ISLR_Chapter3_Answers

Peiyun Zhou

5/13/2019

3.7 Exercises

Q1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

- The null hypothesese is that TV, radio, and newspaper advertising budgets do not predict the product sales. Based on the P values, we reject the null hypothesis and argue that TV and radio advertising budgets significantly predicted the product sales. There was postive relationship between these two predictors and product sales. The higher advertising budget on TV and radio, the more the product sales.

Q2. Carefully explain the differences between the KNN classifier and KNN regression methods.

- KNN classifier is an approach that attempts to estimate the conditional distribution of Y given X, and then classify a given observation to the class with the highest estimated probability. Given a positive integer K and a test observation x0, the KNN classifer first identifies the K points in the training data that are cloest to X0. Then it estimates the confitional probability for class j as the faction of points in N0 whose reponse values equal j. Finally, KNN applies Nayes rule and classifies the test observation x0 to the class with the largest probability.
- KNN regression method is simliar to KNN classifier method. Given a value for K and a prediction point x0, KNN regression first identifies the K training observations that are closest to x0, represented by N0. It then estimates f(x0) using the average of all the training responses in N0.
- Thus, the differences between KNN classifier and KNN regression is that the former makes the decision on the probability while the later is using the average of all training responses.

Q3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get beta0 = 50, beta1 = 2 0 , beta2 = 0 . 0 7 , beta3 = 3 5 , beta4 = 0 . 0 1 , beta5 = -1 0 .

(a) Which answer is correct, and why?

- 2 is correct, because the coefficient for the gender is positive in the dataset.
- 3 is correct, because the coefficient for the interaction between gender and GPA is negative.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

- Salary=50+ 4x20+ 0.07 x 110 + 35 x 1 + 0.01 (110 x 4)+(1x4)x (-10)

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

- No, the magnitude of the effect depend both on the coefficient and the predictors. Although the coefficient is small, the predictor (GPA*IQ) is in the range of ~100. Multiplied with 0.01, the effects on salary is still around a couple thousand dollars.

Q4. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression. For example, Y = beta0 +beta1X +beta2X2 +beta3X3 +error.

(a) Suppose that the true relationship between X and Y is linear, such as Y = Beta0 + Beta1X + error. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

Q6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point ( ).

Q7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case.

Q8

```
Auto<-read.csv("Auto.csv",header=T)
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   150    : 22
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.0   90     : 20
##  Median :23.00   Median :4.000   Median :146.0   88     : 19
##  Mean   :23.52   Mean   :5.458   Mean   :193.5   110    : 18
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   100    : 17
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   75     : 14
##                                                  (Other):287
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2223   1st Qu.:13.80   1st Qu.:73.00   1st Qu.:1.000
##  Median :2800   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2970   Mean   :15.56   Mean   :75.99   Mean   :1.574
##  3rd Qu.:3609   3rd Qu.:17.10   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##              name
##  ford pinto    :  6
##  amc matador   :  5
##  ford maverick :  5
```

```
##  toyota corolla:  5
##  amc gremlin    :  4
##  amc hornet     :  4
##  (Other)        :368
```

```
Auto$horsepower<-as.numeric(Auto$horsepower)
Auto$mpg<-as.numeric(Auto$mpg)
model1<-lm(mpg~horsepower,data=Auto)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3506  -6.0044  -0.3908   4.9519  22.9816
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.80761    0.71129  25.036   <2e-16 ***
## horsepower   0.11080    0.01195   9.273   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.101 on 395 degrees of freedom
## Multiple R-squared:  0.1788, Adjusted R-squared:  0.1767
## F-statistic: 85.99 on 1 and 395 DF,  p-value: < 2.2e-16
```

```
confint(model1)
```

```
##                   2.5 %     97.5 %
## (Intercept) 16.40922729 19.2059968
## horsepower   0.08731313  0.1342963
```

(a)

i. Is there a relationship between the predictor and the response?

- Yes, horsepower significantly predicted the mpg.

ii. How strong is the relationship between the predictor and the response?

- The adjusted R-squared showed that 17.67% of the change in varaince of mpg is being explained by the horsepower.

iii. Is the relationship between the predictor and the response positive or negative?

- Based on the coefficient, the relationship is positive.

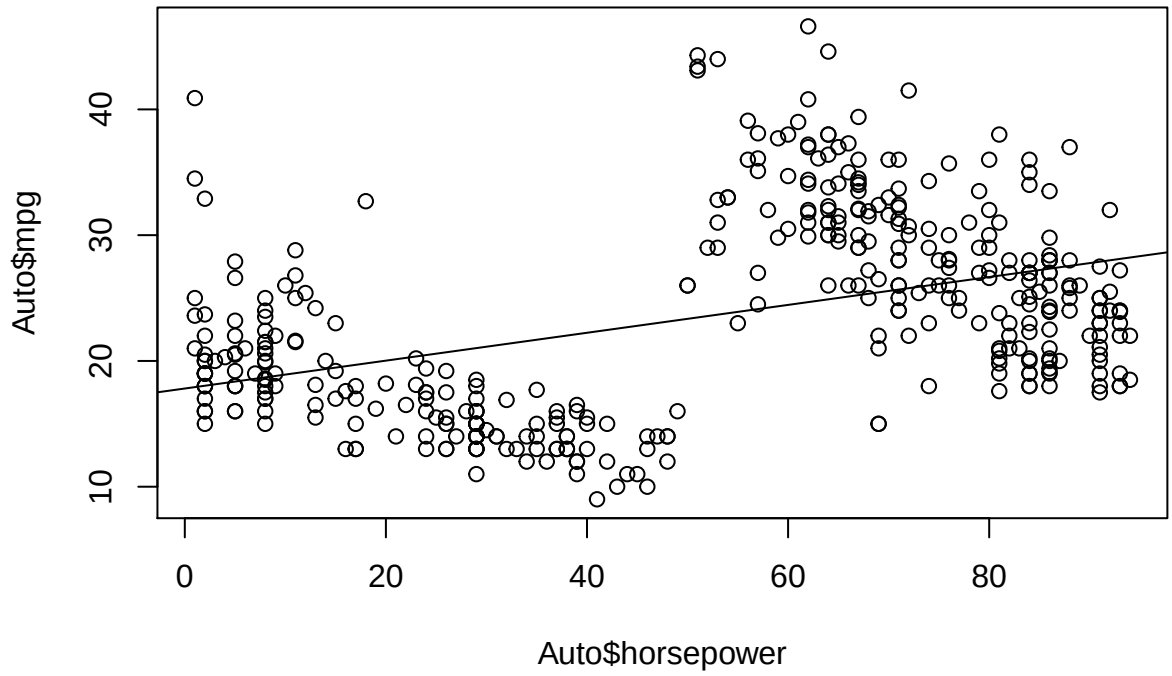
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

- Mpg = 98 x 0.11080 + 17.80761
- CI = hors 0.08731313 0.1342963

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

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

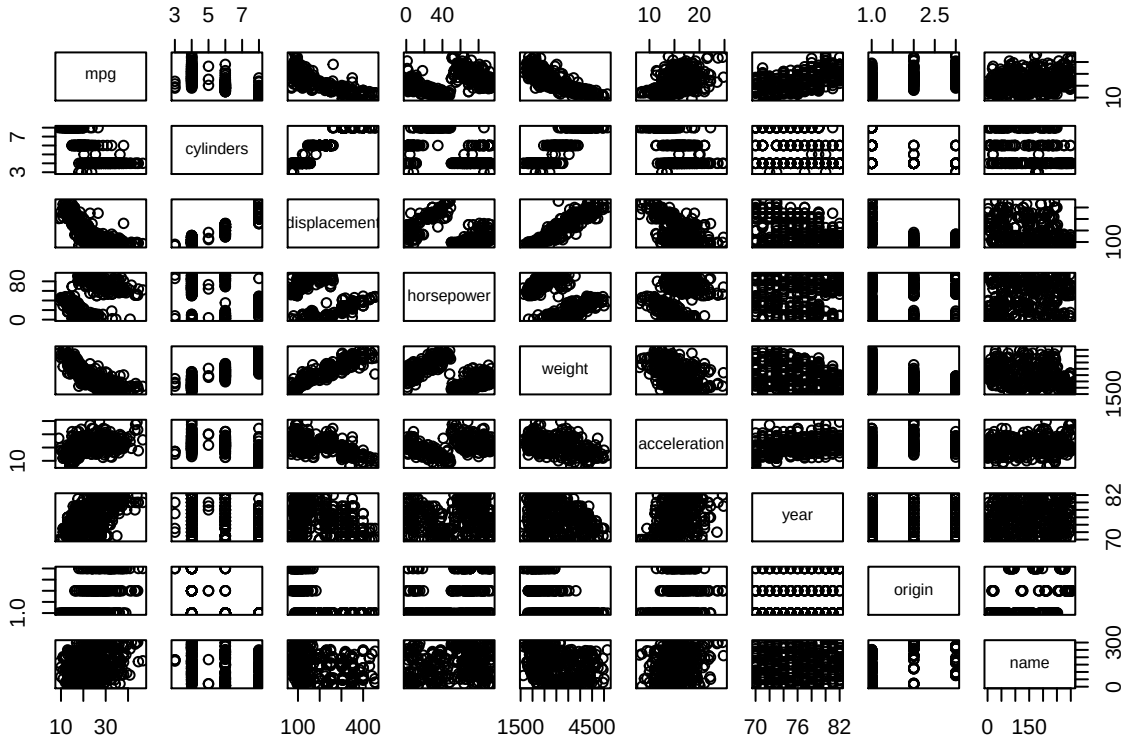
```
plot(Auto$horsepower, Auto$mpg)
abline(model1)
```



Q9 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.

```
pairs(Auto)
```



(b) Com-

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

(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.

- Yes, there are relationships between predictors and respones. Based on the model results, we found all the variables significantly predicted the mpg except cylinders and horsepower. The coefficient for the year suggests a postive relatinship between the year made and mpg. The newest cars tend to have higher mpg.

```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

```
model2=lm(mpg~.-name,data=Auto)
# model2<-lm(mpg~cylinders+displacement+weight+horsepower+acceleration+year+origin,data=Auto)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.629 -2.034 -0.046  1.801 13.010
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.128e+01  4.259e+00  -4.998 8.78e-07 ***
## cylinders    -2.927e-01  3.382e-01  -0.865   0.3874
## displacement  1.603e-02  7.284e-03   2.201   0.0283 *
## horsepower    7.942e-03  6.809e-03   1.166   0.2442
## weight       -6.870e-03  5.799e-04 -11.846  < 2e-16 ***
## acceleration  1.539e-01  7.750e-02   1.986   0.0477 *
## year          7.734e-01  4.939e-02  15.661  < 2e-16 ***
## origin        1.346e+00  2.691e-01   5.004 8.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.331 on 389 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8188
## F-statistic: 256.7 on 7 and 389 DF,  p-value: < 2.2e-16
```

(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?

- Yes, residual plots showed some usually large outliers. The leverage plot also identify some observations with high leverage (14, 327, 394).

(e) Fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```r
model3<-lm(mpg~cylinders+displacement*weight*acceleration*year+origin,data=Auto)
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement * weight * acceleration *
##     year + origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5739 -1.5735  0.0273  1.2608 13.7678
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                         -7.706e+01  1.837e+02  -0.420
## cylinders                           -4.032e-02  3.536e-01  -0.114
## displacement                        -2.059e+00  1.100e+00  -1.872
## weight                               1.050e-01  7.354e-02   1.427
## acceleration                        -3.043e+00  1.103e+01  -0.276
## year                                 1.475e+00  2.439e+00   0.605
## origin                               2.661e-01  2.504e-01   1.063
## displacement:weight                  3.597e-04  2.863e-04   1.256
## displacement:acceleration            1.513e-01  7.595e-02   1.992
## weight:acceleration                 -4.408e-03  4.349e-03  -1.014
## displacement:year                    2.893e-02  1.517e-02   1.907
## weight:year                         -1.484e-03  9.731e-04  -1.525
## acceleration:year                    5.756e-02  1.468e-01   0.392
## displacement:weight:acceleration    -3.084e-05  1.947e-05  -1.584
## displacement:weight:year            -5.130e-06  3.923e-06  -1.308
## displacement:acceleration:year      -2.178e-03  1.043e-03  -2.088
## weight:acceleration:year             5.521e-05  5.747e-05   0.961
## displacement:weight:acceleration:year  4.522e-07  2.668e-07   1.695
##                                     Pr(>|t|)
## (Intercept)                           0.6751
## cylinders                             0.9093
## displacement                          0.0620 .
## weight                                0.1543
## acceleration                          0.7827
## year                                  0.5456
## origin                                0.2886
## displacement:weight                   0.2097
## displacement:acceleration             0.0471 *
## weight:acceleration                   0.3114
## displacement:year                     0.0573 .
## weight:year                           0.1281
## acceleration:year                     0.6952
## displacement:weight:acceleration      0.1140
## displacement:weight:year              0.1917
## displacement:acceleration:year        0.0374 *
## weight:acceleration:year              0.3373
## displacement:weight:acceleration:year  0.0909 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.812 on 379 degrees of freedom
## Multiple R-squared:  0.8764, Adjusted R-squared:  0.8708
## F-statistic: 158.1 on 17 and 379 DF,  p-value: < 2.2e-16
```

(f) Try a few different transformations of the variables, such as log(X), X2. Comment on your findings.

```
model4<-lm(mpg~horsepower+I(horsepower^2),data=Auto)
summary(model4)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9907  -6.0269  -0.2335   4.7160  23.8816
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.8389456  0.9890343  17.026  < 2e-16 ***
## horsepower       0.1801992  0.0507205   3.553 0.000427 ***
## I(horsepower^2) -0.0007355  0.0005225  -1.408 0.160009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.092 on 394 degrees of freedom
## Multiple R-squared:  0.1829, Adjusted R-squared:  0.1787
## F-statistic: 44.09 on 2 and 394 DF,  p-value: < 2.2e-16
```

```
model5<-lm(mpg~cylinders+I(cylinders^2),data=Auto)
summary(model5)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + I(cylinders^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5877  -2.8514  -0.7632   2.4486  17.7486
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     50.7109     4.6674  10.865  < 2e-16 ***
## cylinders       -6.4363     1.7140  -3.755 0.000199 ***
## I(cylinders^2)   0.2429     0.1447   1.678 0.094083 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.928 on 394 degrees of freedom
## Multiple R-squared:  0.6054, Adjusted R-squared:  0.6034
## F-statistic: 302.2 on 2 and 394 DF,  p-value: < 2.2e-16
```