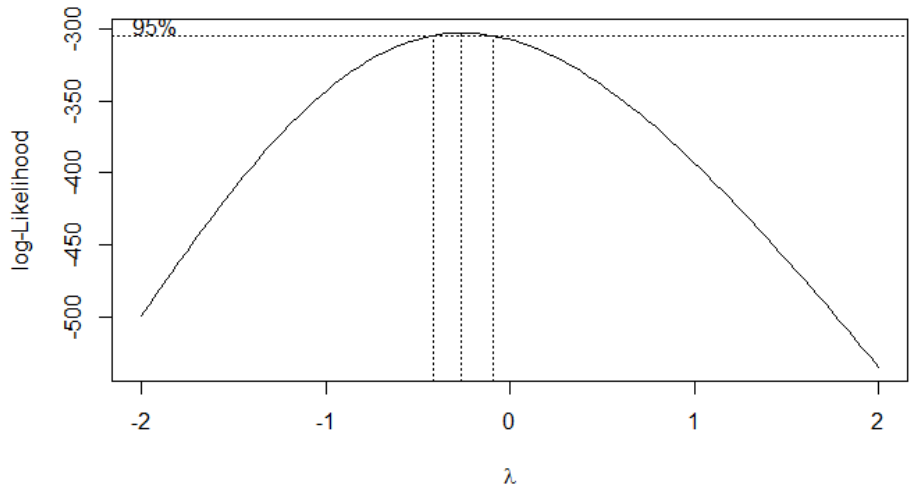
(d) Draw histogram for each quantitative variable. Do you think any transformation is needed ? If so, make the transformation before proceeding to the next question.

Figure 2: Histograms for quantitative variables



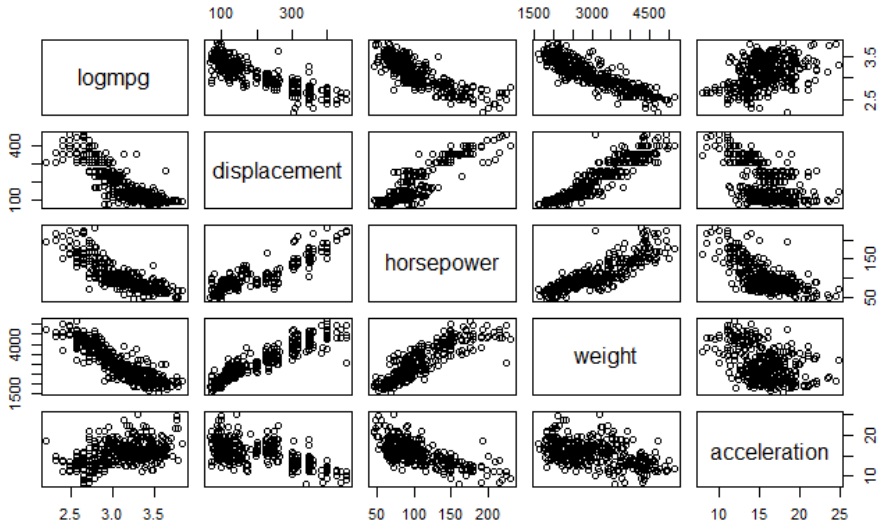
Here we need to make a transformation for some of the variables, particularly mpg since it is our response variable and we are making a model for it. From the box-cox plot we can see that $\lambda = 0$ (log-transformation) is a suitable transformation.

Figure 3: Box-Cox Plot for mpg



(e) Draw the scatter plot matrix among quantitative variables (possibly transformed). Do you observe any nonlinear relationship with the response variable? If so, what should you do?

Figure 4: Scatter Plot matrix among transformed quantitative variables



There appears to be a slight non-linear, in fact quadratic relationship in between log(mpg) and horsepower and displacement. In order to deal with this problem the second order term for each of these variables ahould also be in the model.

(f) Draw pie chart for each categorical variable. Draw side-by-side box plots for the response variable with respect to each categorical variable. What do you observe?

Figure 5: Pie chart for qualitative variables

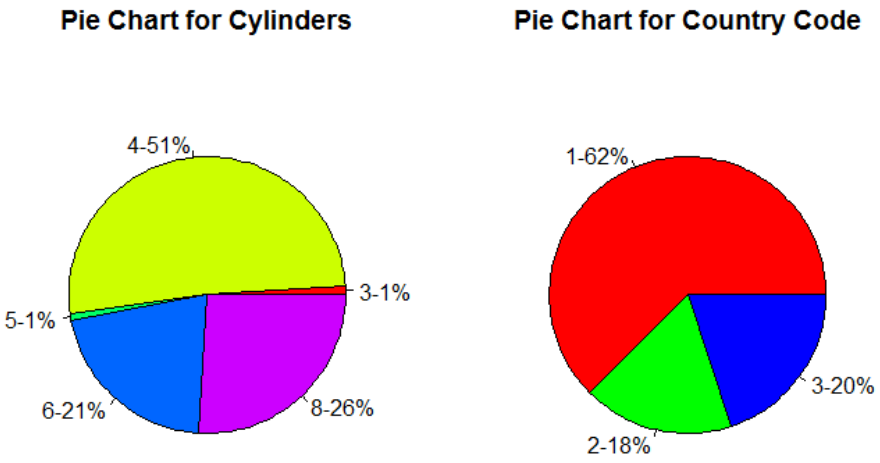


Figure 6: Multiple Box-Plot for Cylinders

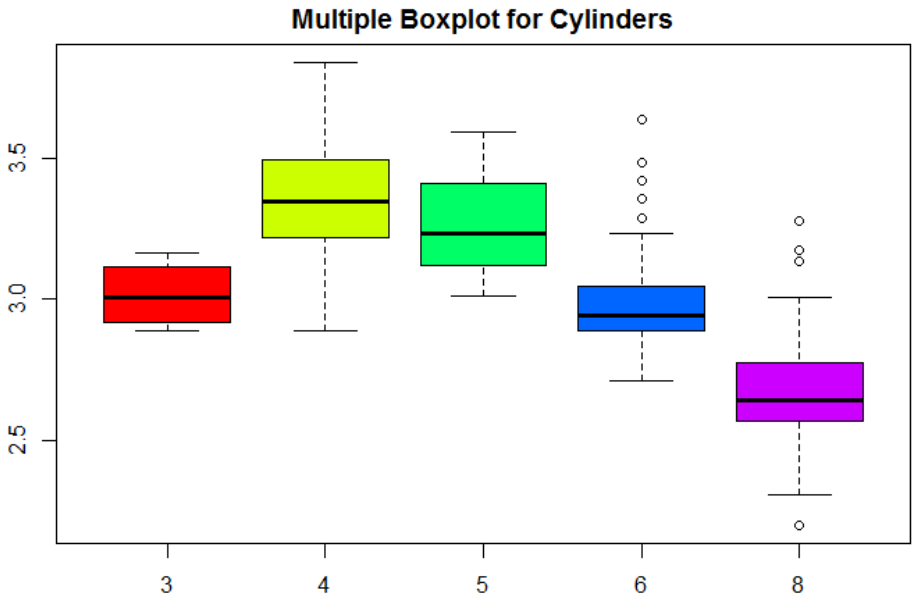
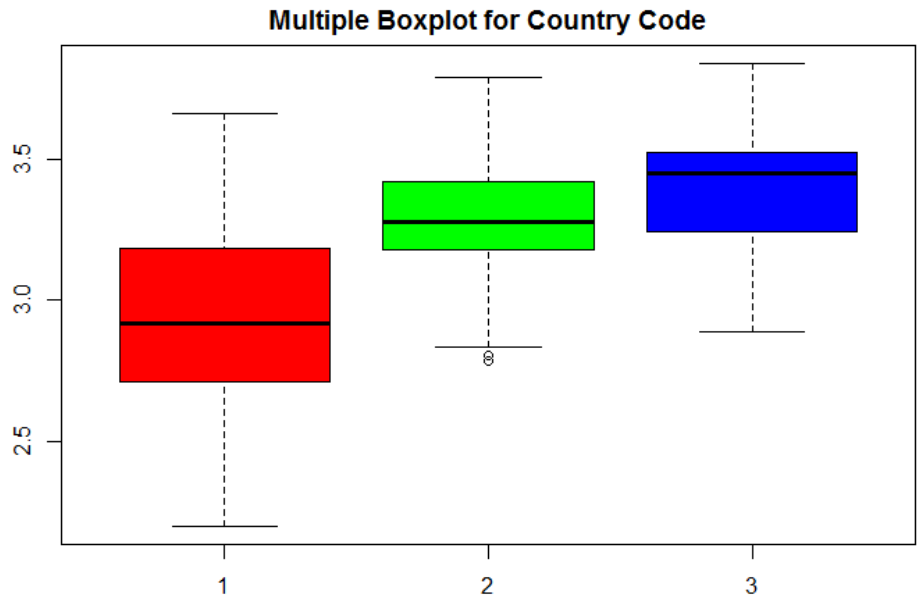


Figure 7: Multiple Box-Plot for Country Code



Cars with 3,6 and 8 cylinders tend to have lower mpg compared to others. Country code 3 appears to have higher mpg, followed by 2 followed by 1.

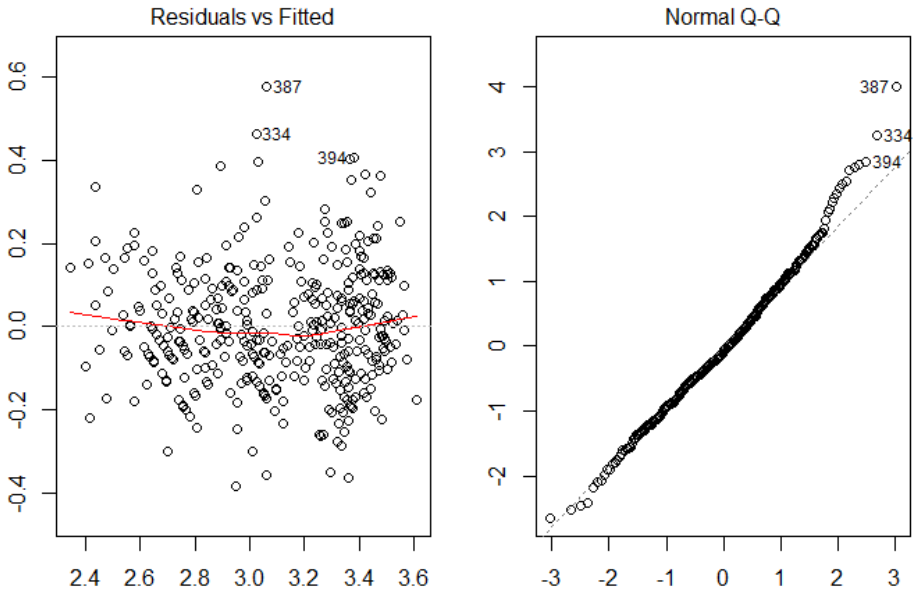
(g) Decide on a model for further investigation. Fit this model and draw residual plots. Does the model seem to be adequate? If not, try to make adjustments and fit an updated model. Repeat this process until you think you have found an adequate model. What would be your next step then? Here considering all the variables except year of make we have:

Models without quadratic term:

(1)
```
Call:
lm(formula = logmpg ~ cylinders + displacement + horsepower +
weight + acceleration + country.code, data = d)

Residuals:
Min       1Q   Median       3Q      Max
-0.38276 -0.08989 -0.00936  0.09033  0.57627

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.779e+00  1.154e-01  32.747  < 2e-16 ***
cylinders4     3.294e-01  7.818e-02   4.214 3.14e-05 ***
cylinders5     4.381e-01  1.187e-01   3.692 0.000255 ***
cylinders6     1.806e-01  8.639e-02   2.091 0.037183 *
cylinders8     2.046e-01  9.984e-02   2.050 0.041079 *
```

```
displacement   1.021e-04  3.399e-04   0.300 0.764096
horsepower    -3.624e-03  6.124e-04  -5.917 7.29e-09 ***
weight        -1.574e-04  2.931e-05  -5.369 1.38e-07 ***
acceleration  -8.397e-03  4.394e-03  -1.911 0.056718 .
country.code2 -8.913e-03  2.559e-02  -0.348 0.727815
country.code3  7.889e-02  2.499e-02   3.157 0.001722 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1468 on 381 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.8183,  Adjusted R-squared:  0.8136
F-statistic: 171.6 on 10 and 381 DF,  p-value: < 2.2e-16
```

Figure 8: Residual Plot and QQ Plot



Model 1.

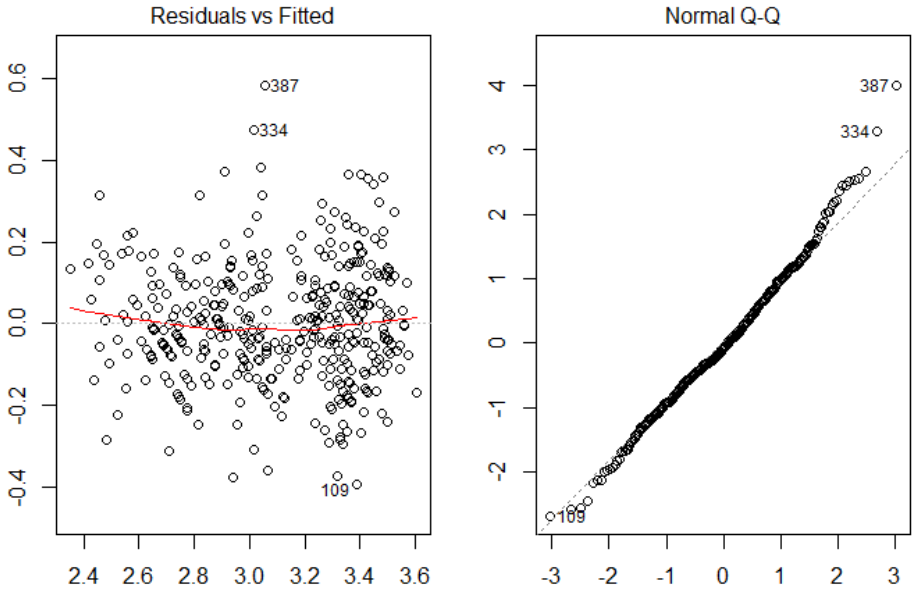
By examining the residual plots and the R^2, this model appears to be reasonable although there appears to be some nonlinearity in the residual vs. fitted value plot. Since the coefficients of displacement and acceleration are not significant at 0.05 level, we decide to try a model without these two terms.

(2)
```
Call:
lm(formula = logmpg ~ cylinders + horsepower + weight + country.code,
data = d)
```

```
Residuals:
Min       1Q   Median       3Q      Max
-0.39266 -0.08824 -0.01006  0.09397  0.58299

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.653e+00  9.619e-02  37.970  < 2e-16 ***
cylinders4     3.181e-01  7.606e-02   4.183 3.58e-05 ***
cylinders5     4.286e-01  1.165e-01   3.680 0.000266 ***
cylinders6     1.812e-01  7.927e-02   2.287 0.022766 *
cylinders8     2.196e-01  8.451e-02   2.599 0.009723 **
horsepower    -2.830e-03  4.444e-04  -6.367 5.52e-10 ***
weight        -1.787e-04  2.363e-05  -7.560 3.02e-13 ***
country.code2 -1.377e-02  2.424e-02  -0.568 0.570432
country.code3  7.391e-02  2.387e-02   3.096 0.002103 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1472 on 383 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.8165,  Adjusted R-squared:  0.8126
F-statistic:   213 on 8 and 383 DF,  p-value: < 2.2e-16
```

Model 2. By examining the residual plots and the R^2, this model appears to be reasonable (except for the nonlinearity in residual vs. fitted values plot). Particularly, R^2 does not decrease much on removing displacement and acceleration and now all the coefficients are significant at 0.05 level.

Figure 9: Residual Plot and QQ Plot



Models with quadratic term:

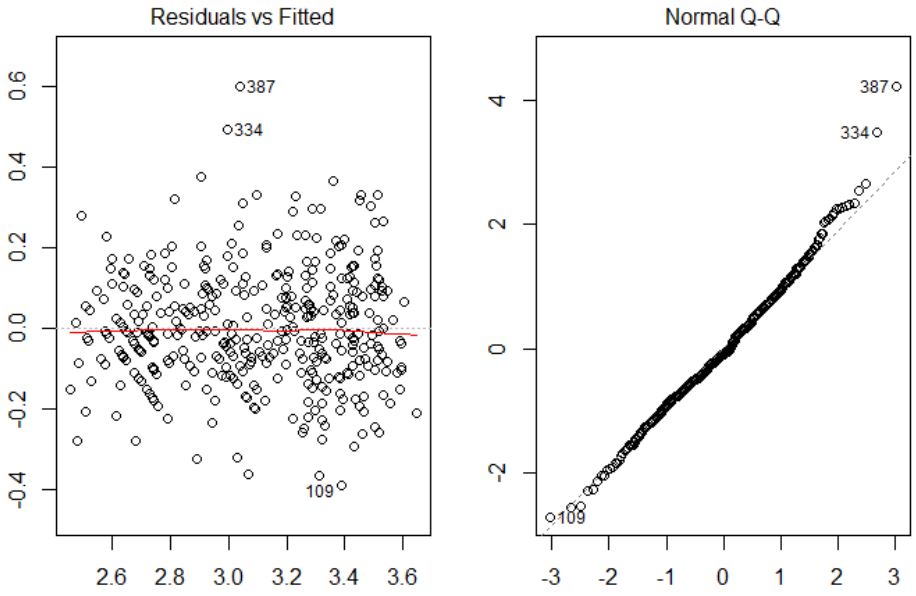
(3)
```
Call:
lm(formula = logmpg ~ cylinders + displacement + I(displacement^2) +
horsepower + I(horsepower^2) + weight + country.code, data = d)

Residuals:
Min       1Q   Median       3Q      Max
-0.39004 -0.09240 -0.01282  0.09155  0.59940

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.854e+00  1.178e-01  32.720  < 2e-16 ***
cylinders4         3.556e-01  8.035e-02   4.425 1.26e-05 ***
cylinders5         4.976e-01  1.224e-01   4.064 5.87e-05 ***
cylinders6         3.011e-01  9.705e-02   3.102  0.00206 **
cylinders8         3.307e-01  1.076e-01   3.073  0.00227 **
displacement      -2.477e-03  9.921e-04  -2.497  0.01296 *
I(displacement^2)  4.219e-06  1.702e-06   2.479  0.01360 *
horsepower        -4.251e-03  1.632e-03  -2.604  0.00957 **
I(horsepower^2)    3.835e-06  6.274e-06   0.611  0.54138
weight            -1.407e-04  2.793e-05  -5.037 7.31e-07 ***
country.code2     -4.475e-02  2.726e-02  -1.642  0.10148
country.code3      4.157e-02  2.703e-02   1.538  0.12489
---
```

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

Residual standard error: 0.145 on 380 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.8233,  Adjusted R-squared:  0.8181
F-statistic: 160.9 on 11 and 380 DF,  p-value: < 2.2e-16
```

Model 3.

By examining the residual plots and the R^2, this model appears to be adequate. Particularly, the nonlinearity in residual vs. fitted value plot is gone now. In this model, the quadratic term against displacement is significant at 5 % level while that of horse power is not. Also country code turns out to be not significant at 0.05 level. So we decided to try a model without these terms.

Figure 10: Residual Plot and QQ Plot



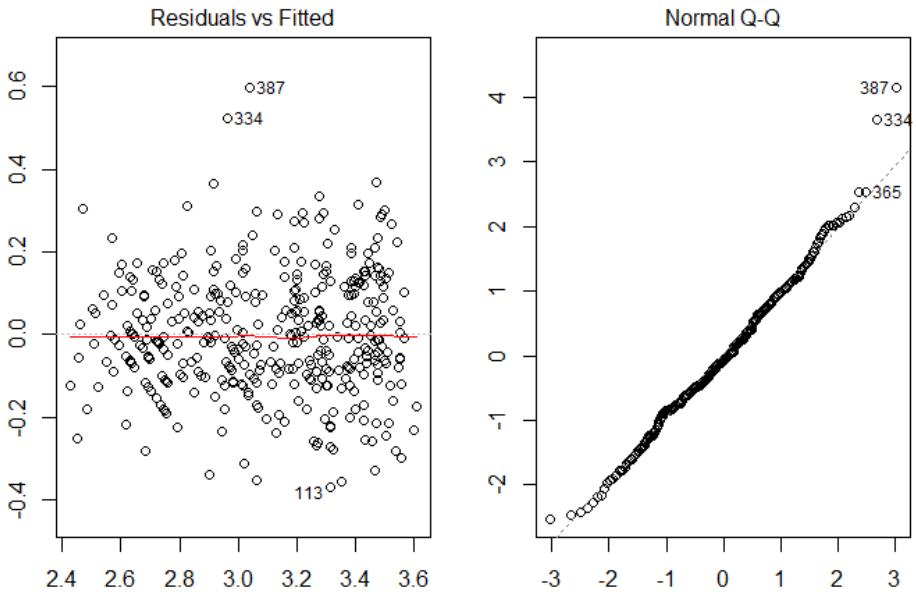
(4)
```
Call:
lm(formula = logmpg ~ cylinders + displacement + I(displacement^2) +
horsepower + weight, data = d)

Residuals:
Min       1Q   Median       3Q      Max
-0.36950 -0.09150 -0.01014  0.09760  0.59760

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)         3.890e+00  9.774e-02  39.797  < 2e-16 ***
cylinders4          3.257e-01  7.783e-02   4.184 3.55e-05 ***
cylinders5          4.369e-01  1.170e-01   3.734 0.000217 ***
cylinders6          2.834e-01  9.334e-02   3.036 0.002558 **
cylinders8          3.137e-01  1.048e-01   2.993 0.002943 **
displacement       -2.754e-03  7.423e-04  -3.710 0.000238 ***
I(displacement^2)   4.917e-06  1.199e-06   4.102 5.01e-05 ***
horsepower         -3.256e-03  5.030e-04  -6.473 2.94e-10 ***
weight             -1.567e-04  2.751e-05  -5.696 2.45e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1467 on 383 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.8176,  Adjusted R-squared:  0.8138
F-statistic: 214.6 on 8 and 383 DF,  p-value: < 2.2e-16
```

Model 4. By examining the residual plots and the R^2, this model appears to be adequate.

Particularly, R^2 does not decrease much on removing the horse power quadratic term or country code and now all coefficients are significant at 0.05 level.

Figure 11: Residual Plot and QQ Plot



This process results in several competing models (e.g., Model 3 and Model 4) that may turn out to be equally adequate. So the next step would be to perform model selection and model validation.