Linear Regression

Wendy Liang

2/5/2021

3.1, 3.2, 3.5, 3.6,3.11, 3.12, 3.13, 3.14

3.1 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 than in terms of the coefficients of the linear model.

Answer

$H0_1 : \beta_1 = 0$, $H0_2 : \beta_1 = 0$, $H0_3 : \beta_1 = 0$

TV and radio are significant and newspaper is not significant according to the p values, so we reject $H0_1$ and $H0_2$ and accept $H0_3$. In other wowrd, newspaper do not affect sales.

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

Answer:

1) The KNN classifier: solve classification problems by identifying the neighborhood of x_0 and then estimating the conditional probability P(Y=j|X=x_0) for class j as the fraction of points in the neighborhood whose response values equal j.
2) The KNN regression: solve regression problems by identifying the neighborhood of x_0 and then estimating f(x_0) as the average of all the training responses in the neighborhood.

3.5

$$\hat{y}_i = x_i \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{k=1}^{n} x_k^2} = \sum_{j=1}^{n} \underbrace{\boxed{\frac{x_i x_j}{\sum_{k=1}^{n} x_k^2}}}_{= a_{i'}} \cdot y_j$$

3.6 Using (3.4), argue that in the case of simple linear regression, the

least squares line always passes through the point $(xbar, ybar)$.

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$$

3.11

In this problem we will investigate the t-statistic for the null hypothesis H0: beta=0 in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
set.seed(1)
x=rnorm(100)
y=2*x+rnorm (100)
```

```
fit1 <- lm(y ~ x + 0)
summary(fit1)
```

a.

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
```

```
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

coefficient = 1.9939, standard error = 0.1065 , t = 18.73, p-value = 2.2e-16 $<$ 0.05, so we reject H0.

```
fit2<- lm(x ~ y + 0)
summary(fit2)
```

b.

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

coefficient = 0.39111, the standard error = 0.02089, t = 18.73, p-value = 2.2e-16 $<$ 0.05, so we reject H0.

c. We obtain the same value for the t-statistic and consequently the same value for the corresponding p-value.

```
n <- length(x)
t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
as.numeric(t)
```

d.

```
## [1] 18.72593
```

e. If we replace x_i by y_i in the formula for the t-statistic, the result would be the same.

```
fit3 <- lm(y ~ x)
summary(fit3)
```

f.

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389    0.698
## x            1.99894    0.10773  18.556   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
fit4 <- lm(x ~ y)
summary(fit4)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91    0.365
## y            0.38942    0.02099   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

3.12 This problem involves simple linear regression without an intercept.

a.