Statistics 206

Homework 7

Due : Nov. 20, 2019, In Class

1. Tell true or false of the following statements.

   (a) To quantify a qualitative variable with three classes C_1, C_2, C_3, we need the following indicator variables:

   $$X_1 = \begin{cases} 1 & if \quad C_1 \\ 0 & if \quad otherwise \end{cases} \quad X_2 = \begin{cases} 1 & if \quad C_2 \\ 0 & if \quad otherwise \end{cases} \quad X_3 = \begin{cases} 1 & if \quad C_3 \\ 0 & if \quad otherwise \end{cases}$$

   FALSE. We only need X_1 and X_2. C_3 is represented by both $X_1 = X_2 = 0$. Indeed, $X_1 + X_2 + X_3 \equiv 1$, so three of them are in perfect intercorrelation. If all three are included in a model, the LS estimators will not be defined.

   (b) Polynomial regression models with higher-order powers (e.g., higher than the third power) are preferred since they provide better approximations to the regression relation.

   FALSE. Polynomial regression models with higher-order powers could be highly variable and hard to generalize.

   (c) In interaction regression models, the effect of one variable depends on the value of another variable with which it appears together in a cross-product term.

   TRUE.

   (d) With a qualitative variable, the best way is to fit separate regression models under each of its classes.

   FALSE. This usually would not be as efficient as fitting one regression model using indicator variables due to loss of degrees of freedom (since each class will have a smaller sample size and more parameters are being fitted).

   (e) With many potential X variables, we can first fit a model with all these variables, and then drop those having non-significant regression coefficients by t-tests.

   FALSE. This would not work in presence of multicollinearity since then important X variables could be all dropped. This is because, with multicollinearity, T-tests for individual X variables could be all non-significant, yet there is a significant regression relation between the response variable and the set of X variables.

   (f) A correct model must be a good model.

   FALSE. A correct model is one has little bias, but it may have large variance.

   (g) With too many nuisance X variables, the model tends to have a large model bias.

   FALSE. A model with too many nuisance X variables tends to have large variance.

2. **(Commercial Property Cont'd) Partial coefficients and added-variable plots.**

(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.)

(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.

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

(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?

(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?

(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?

(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?

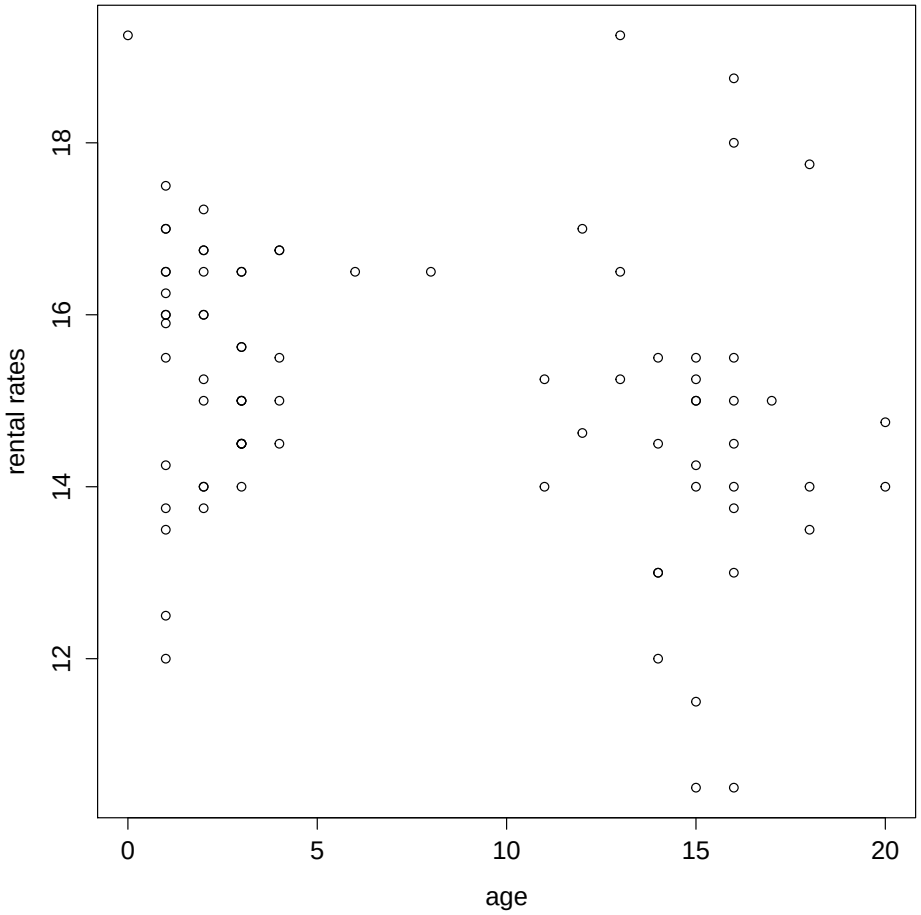
See the separate pdf file generated by RMarkdown.

3. **(Homework 5 Continued). Polynomial Regression.** You should use R and the *lm()* function and its associated functions (e.g., *summary(), anova(), confint(), predict.lm()*) to do this problem. Please also attach your R codes and 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 canvas under Files/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

Based on the analysis from Homework 5 the vacancy rate (X_3) is not important in explaining the rental rates (Y) when age (X_1), operating expenses (X_2) and square footage (X_4) are included in the model. So here we will use the latter three variables to build a regression for rental rates.

(a) Plot rental rates (Y) against the age of property (X_1) and comment on the shape of their relationship.



The age of property (X_1) exhibits some curvilinear relation when plotted against the rental rates (Y)

(b) Fit a polynomial regression model with linear terms for centered age of property ($\tilde{X}_1$), operating expenses (X_2), and square footage (X_4), and a quadratic term for centered age of property ($\tilde{X}_1$). Write down the model equation. Obtain the fitted regression function and also express it in terms of the original age of property X_1. Draw the observations Y against the fitted values $\widehat{Y}$ plot. Does the model provide a good fit?

Model equation:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_{i1} + \beta_2 X_{i2} + \beta_3 X_{i4} + \beta_4 \tilde{X}_{i1}^2, \quad i = 1, \cdots, 81$$

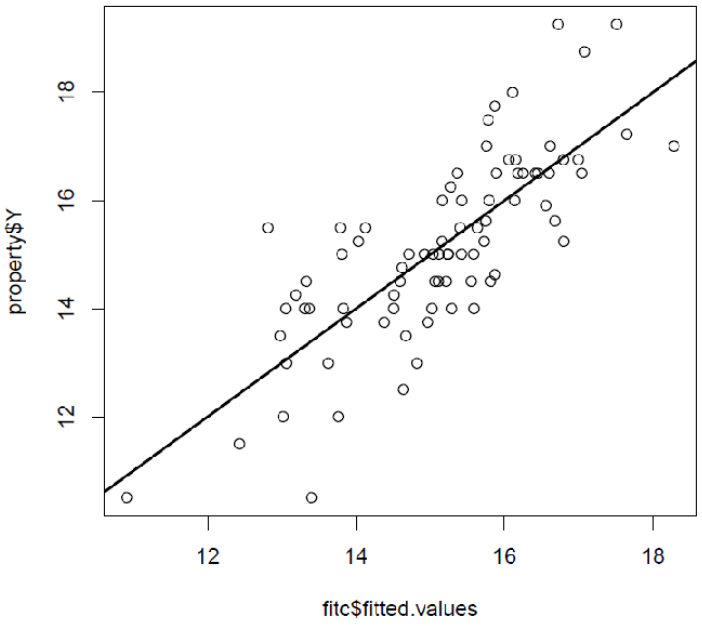
```
> summary(fitc)
```

```
Call:
lm(formula = Y ~ X1 + X2 + X4 + I(X1^2), data = property.c)

Residuals:
Min       1Q   Median       3Q      Max
-2.89596 -0.62547 -0.08907  0.62793  2.68309

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.019e+01  6.709e-01  15.188  < 2e-16 ***
X1          -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
X2           3.140e-01  5.880e-02   5.340 9.33e-07 ***
X4           8.046e-06  1.267e-06   6.351 1.42e-08 ***
I(X1^2)      1.415e-02  5.821e-03   2.431   0.0174 *
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 1.097 on 76 degrees of freedom
Multiple R-squared: 0.6131,Adjusted R-squared: 0.5927
F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

Fitted regression function:

$$\begin{aligned}\widehat{Y} &= 10.19 - 0.1818\tilde{X}_1 + 0.314X_2 + 8.046 \times 10^{-6}X_4 + 0.01415\tilde{X}_1^2 \\ &= 10.19 - 0.1818(X_1 - 7.8642) + 0.314X_2 + 8.046 \times 10^{-6}X_4 + 0.01415(X_1 - 7.8642)^2.\end{aligned}$$

The model provides a fairly good fit.



(c) Compare R^2, R_a^2 of the above model with those of Model 2 from Homework 5 ($Y \sim X_1 + X_2 + X_4$). What do you find?

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

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 14.819  14.819 12.3036 0.0007627 ***
X2         1 72.802  72.802 60.4463 2.968e-11 ***
X4         1 50.287  50.287 41.7522 8.907e-09 ***
I(X1^2)    1  7.115   7.115  5.9078 0.0174321 *
Residuals 76 91.535   1.204
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
```

For the above model: $R^2 = 0.6131$, $R_a^2 = 0.5927$.
For Model 2 from homework 5: $R^2 = 0.583$, $R_a^2 = 0.5667$. So the model here has a better fit of the data than Model 2 of homework 5.

(d) Test whether or not the quadratic term for centered age of property ($\tilde{X}_1$) may be dropped from the model at level 0.05. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

Null and alternative hypotheses:

$$H_0 : \beta_4 = 0, \ vs. \ H_a : \beta_4 \neq 0$$

Test statistic:

$$T^* = \frac{\hat{\beta}_4}{se(\hat{\beta}_4)} = 2.431$$

Under H_0, $T^* \sim t_{76}$, Since $2.431 > 1.99 = t(0.975; 76)$(or p-value=0.0174 < 0.05), reject H_0 and conclude that the quadratic term for centered age of property can not be dropped.

(e) Predict the rental rates for a property with $X_1 = 4, X_2 = 10, X_4 = 80,000$. Construct a 99% prediction interval and compare it with the prediction interval from Model 2 of Homework 5.

```
> newX=data.frame(X1=4-mean(property$X1),X2=10,X4=80000)
> predict.lm(fitc, newX, interval="prediction", level=0.99, se.fit=TRUE)
$fit
fit      lwr      upr
1 14.88699 11.93875 17.83524

$se.fit
[1] 0.201945

$df
[1] 76

$residual.scale
[1] 1.097455
```

Recall from homework 5, the prediction interval given by Model 2 is (12.09134, 18.14836) with the fitted value (center of the interval) being 15.11985. The current interval is likely to be less biased due to the inclusion of the quadratic term.

Moreover, the above prediction interval is slightly narrower than the one from Model 2, which is due to smaller MSE of the current model (1.204 here vs. 1.281 of Model 2). The SE of the fitted value is actually slightly larger in the current Model compared with that of Model 2 (0.201945 vs. 0.1833524). But this is more than compensated for by the smaller MSE for the prediction SE:

$$s(pred) = \sqrt{s^2(fitted) + MSE}.$$

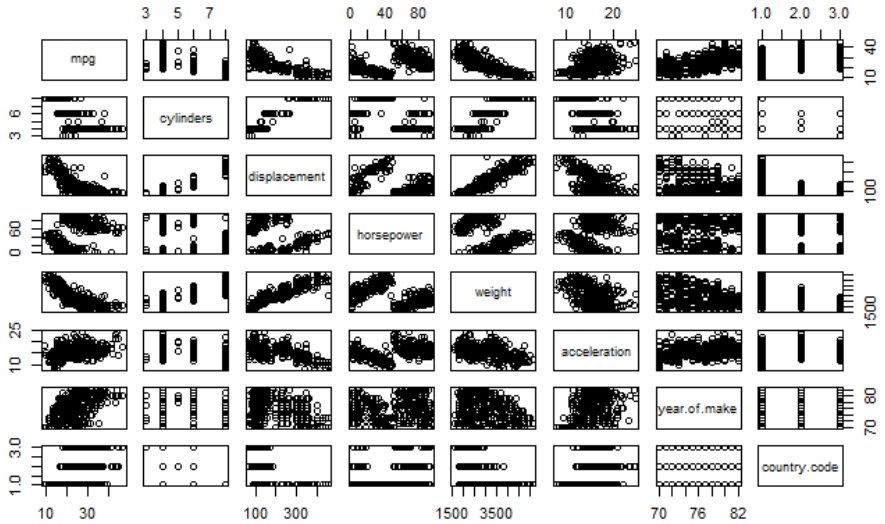
4. **Exploratory data analysis and preliminary investigation**. Read the data in "Car.csv" on canvas/Files/Homework into R:

```
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



Unusual observations: scatter plots involving horsepower look weird with two clusters; scatter plots involving cylinders and country code look like dots located on lines.

(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"
```

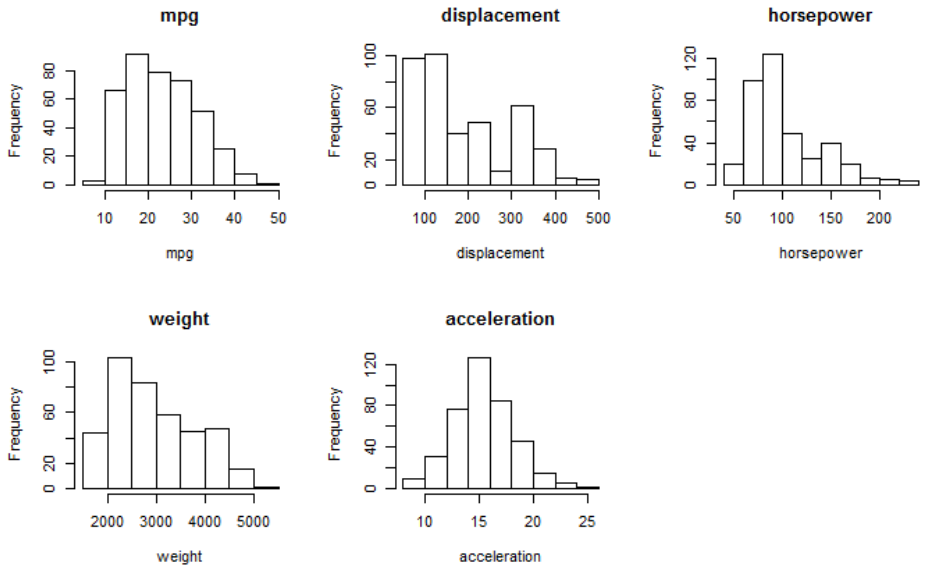
It is unusual that horsepower is a factor, it should have been numeric. (Probably because of the presence of "?" for some of the values).
Year of make is an ordinal variable.
Country code and cylinders should be treated as categorical variables and the rest as quantitative variables.

(c) Fix the problems that you have identified (if any) before proceeding to the next question.

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.

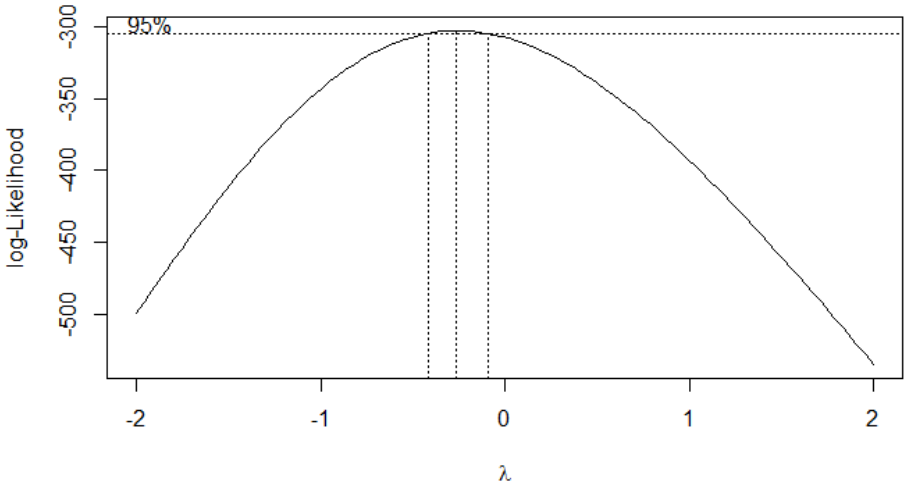
(d) Draw histogram for each quantitative variable. Do you think any transformation is needed ? If so, make the transformation before proceeding to the next question.

Figure 2: Histograms for quantitative variables



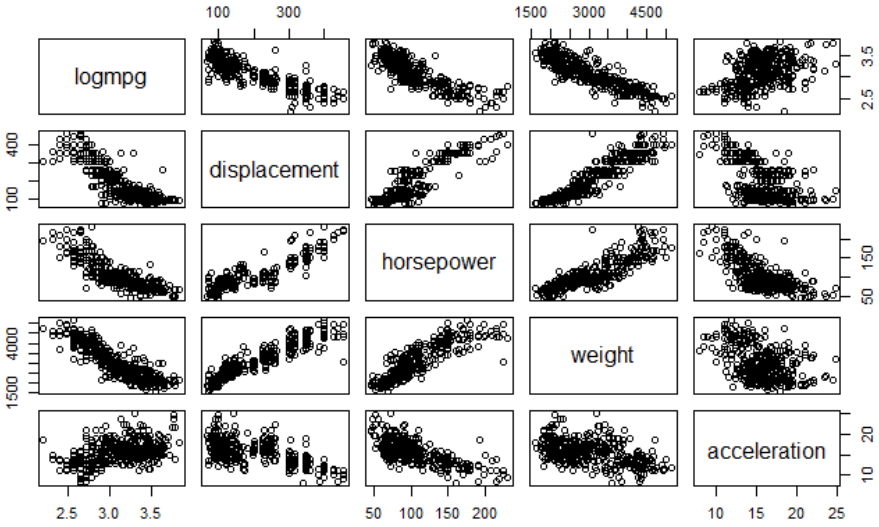
Here we need to make a transformation for some of the variables, particularly mpg since it is our response variable and we are making a model for it. From the box-cox plot we can see that $\lambda = 0$ (log-transformation) is a suitable transformation.

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 should 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

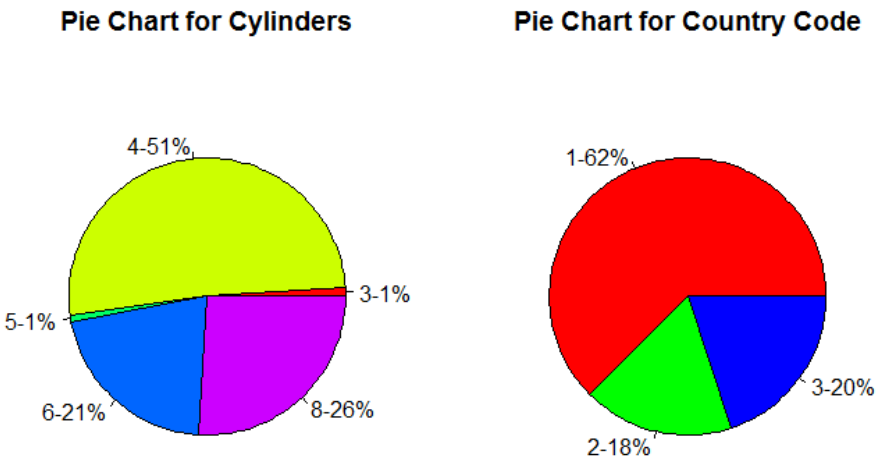


Figure 6: Multiple Box-Plot for Cylinders

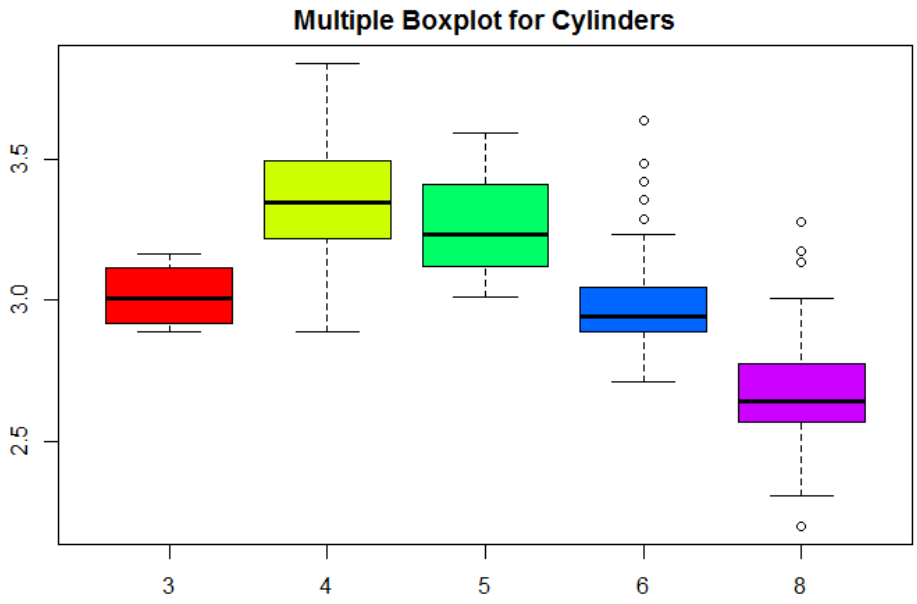
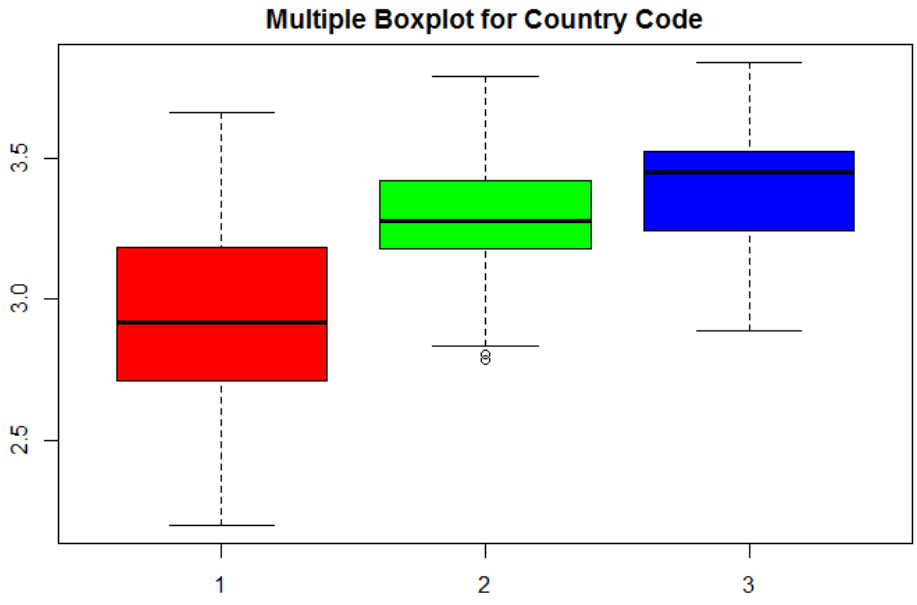


Figure 7: Multiple Box-Plot for Country Code



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.

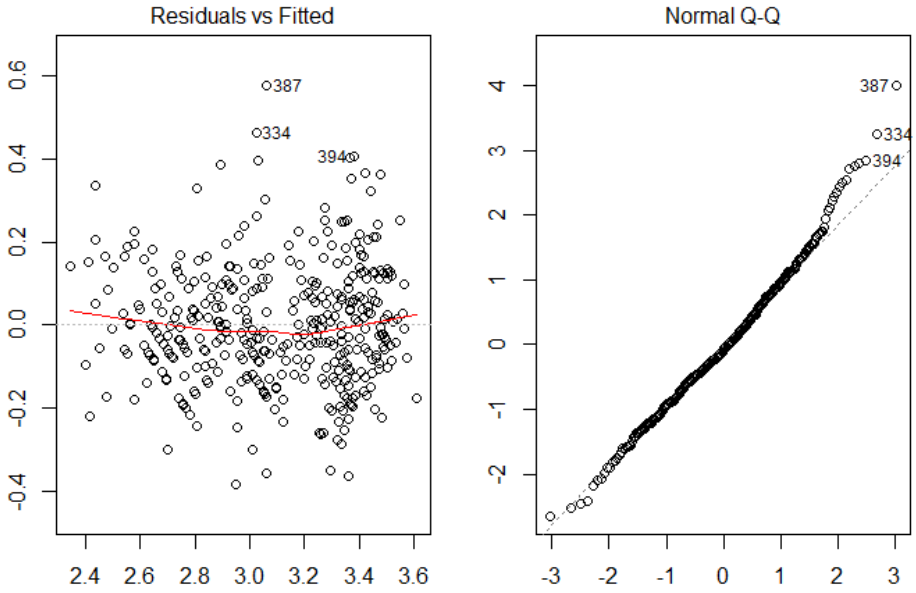
(g) Decide on a model for further investigation. Fit this model and draw residual plots. Does the model seem to be adequate? If not, try to make adjustments and fit an updated model. Repeat this process until you think you have found an adequate model. What would be your next step then?

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.

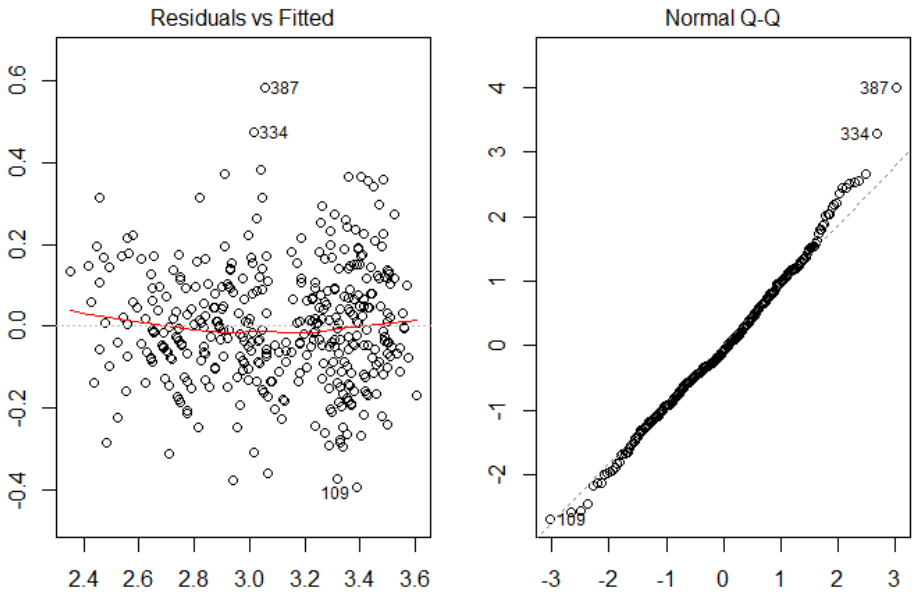
By examining the residual plots and the R^2, this model appears to be reasonable although there appears to be some nonlinearity in the residual vs. fitted value plot. Since the coefficients of displacement and acceleration are not significant at 0.05 level, we decide to try a model without these two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

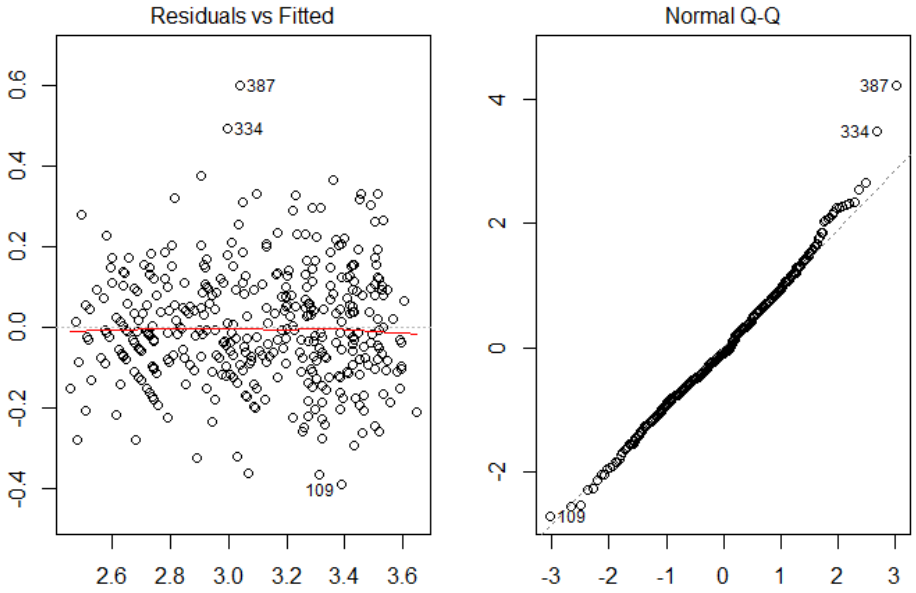
(3)
```
Call:
lm(formula = logmpg ~ cylinders + displacement + I(displacement^2) +
horsepower + I(horsepower^2) + weight + country.code, data = d)

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



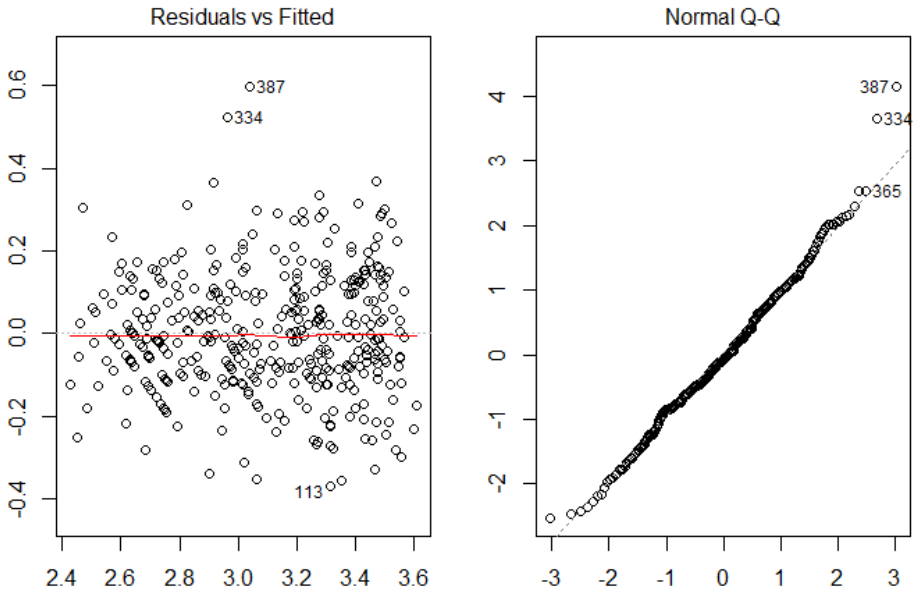
(4)
```
Call:
lm(formula = logmpg ~ cylinders + displacement + I(displacement^2) +
horsepower + weight, data = d)

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.

5. **Polynomial Regression**. Write down model equations for the following models.

   (a) A third-order polynomial regression model with one predictor.

   Model equation:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \beta_3 \tilde{X}_i^3 + \epsilon_i, \quad i = 1, 2, .., n$$

   where $\tilde{X}_i = X_i - \bar{X}$.

   (b) A second-order polynomial regression model with K predictors.

   Model equation:

$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \tilde{X}_{ik} + \sum_{k=1}^{K} \beta_{kk} \tilde{X}_{ik}^2 + \sum_{1 \le k \ne k' \le K} \beta_{kk'} \tilde{X}_{ik} \tilde{X}_{ik'} + \epsilon_i, i = 1, 2, .., n$$

   where $\tilde{X}_{ik} = X_{ik} - \bar{X}_k$.

6. **Bias-variance trade-off**. Consider the following simulation study. You can modify the codes in bias-variance-trade-off-simulation2.R under the canvas/Files/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.

7. **(Optional problem). Simultaneous confidence bands of the regression function.** Under the Normal error model, derive the simultaneous confidence bands of the regression function by the following steps.

(a) Show that

$$\frac{(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta)}{MSE} \sim pF_{p,n-p}.$$

Proof. Under the normal error model, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$. Thus,

$$(X^TX)^{1/2}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \sigma^2 I),$$

and

$$(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta) \sim \sigma^2 \chi^2_p.$$

□

From the lecture notes, we know that $pMSE \sim \sigma^2 \chi^2_{n-p}$. Since

$$\begin{aligned} \text{Cov}(\hat{\beta}, (I-H)Y) &= \text{Cov}(HY, (I-H)Y) \\ &= 0, \end{aligned}$$

and $\hat{\beta}$ and $(I-H)Y$ are jointly normal, $\hat{\beta}$ and $(I-H)Y$ are independent. Thus,

$$\frac{(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta)}{MSE} \sim pF_{p,n-p}.$$

(b) Show that for a constant $C \geq 0$, $|x^T\beta - x^T\hat{\beta}| \leq \sqrt{Cx^T(X^TX)^{-1}x}$ for all $x \in \mathbb{R}^p$ if and only if $(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta) \leq C$.

Proof. Let $y = (X^TX)^{-1/2}x$. Using CauchySchwarz inequality, we have

$$\begin{aligned} |x^T\beta - x^T\hat{\beta}|^2 &= |y^T(X^TX)^{1/2}(\hat{\beta} - \beta)|^2 \\ &\leq (y^Ty)(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta) \\ &= x^T(X^TX)^{-1}x(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta). \end{aligned}$$

Since the equality can be achieved, this suggests the equivalency. □

(c) Show that the $(1-\alpha)$ simultaneous confidence bands for the regression function, $x^T\beta, x \in \mathbb{R}^p$, are:

$$x^T\hat{\beta} \pm \sqrt{pF(1-\alpha; p, n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \quad x \in \mathbb{R}^p,$$

i.e.,

$$P(x^T\beta \in x^T\hat{\beta} \pm \sqrt{pF(1-\alpha; p, n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \text{ for all } x \in \mathbb{R}^p) = 1-\alpha.$$

Proof. To show the probability equals $1-\alpha$.

$$\begin{aligned} LHS &= P(|x^T\beta \\ &\qquad - x^T\hat{\beta}| \leq \sqrt{pF(1-\alpha; p, n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \quad \text{for all } x) \\ &= P((\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta) \leq pF(1-\alpha; p, n-p)MSE) \quad \text{Using (b)} \\ &= P(\frac{(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta)}{MSE} \leq pF(1-\alpha; p, n-p)) \\ &= P(F_{p,n-p} \leq F(1-\alpha; p, n-p)) \quad \text{Using (a)} \\ &= 1-\alpha. \end{aligned}$$

□

8. **(Optional problem). Regression coefficients as partial coefficients.** Let $X = (X_1, X_2)$ where $X_1 \in \mathbb{R}^{n\times s}, X_2 \in \mathbb{R}^{n\times t}$. Write the LS fitted regression coefficients as $\hat{\beta} = \begin{pmatrix} \hat{\beta}^{(1)} \\ \hat{\beta}^{(2)} \end{pmatrix}$. Show that:

   (a) The LS fitted regression coefficients of X_2 is

   $$\hat{\beta}^{(2)} = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T (Y - \hat{Y}(X_1)), \quad \tilde{X}_2 = X_2 - \hat{X}_2(X_1),$$

   i.e., $\hat{\beta}^{(2)}$ **is the LS fitted regression coefficients by regressing** Y **(or** $Y - \hat{Y}(X_1)$**) onto** $X_2 - \hat{X}_2(X_1)$**.** Such coefficients are called **partial coefficients.**

   Proof. Since $X_2 - \hat{X}_2(X_1)$ is orthogonal to $span\{X_1\}$, and $\hat{Y}(X_1)$ is in the space of $span\{X_1\}$, we can show that $\tilde{X}_2^T \hat{Y}(X_1) = 0$. This tells us that $(\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T (Y - \hat{Y}(X_1))$.
   Now, to show $\hat{\beta}^{(2)} = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y$. We know

   $$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix}.$$

   Denote

   $$(X^T X)^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

   where $B_{22} = (X_2^T X_2 - X_2^T H(X_1) X_2)^{-1}$, and $B_{21} = -B_{22} X_2^T X_1 (X_1^T X_1)^{-1}$. Since $\hat{\beta} = (X^T X)^{-1} X^T Y$, we have

   $$\begin{aligned} LHS &= B_{21} X_1^T Y + B_{22} X_2^T Y \\ &= B_{22}(X_2^T - X_2^T X_1 (X_1^T X_1)^{-1} X_1^T) Y \\ &= B_{22} X_2^T (I - H(X_1)) Y \\ &= (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y \\ &= RHS. \end{aligned}$$

   □

   (b) If $X_1 \perp X_2$ (i.e., the columns of X_1 and the columns of X_2 are orthogonal), then

   $$\hat{\beta}^{(2)} = (X_2^T X_2)^{-1} X_2^T Y, \quad if, \quad X_1 \perp X_2,$$

   i.e., the LS fitted regression coefficients by regressing Y onto X_2 alone.

   Proof. $X_1 \perp X_2$ indicates that $X_1^T X_2 = 0$. Thus,

   $$(X^T X)^{-1} = \begin{pmatrix} (X_1^T X_1)^{-1} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{pmatrix}.$$

   Plug it into the expression of $\hat{\beta}$, we can obtain

   $$\hat{\beta}^{(2)} = (X_2^T X_2)^{-1} X_2^T Y.$$

   □