**Project #2**

**by**

**Dipika Mahashabde**

**(Cwid:803004274)**

**CPSC 483**

**Professor: Kenytt Avery**



**Department of Computer Science**

**California State University, Fullerton**

**Spring 2016**

Part 1

Installed R

Part 2

1. Transcripts are saved in one file attached in this zip file.

2. Plots and contours are saved in pdf file.

3. Auto Data:

<http://www-bcf.usc.edu/~gareth/ISL/Auto.data>

Screenshot: In the Zip file

Part 3 (Ex- chapter 3)

8. (a). Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.

**Answer (a):**

Command:-

>lm.fit3=lm(mpg~horsepower,data=Auto)

>summary(lm.fit3)

|  |
| --- |
| Call:  lm(formula = mpg ~ horsepower, data = Auto)  Residuals:  Min 1Q Median 3Q Max  -13.5710 -3.2592 -0.3435 2.7630 16.9240  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.906 on 390 degrees of freedom  Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049  F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16 |

(I). Is there a relationship between the predictor and the response?

**Answer**: After doing this we get to know that there is a relationship between the response and predictor. When we check if the there is a relation present or no we do it by checking the hypotenuse, if the value is near 1 the relationship is not good. We discard the parameter but here we see that the value is far from 1 so there exist relation between response and predictor.

(II). How strong is the relationship between the predictor and the response?

**Answer:** R-squared is 0.6059 in percentage is 60.59% stronger.

(III). Is the relationship between the predictor and the response positive or negative?

**Answer:** The relation between the predictor and the response is negative because Coefficient of horsepower is negative. So when the horsepower increases the mpg reduces.

(IV). What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

**Answer:**

1. Confidence intervals:

Command:-

> predict(lm.fit3,data.frame(horsepower=c(98)),interval = "confidence")

|  |
| --- |
| fit lwr upr  1 24.46708 23.97308 24.96108 |

2. Prediction intervals:

Command:-

> predict(lm.fit3,data.frame(horsepower=c(98)),interval = "prediction")

|  |
| --- |
| fit lwr upr  1 24.46708 14.8094 34.12476 |

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

**Answer:** Attached in Ex\_pdfs named as a8b.pdf

Command:

pdf("a8b.pdf")

attach(Auto)

plot(horsepower,mpg)

abline(lm.fit3)

dev.off()

(C) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

**Answer:** Attached in Ex\_pdfs named as a8c.pdf

**Comment:**

1. Yes there is an unusual large outlier Non-linear graph in the residual plot.
2. Yes there is an unusual high leverage plot which is point no 94 and 117 even when his standardized residual is less.

9. This question involves the use of multiple linear regression on the Auto data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

**Answer:** Attached in Ex\_pdfs named as a9a.pdf

(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

**Answer:**

Command:

> cor(Auto[1:8])

|  |
| --- |
| mpg cylinders displacement horsepower  mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268  cylinders -0.7776175 1.0000000 0.9508233 0.8429834  displacement -0.8051269 0.9508233 1.0000000 0.8972570  horsepower -0.7784268 0.8429834 0.8972570 1.0000000  weight -0.8322442 0.8975273 0.9329944 0.8645377  acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955  year 0.5805410 -0.3456474 -0.3698552 -0.4163615  origin 0.5652088 -0.5689316 -0.6145351 -0.4551715  weight acceleration year origin  mpg -0.8322442 0.4233285 0.5805410 0.5652088  cylinders 0.8975273 -0.5046834 -0.3456474 -0.5689316  displacement 0.9329944 -0.5438005 -0.3698552 -0.6145351  horsepower 0.8645377 -0.6891955 -0.4163615 -0.4551715  weight 1.0000000 -0.4168392 -0.3091199 -0.5850054  acceleration -0.4168392 1.0000000 0.2903161 0.2127458  year -0.3091199 0.2903161 1.0000000 0.1815277  origin -0.5850054 0.2127458 0.1815277 1.0000000 |

(c) Use the lm() function to perform a multiple linear regression

with mpg as the response and all other variables except name as

the predictors. Use the summary() function to print the results.

Comment on the output. For instance:

**Answer:**

Command:

> lm.fit4=lm(mpg~.-name,data=Auto)

> summary(lm.fit4)

|  |
| --- |
| Call:  lm(formula = mpg ~ . - name, data = Auto)  Residuals:  Min 1Q Median 3Q Max  -9.5903 -2.1565 -0.1169 1.8690 13.0604  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*  cylinders -0.493376 0.323282 -1.526 0.12780  displacement 0.019896 0.007515 2.647 0.00844 \*\*  horsepower -0.016951 0.013787 -1.230 0.21963  weight -0.006474 0.000652 -9.929 < 2e-16 \*\*\*  acceleration 0.080576 0.098845 0.815 0.41548  year 0.750773 0.050973 14.729 < 2e-16 \*\*\*  origin 1.426141 0.278136 5.127 4.67e-07 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.328 on 384 degrees of freedom  Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182  F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16 |

I. Is there a relationship between the predictors and the response?

**Answer:** Yes there is a relationship between the response and predictor. 82.15%

II. Which predictors appear to have a statistically significant relationship to the response?

**Answer:** Cylinder, Horsepower, Acceleration are not significant to the response.

III. What does the coefficient for the year variable suggest?

**Answer:** Year variable suggest that the mpq increases 0.75 times every year.

(d) Use the plot() function to produce diagnostic plots of the linear

regression fit. Comment on any problems you see with the fit.

Do the residual plots suggest any unusually large outliers? Does

the leverage plot identify any observations with unusually high

leverage?

**Answer:**

1. Yes there is an unusual large outlier Non-linear graph in the residual plot.
2. Yes there is an unusual high leverage plot which is point no 14 even when his standardized residual is less.

(e) Use the \* and : symbols to fit linear regression models with

interaction effects. Do any interactions appear to be statistically

significant?

**Answer:**

Command:

Auto2=Auto[,1:8]

lm.fit6=lm(mpg~.\*.,data=Auto2)

summary(lm.fit6)

|  |
| --- |
| Call:  lm(formula = mpg ~ . \* ., data = Auto2)  Residuals:  Min 1Q Median 3Q Max  -7.6303 -1.4481 0.0596 1.2739 11.1386  Coefficients:  Estimate Std. Error t value  (Intercept) 3.548e+01 5.314e+01 0.668  cylinders 6.989e+00 8.248e+00 0.847  displacement -4.785e-01 1.894e-01 -2.527  horsepower 5.034e-01 3.470e-01 1.451  weight 4.133e-03 1.759e-02 0.235  acceleration -5.859e+00 2.174e+00 -2.696  year 6.974e-01 6.097e-01 1.144  origin -2.090e+01 7.097e+00 -2.944  cylinders:displacement -3.383e-03 6.455e-03 -0.524  cylinders:horsepower 1.161e-02 2.420e-02 0.480  cylinders:weight 3.575e-04 8.955e-04 0.399  cylinders:acceleration 2.779e-01 1.664e-01 1.670  cylinders:year -1.741e-01 9.714e-02 -1.793  cylinders:origin 4.022e-01 4.926e-01 0.816  displacement:horsepower -8.491e-05 2.885e-04 -0.294  displacement:weight 2.472e-05 1.470e-05 1.682  displacement:acceleration -3.479e-03 3.342e-03 -1.041  displacement:year 5.934e-03 2.391e-03 2.482  displacement:origin 2.398e-02 1.947e-02 1.232  horsepower:weight -1.968e-05 2.924e-05 -0.673  horsepower:acceleration -7.213e-03 3.719e-03 -1.939  horsepower:year -5.838e-03 3.938e-03 -1.482  horsepower:origin 2.233e-03 2.930e-02 0.076  weight:acceleration 2.346e-04 2.289e-04 1.025  weight:year -2.245e-04 2.127e-04 -1.056  weight:origin -5.789e-04 1.591e-03 -0.364  acceleration:year 5.562e-02 2.558e-02 2.174  acceleration:origin 4.583e-01 1.567e-01 2.926  year:origin 1.393e-01 7.399e-02 1.882  Pr(>|t|)  (Intercept) 0.50475  cylinders 0.39738  displacement 0.01192 \*  horsepower 0.14769  weight 0.81442  acceleration 0.00735 \*\*  year 0.25340  origin 0.00345 \*\*  cylinders:displacement 0.60051  cylinders:horsepower 0.63157  cylinders:weight 0.69000  cylinders:acceleration 0.09584 .  cylinders:year 0.07389 .  cylinders:origin 0.41482  displacement:horsepower 0.76867  displacement:weight 0.09342 .  displacement:acceleration 0.29853  displacement:year 0.01352 \*  displacement:origin 0.21875  horsepower:weight 0.50124  horsepower:acceleration 0.05325 .  horsepower:year 0.13916  horsepower:origin 0.93931  weight:acceleration 0.30596  weight:year 0.29182  weight:origin 0.71623  acceleration:year 0.03033 \*  acceleration:origin 0.00365 \*\*  year:origin 0.06062 .  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2.695 on 363 degrees of freedom  Multiple R-squared: 0.8893, Adjusted R-squared: 0.8808  F-statistic: 104.2 on 28 and 363 DF, p-value: < 2.2e-16 |

(f) Try a few

√ different transformations of the variables, such as

log(X), X, X 2. Comment on your findings.

**Answer:**

Command:

> lm.fit7=lm(mpg~sqrt(horsepower)+sqrt(displacement)+sqrt(weight)+sqrt(year),data=Auto)

> summary(lm.fit7)

|  |
| --- |
| Call:  lm(formula = mpg ~ sqrt(horsepower) + sqrt(displacement) + sqrt(weight) +  sqrt(year), data = Auto)  Residuals:  Min 1Q Median 3Q Max  -8.9665 -2.1092 -0.1462 1.8446 13.9989  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -48.59750 7.99961 -6.075 2.97e-09 \*\*\*  sqrt(horsepower) -0.29967 0.21416 -1.399 0.163  sqrt(displacement) -0.03777 0.14617 -0.258 0.796  sqrt(weight) -0.67234 0.06630 -10.140 < 2e-16 \*\*\*  sqrt(year) 12.84047 0.87250 14.717 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.279 on 387 degrees of freedom  Multiple R-squared: 0.8253, Adjusted R-squared: 0.8235  F-statistic: 457.2 on 4 and 387 DF, p-value: < 2.2e-16 |

> lm.fit7=lm(mpg~log(horsepower)+log(displacement)+log(weight)+log(year),data=Auto)

> summary(lm.fit7)

|  |
| --- |
| Call:  lm(formula = mpg ~ log(horsepower) + log(displacement) + log(weight) +log(year), data = Auto)  Residuals:  Min 1Q Median 3Q Max  -9.195 -1.912 -0.075 1.726 13.594  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -68.8180 17.7967 -3.867 0.000129 \*\*\*  log(horsepower) -2.4891 1.0407 -2.392 0.017251 \*  log(displacement) -0.8543 0.9446 -0.904 0.366327  log(weight) -16.5279 1.8177 -9.093 < 2e-16 \*\*\*  log(year) 55.3470 3.6292 15.250 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.148 on 387 degrees of freedom  Multiple R-squared: 0.839, Adjusted R-squared: 0.8373  F-statistic: 504 on 4 and 387 DF, p-value: < 2.2e-16 |

> lm.fit7=lm(mpg~I(horsepower^2)+I(displacement^2)+I(weight^2)+I(year^2),data=Auto)

> summary(lm.fit7)

|  |
| --- |
| Call:  lm(formula = mpg ~ I(horsepower^2) + I(displacement^2) + I(weight^2) +I(year^2), data=Auto)  Residuals:  Min 1Q Median 3Q Max  -9.6569 -2.7232 -0.1091 2.2742 14.9837  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 3.568e+00 2.317e+00 1.540 0.124  I(horsepower^2) 4.334e-06 4.710e-05 0.092 0.927  I(displacement^2) 1.211e-05 1.226e-05 0.987 0.324  I(weight^2) -1.084e-06 8.416e-08 -12.880 <2e-16 \*\*\*  I(year^2) 5.120e-03 3.775e-04 13.563 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3.789 on 387 degrees of freedom  Multiple R-squared: 0.7668, Adjusted R-squared: 0.7643  F-statistic: 318.1 on 4 and 387 DF, p-value: < 2.2e-16 |

Findings :-

1. The variable Year is always positive with respect to mpg.

2. The variables weight year are always significant with respect tompg

3. The relationship with mpg ranges from 75%-84%