HW2

Zhengyu Lu

2022-09-23

## HW2 Question2

### (a)Fit a multiple regression model to predict Sales using Price, Urban, and US.

library(ISLR)  
attach(Carseats)  
contrasts(Urban)

## Yes  
## No 0  
## Yes 1

contrasts(US)

## Yes  
## No 0  
## Yes 1

head(Carseats)

## Sales CompPrice Income Advertising Population Price ShelveLoc Age Education  
## 1 9.50 138 73 11 276 120 Bad 42 17  
## 2 11.22 111 48 16 260 83 Good 65 10  
## 3 10.06 113 35 10 269 80 Medium 59 12  
## 4 7.40 117 100 4 466 97 Medium 55 14  
## 5 4.15 141 64 3 340 128 Bad 38 13  
## 6 10.81 124 113 13 501 72 Bad 78 16  
## Urban US  
## 1 Yes Yes  
## 2 Yes Yes  
## 3 Yes Yes  
## 4 Yes Yes  
## 5 Yes No  
## 6 No Yes

lm.fit <- lm(Sales~Price+Urban+US, data=Carseats)  
summary(lm.fit)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9206 -1.6220 -0.0564 1.5786 7.0581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*  
## Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*  
## UrbanYes -0.021916 0.271650 -0.081 0.936   
## USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

### (b) Provide an interpretation of each coeffcient in the model. Be carefulsome of the variables in the model are qualitative!

If all other variables are constant, for a one unit increase in the price, the sales of car decreases by 0.054459 unit.

If all other variables are constant, and if the location is in urban region, the sales of car decreases by 0.021916 unit. However, this relationship is not statistically significant.

If all other variables are constant, and if the location is in US is Yes, the sales of car increases by 1.200573.

### (c) Write out the model in equation form, being careful to handle the qualitative variables properly.

Y=13.04-0.05*Price-0.02*UrbanYes+1.2\*USYes

### (d) For which of the predictors can you reject the null hypothesis H0 : βj = 0? Use the significance level 0.05 for the hypothesis test.

For variables Price and and US, we can reject the null hypotheses because the p-value are less than 0.05. The p-value of Urban is 0.936 which is much greater than 0.05, so we can not reject the null hypotheses.

### (e) On the basis of your response to question (d), fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

lm.fit2 <- lm(Sales~Price+US, data=Carseats)  
summary(lm.fit2)

##   
## Call:  
## lm(formula = Sales ~ Price + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9269 -1.6286 -0.0574 1.5766 7.0515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*  
## Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*  
## USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.469 on 397 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354   
## F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

### (f) What are the values of R squared for models in (a) and (e)? Does larger R squared mean that it is a better model?

The multiple R squared and adjusted R squared for (a) are 0.2393, 0.2335; For (e), they are 0.2393, 0.2354. These R squared means that only about 23.93% of the Sales variabels can be explained by Price, US, (Urban).

Larger R squared does not mean that it is a better model because it favors a more flexible model, which can be overfitted and is not a good model.

### (g) Using the model from (e), obtain 95 % confidence intervals for the coeficient(s).

confint(lm.fit2)

## 2.5 % 97.5 %  
## (Intercept) 11.79032020 14.27126531  
## Price -0.06475984 -0.04419543  
## USYes 0.69151957 1.70776632

### (h) Fit linear regression models in (e) with interaction effects. Provide an interpretation of each coefficient in the model.

lm.fit3 <- lm(Sales~Price\*US, data=Carseats)  
summary(lm.fit3)

##   
## Call:  
## lm(formula = Sales ~ Price \* US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9299 -1.6375 -0.0492 1.5765 7.0430   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.974798 0.953079 13.614 < 2e-16 \*\*\*  
## Price -0.053986 0.008163 -6.613 1.22e-10 \*\*\*  
## USYes 1.295775 1.252146 1.035 0.301   
## Price:USYes -0.000835 0.010641 -0.078 0.937   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

If all other variables are constant, for a one unit increase in the price, the sales of car decreases by 0.053986 unit.

If all other variables are constant, and if the variable US is Yes, the sales of car increases by 1.295775. However, this relationship is not statistically significant.

Keep the price constant, the effect on Sales of a one unit increase in price is 0.000835 less for those locations which are in US versus those which are not

## HW Question 4(a)

vec=1:100  
for(i in 1:100){x1 <- rnorm(500,30,9)  
error=rnorm(500,0,16)  
y1=10\*x1+error  
m1=lm(x1~y1-1)  
k <- coefficients(m1)  
vec[i]<-k}  
print(vec)

## [1] 0.09965262 0.09977659 0.09988910 0.10000465 0.09950993 0.09984641  
## [7] 0.09966262 0.09963963 0.09976168 0.09939082 0.09971656 0.09945900  
## [13] 0.09946785 0.09955415 0.09985094 0.10017686 0.09937456 0.09975250  
## [19] 0.09963114 0.10019258 0.09961079 0.09977017 0.09997550 0.09983392  
## [25] 0.09973351 0.09964547 0.09982754 0.09972550 0.09967036 0.09995567  
## [31] 0.09979254 0.09969410 0.09974654 0.09954470 0.09951673 0.09960402  
## [37] 0.09963924 0.09936150 0.10015063 0.09976494 0.09964739 0.09979559  
## [43] 0.09977631 0.09943312 0.09945403 0.09937361 0.09980855 0.09989133  
## [49] 0.09937687 0.09954931 0.09929503 0.09989491 0.09977461 0.09980534  
## [55] 0.09977897 0.09982871 0.09987625 0.10011801 0.09996160 0.09989004  
## [61] 0.09973035 0.09998897 0.10018894 0.09968661 0.09978397 0.09973479  
## [67] 0.09964199 0.09964854 0.09989705 0.09998413 0.09987859 0.10007653  
## [73] 0.09991477 0.10009739 0.09963674 0.09998373 0.10001706 0.09932275  
## [79] 0.09982164 0.09944027 0.09949895 0.09975152 0.09977855 0.09895673  
## [85] 0.09939888 0.09970909 0.09960048 0.09948494 0.09979824 0.09957979  
## [91] 0.09998657 0.09995633 0.10017668 0.09963863 0.09976784 0.09979414  
## [97] 0.09970122 0.10014168 0.09995075 0.09928156

mean(vec)

## [1] 0.09973631

According to the result, I think the coefficient estimate for the regression of x onto y is a good estimate of 1=β in the model because the mean of the 100 estimates is very close to 1/10.

### HW Question 4(a) Different setting

vec=1:100  
for(i in 1:100){x2 <- rnorm(500,80,120)  
error2=rnorm(500,0,32)  
y2=35\*x2+error2  
m2=lm(x2~y2-1)  
k2 <- coefficients(m2)  
vec[i]<-k2}  
print(vec)

## [1] 0.02855595 0.02856607 0.02859068 0.02856676 0.02855708 0.02858294  
## [7] 0.02857415 0.02858004 0.02857375 0.02857304 0.02855729 0.02856435  
## [13] 0.02856181 0.02856718 0.02857366 0.02857658 0.02856272 0.02856508  
## [19] 0.02857161 0.02856788 0.02857755 0.02857141 0.02856819 0.02857694  
## [25] 0.02857157 0.02856715 0.02856316 0.02856440 0.02857725 0.02858637  
## [31] 0.02856788 0.02856609 0.02855896 0.02857005 0.02856200 0.02857583  
## [37] 0.02856545 0.02857841 0.02855994 0.02856236 0.02856748 0.02856314  
## [43] 0.02857376 0.02856814 0.02856114 0.02855836 0.02855947 0.02856901  
## [49] 0.02855024 0.02857114 0.02856658 0.02858051 0.02856799 0.02857890  
## [55] 0.02859761 0.02857638 0.02856684 0.02857520 0.02855293 0.02857696  
## [61] 0.02857252 0.02857399 0.02855910 0.02857244 0.02857604 0.02857338  
## [67] 0.02856848 0.02855712 0.02858831 0.02858306 0.02855884 0.02857480  
## [73] 0.02858666 0.02857430 0.02856003 0.02857331 0.02857501 0.02857868  
## [79] 0.02855182 0.02856798 0.02857304 0.02856668 0.02856042 0.02856621  
## [85] 0.02856578 0.02856710 0.02857342 0.02857519 0.02856812 0.02857167  
## [91] 0.02856755 0.02856713 0.02857480 0.02857036 0.02856658 0.02856119  
## [97] 0.02857658 0.02856705 0.02858647 0.02856537

mean(vec)

## [1] 0.02856978

The mean of the 100 estimates is also very close to 1/35.