An Introduction to Statistical Learning: Chapter 3

Due on January 2,2016 at 3:10pm

G. James et.al.

Tony Jiang

1. 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 in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	<0.8599

Answer:
H_0: TV ads has no effect on sales. A small p-value makes us reject the null hypotheses and accept the alternative, TV ads has a significant effect on sales.
H_0: radio ads has no effect on sales. A small p-value makes us reject the null hypotheses and accept the alternative, TV ads has a significant effect on sales.
H_0: newspaper ads has no effect on sales. A large p-value makes us retain the null hypotheses. Newspaper ads has no effect on sales.

2. Carefully explain the differences between the KNN classifier and KNN regression methods.
Answer:
The underlying algorithms are the same but the outputs are different: KNN classifier gives a qualitative output while KNN regression methods yields a quantitative one.

3. Suppose we have a data set with five predictors, X_1 =GPA, X_2 =IQ,X_3 =Gender (1 for Female and 0 for Male),X_4 =Interaction between GPA and IQ, and X_5=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 $\hat{\beta}_0$=50,$\hat{\beta}_1$=20,$\hat{\beta}_2$=0.07,$\hat{\beta}_3$=35,$\hat{\beta}_4$=0.01,$\hat{\beta}_5 = -10$.
(a) Which answer is correct, and why?

i For a fixed value of IQ and GPA, males earn more on average than females.
ii For a fixed value of IQ and GPA, females earn more on average than males.
iii For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
iv For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

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

Answer:
(a) iii is correct.
From the fitted model, we have

$$Sales = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \times IQ - 10GPA \times Gender$$

So the difference of salary between Female and Male can be expressed by:

$$Sales_{Female} - Sales_{male} = \big(50 + 20GPA + 0.07IQ + 35 \times 1 + 0.01GPA \times IQ - 10GPA \times \\ 1\big) - \big(50 + 20GPA + 0.07IQ + 35 \times 0 + 0.01GPA \times IQ - 10GPA \times 0\big)$$

Rearrange:

$$Sales_{Female} - Sales_{male} = 35 - 10GPA$$

We can easily see that if GPA¿3.5, Female will make less salary than male on average. So the answer is iii.
(b) Substituting the IQ and GPA into the formula in (a), we can show that the salary will be \$137.1K.
(c) False. The significance of an effect depends on the p-value associated with the testing of the hypothesis on the parameter not on the value of the parameter itself.

4. 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, i.e. $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}^2_X + \hat{\beta}_3 X^3 + \epsilon$ (a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. 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.
Answer:
(a) cubic regression will have a smaller training RSS than linear model because cubic has more flexibility and can provide a better fit.
(b) cubic regression will have a larger test RSS than linear model as there is really no non-linear relationship thus cubic regression over fit the model.
(c) Regardless the linearity in the true relationship, cubic regression will have a smaller training RSS than linear model because cubic has more flexibility and can provide a better fit.

(d) More information are required.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the *ith* fitted value takes the form

$$\hat{y_i} = x_i\hat{\beta},$$

where

$$\hat{\beta} = \Big(\sum_{i=1}^{n} x_i y_i\Big)/\Big(\sum_{i'=1}^{n} x_{i'}^2\Big)$$

Show that we can write

$$\hat{y_i} = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

What is $a_{i'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

Answer:

Substitute $\hat{\beta}$ back to the first equation

$$\begin{aligned}\hat{y_i} &= x_i\hat{\beta}\\ &= x_i\Big(\sum_{i=1}^{n} x_i y_i\Big)/\Big(\sum_{i'=1}^{n} x_{i'}^2\Big)\\ &= \frac{x_i\sum_{i'=1}^{n} x_{i'}y_{i'}}{\sum_{i'=1}^{n} x_{i'}^2}\\ &= \sum_{i'=1}^{n}\frac{x_i x_{i'}}{\sum_{i'=1}^{n} x_{i'}^2} y_{i'}\end{aligned}$$

we can observe that $a_{i'} = \frac{x_i x_{i'}}{\sum_{i'=1}^{n} x_{i'}^2}$

6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

Answer:

We first take summation on all fitted values then divide by n. With some rearrangement we can easily show that ($\bar{y}$,$\bar{x}$) also satisfy the equation defined by the line with intercept of $\hat{\beta}_0$ and a slope of $\hat{\beta}_1$.

$$\begin{aligned}y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i\\ \sum_{i=1}^{n} y_i &= \sum_{i=1}^{n}\big(\hat{\beta}_0 + \hat{\beta}_1 x_i\big)\\ \frac{1}{n}\sum_{i=1}^{n} y_i &= \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\beta}_0 + \hat{\beta}_1 x_i\big)\\ \bar{y} &= \hat{\beta}_0 + \hat{\beta}_1\bar{x}\end{aligned}$$

7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.
Answer:

$$\begin{aligned}
R^2 &= 1 - \frac{RSS}{TSS} \\
&= 1 - \frac{\sum (y_i - \hat{\beta} x_i)^2}{\sum (y_i - \bar{y})^2} \\
&= 1 - \frac{\sum (y_i - \hat{\beta} x_i)^2}{\sum {y_i}^2} \\
&= \frac{\sum {y_i}^2 - \sum (y_i - \hat{\beta} x_i)^2}{\sum {y_i}^2} \\
&= \frac{\sum {y_i}^2 - \sum y_i^2 + 2\hat{\beta} \sum x_i y_i - \hat{\beta}^2 \sum x_i^2}{\sum {y_i}^2} \\
&= \frac{\sum {y_i}^2 - \sum y_i^2 + 2\hat{\beta} \sum x_i (x_i \hat{\beta}) - \hat{\beta}^2 \sum x_i^2}{\sum {y_i}^2} \\
&= \frac{\hat{\beta}^2 \sum x_i^2}{\sum {y_i}^2} \\
&= \frac{\hat{\beta}^2 \sum x_i^2}{\sum {y_i}^2} \\
&= \frac{(\frac{\sum x_i y_i}{\sum x_i^2})^2 \sum x_i^2}{\sum {y_i}^2} \\
&= \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum {y_i}^2} \\
&= \Big(\frac{\sum x_i y_i}{\sum x_i^2 \sum y_i^2}\Big)^2 \\
&= \Big(Cor(x, y)\Big)^2
\end{aligned}$$

Note: this only applies to simple linear regression model with no intercept (so we only need to deal with one β.