Chapter 3 HW

Zahraa Alshalal

03/09/2023

## libraries:

# packages  
library(ISLR)  
tinytex::install\_tinytex()

## Conceptual Questions

#### Exercise 1:

* : advertising budgets of “TV”, “radio” or “newspaper” do not have an effect on sales.
* .
* The corresponding p-values are highly significant for “TV” and “radio” and not significant for “newspaper”. We may conclude that newspaper advertising budget do not affect sales.

#### Exercise 3:

#### a)

* Salary (high school):
* Salary (college):
* when GPA is higher than 3.5 high school graduates earn more. iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

#### b)

* Salary:

#### c)

* False: the magnitude of coefficient is not an indicator of statistical significance. We must examine the p-value of the regression coefficient to determine if the interaction term is statistically significant.

## Applied Questions:

#### Exercise 10:

#### a)

summary(Carseats)

## Sales CompPrice Income Advertising   
## Min. : 0.000 Min. : 77 Min. : 21.00 Min. : 0.000   
## 1st Qu.: 5.390 1st Qu.:115 1st Qu.: 42.75 1st Qu.: 0.000   
## Median : 7.490 Median :125 Median : 69.00 Median : 5.000   
## Mean : 7.496 Mean :125 Mean : 68.66 Mean : 6.635   
## 3rd Qu.: 9.320 3rd Qu.:135 3rd Qu.: 91.00 3rd Qu.:12.000   
## Max. :16.270 Max. :175 Max. :120.00 Max. :29.000   
## Population Price ShelveLoc Age Education   
## Min. : 10.0 Min. : 24.0 Bad : 96 Min. :25.00 Min. :10.0   
## 1st Qu.:139.0 1st Qu.:100.0 Good : 85 1st Qu.:39.75 1st Qu.:12.0   
## Median :272.0 Median :117.0 Medium:219 Median :54.50 Median :14.0   
## Mean :264.8 Mean :115.8 Mean :53.32 Mean :13.9   
## 3rd Qu.:398.5 3rd Qu.:131.0 3rd Qu.:66.00 3rd Qu.:16.0   
## Max. :509.0 Max. :191.0 Max. :80.00 Max. :18.0   
## Urban US   
## No :118 No :142   
## Yes:282 Yes:258   
##   
##   
##   
##

attach(Carseats)  
lm.fit = lm(Sales~Price+Urban+US)  
summary(lm.fit)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US)  
##   
## 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)

* Price: The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases.
* Urban: There isn’t a relationship between the location of the store and the number of sales based on the high p-value of the t-statistic.
* US: There is a relationship between whether the store is in the US or not and the amount of sales. The coefficient states a positive relationship between US and Sales. On average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

#### c)

#### d)

* We can reject the null hypothesis for the “Price” and “US” variables.

#### e)

lm.fit2 = lm(Sales ~ Price + US)  
summary(lm.fit2)

##   
## Call:  
## lm(formula = Sales ~ Price + US)  
##   
## 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)

* Based on the RSE and R^2 of the linear regressions, they both fit the data similarly, with linear regression from (e) fitting the data slightly better. about 23.9262888% of the variability is explained by the second model.

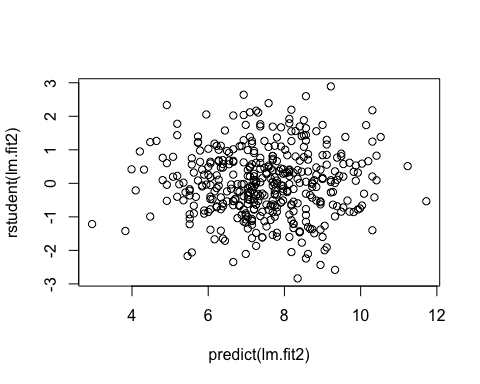
#### g)

confint(lm.fit2)

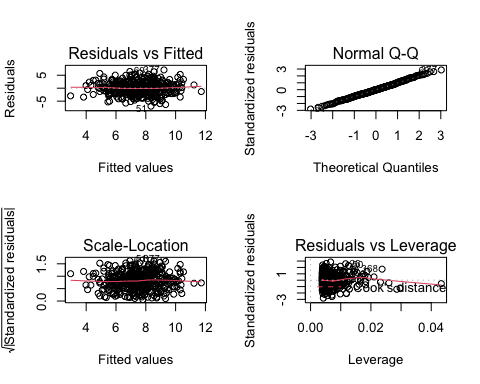
## 2.5 % 97.5 %  
## (Intercept) 11.79032020 14.27126531  
## Price -0.06475984 -0.04419543  
## USYes 0.69151957 1.70776632

#### h)

plot(predict(lm.fit2), rstudent(lm.fit2))



par(mfrow=c(2,2))  
plot(lm.fit2)



* All studentized residuals appear to be bounded by (-3 to 3), so not potential outliers are suggested from the linear regression.
* There are a few observations that greatly exceed (p+1)/n on the leverage-statistic plot that suggest that the corresponding points have high leverage.

#### Exercise 13:

#### a)

set.seed(1)  
x = rnorm(100)

#### b)

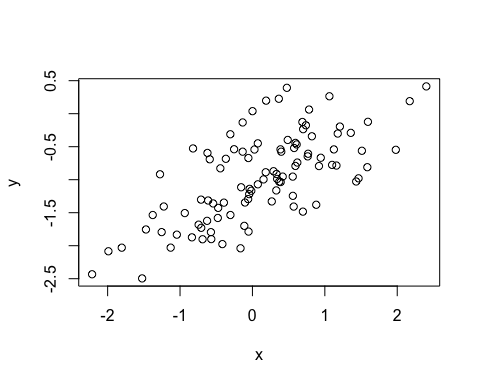
eps = rnorm(100, 0, sqrt(0.25))

#### c)

y = -1 + 0.5\*x + eps

#### d)

plot(x, y)



* A linear relationship between x and y with a positive slope, with a variance as is to be expected.

#### e)

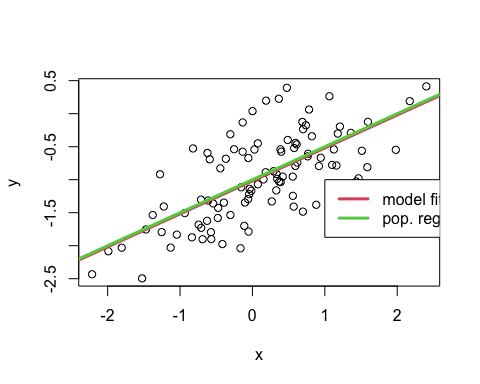
lm.fit = lm(y~x)  
summary(lm.fit)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.93842 -0.30688 -0.06975 0.26970 1.17309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.01885 0.04849 -21.010 < 2e-16 \*\*\*  
## x 0.49947 0.05386 9.273 4.58e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4814 on 98 degrees of freedom  
## Multiple R-squared: 0.4674, Adjusted R-squared: 0.4619   
## F-statistic: 85.99 on 1 and 98 DF, p-value: 4.583e-15

* The linear regression fits a model close to the true value of the coefficients as was constructed. The model has a large F-statistic with a near-zero p-value so the null hypothesis can be rejected.
* similar to β0 and β1.

#### f)

plot(x, y)  
abline(lm.fit, lwd=3, col=2)  
abline(-1, 0.5, lwd=3, col=3)  
legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)



#### g)

lm.fit\_sq = lm(y~x+I(x^2))  
summary(lm.fit\_sq)

##   
## Call:  
## lm(formula = y ~ x + I(x^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98252 -0.31270 -0.06441 0.29014 1.13500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.97164 0.05883 -16.517 < 2e-16 \*\*\*  
## x 0.50858 0.05399 9.420 2.4e-15 \*\*\*  
## I(x^2) -0.05946 0.04238 -1.403 0.164   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.479 on 97 degrees of freedom  
## Multiple R-squared: 0.4779, Adjusted R-squared: 0.4672   
## F-statistic: 44.4 on 2 and 97 DF, p-value: 2.038e-14

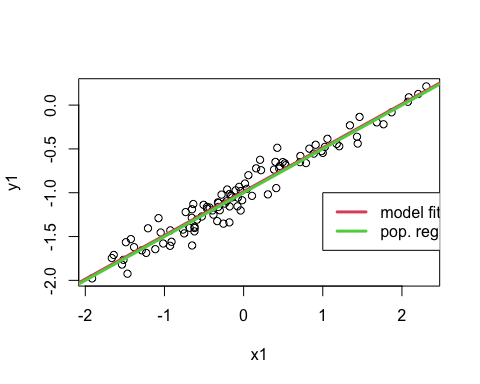
* There is evidence that model fit has increased over the training data given the slight increase in R^2 and RSE.

#### h)

set.seed(1)  
eps1 = rnorm(100, 0, 0.125)  
x1 = rnorm(100)  
y1 = -1 + 0.5\*x1 + eps1  
plot(x1, y1)  
lm.fit1 = lm(y1~x1)  
summary(lm.fit1)

##   
## Call:  
## lm(formula = y1 ~ x1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.29052 -0.07545 0.00067 0.07288 0.28664   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.98639 0.01129 -87.34 <2e-16 \*\*\*  
## x1 0.49988 0.01184 42.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1128 on 98 degrees of freedom  
## Multiple R-squared: 0.9479, Adjusted R-squared: 0.9474   
## F-statistic: 1782 on 1 and 98 DF, p-value: < 2.2e-16

abline(lm.fit1, lwd=3, col=2)  
abline(-1, 0.5, lwd=3, col=3)  
legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)



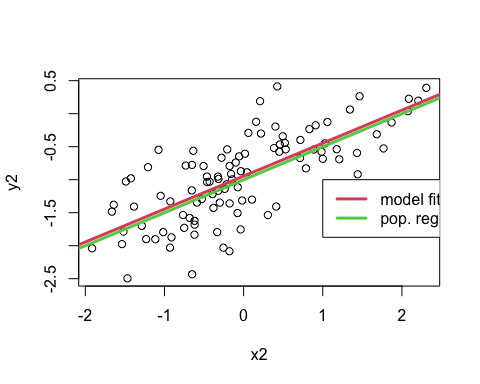
* As expected, the error observed in the values of R^2 decreases considerably.

#### i)

set.seed(1)  
eps2 = rnorm(100, 0, 0.5)  
x2 = rnorm(100)  
y2 = -1 + 0.5\*x2 + eps2  
plot(x2, y2)  
lm.fit2 = lm(y2~x2)  
summary(lm.fit2)

##   
## Call:  
## lm(formula = y2 ~ x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.16208 -0.30181 0.00268 0.29152 1.14658   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.94557 0.04517 -20.93 <2e-16 \*\*\*  
## x2 0.49953 0.04736 10.55 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4514 on 98 degrees of freedom  
## Multiple R-squared: 0.5317, Adjusted R-squared: 0.5269   
## F-statistic: 111.2 on 1 and 98 DF, p-value: < 2.2e-16

abline(lm.fit2, lwd=3, col=2)  
abline(-1, 0.5, lwd=3, col=3)  
legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)



* As expected, the error observed in R^2 and increases considerably.

#### j)

confint(lm.fit)

## 2.5 % 97.5 %  
## (Intercept) -1.1150804 -0.9226122  
## x 0.3925794 0.6063602

confint(lm.fit1)

## 2.5 % 97.5 %  
## (Intercept) -1.008805 -0.9639819  
## x1 0.476387 0.5233799

confint(lm.fit2)

## 2.5 % 97.5 %  
## (Intercept) -1.0352203 -0.8559276  
## x2 0.4055479 0.5935197

* All intervals seem to be centered on approximately 0.5, with the second fit’s interval being narrower than the first fit’s interval and the last fit’s interval being wider than the first fit’s interval.