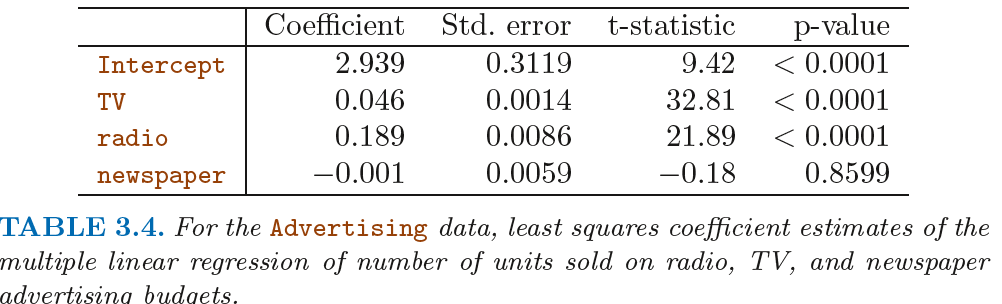
***Lab 02 – Exercises***

*Conceptual*

1. ***Describe the null hypotheses to which the p-values in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Explain you answer in terms of sales, TV, radio, and newspaper rather than in terms of the coefficients of the linear model.***



From the figure, the p-values of TV and radio suggest that there is likely a change in sales as a response to changes in TV and radio that is statistically significant (<0.05). Therefore, for TV and Radio, H0 0, which supports rejection of the null hypothesis. With regard to newspaper, its associated p-value suggests that there is no relationship between newspaper and sales, supporting the null hypothesis H0.

1. ***Carefully explain the differences between the KNN classifier and KNN regression methods.***

The primary distinction between the KNN classifier and the KNN regression is that the KNN classifier uses the nearest neighbor method to categorize each observation. The KNN regression, however, uses the nearest neighbor method to estimate an average value.

1. ***Suppose we have a data set with 5 predictors, X1 = GPA, X2 = IQ, X3 = Gender (1Female 0Male), 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: B0=50, B1=20,B2=0.07,B3=35,B4=0.01,B5=-10.***
2. *Which answers are correct and why?*
3. *For a fixed value of IQ and GPA, males earn more on average than females.*

This statement is **correct** because holding IQ and GPA fixed presents the scenario where:

If Gender = 0, ^y = 50 + GPA(20 + 0.1IQ **– 10(0)**) + 0.07IQ + 35Gender

1. *For a fixed value of IQ and GPA, females earn more on average than males.*

This statement is **incorrect** because holding IQ and GPA fixed presents the scenario where:

If Gender = 0, ^y = 50 + GPA(20 + 0.1IQ **– 10(1)**) + 0.07IQ + 35Gender

1. *For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.*

This statement is **correct**, again, due to the fact that when gender = 1, the value of Income is going to be less than if gender = 0 because of the - 10\*Gender term in the formula.

1. *For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.*

This statement is **incorrect**, again, due to the fact that when gender = 1, the value of Income is going to be less than if gender = 0 because of the - 10\*Gender term in the formula.

1. *Predict the salary of a female with IQ 110 and a GPA of 4.0.*

$176,000

1. *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.*

This is false because a value of 0.1 suggests that there is some change in income resulting from changes in both GPA and IQ. This would effectively reject the null hypothesis.

1. ***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 (Y = B0 + B1X + B2X2 + B3X3 + E).***
2. *Suppose that the true relationship between X and Y is linear (Y = B0 + B1X + E). 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.*

We should expect the training RSS of the cubic regression model to be lower as it will have overfit the data.

1. *Answer (a) using the test RSS values rather than the training RSS values.*

We should expect the test RSS or the linear regression model to be lower given 1) the true relationship *is* linear and 2) the cubic regression would be overfit from its training sets.

1. *Suppose that the true relationship between X and Y is non-linear, but we don’t know how far it is from linear. Consider the training RSS values for the linear and cubic regressions. 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.*

For the training data, we should expect the RSS of the cubic regression to be lower as it will better fit non-linear data than the simple linear model.

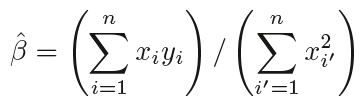
1. *Answer (c) using the test RSS values rather than the training RSS values.*

There is not enough information to tell whether the cubic or linear RSS value will be lower given that we don’t know *how* far from linear the true relationship is.

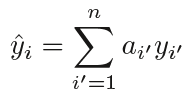
1. ***Consider the fitted values that result from performing linear regression without an intercept****.* ***In this setting, the i-th fitted value takes the form:***



where

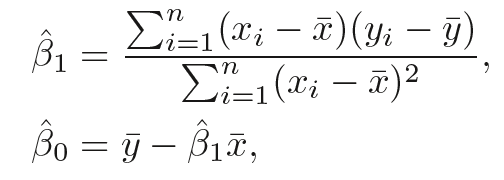
.

***Show that we can write***

******

***What is ai’ ?***

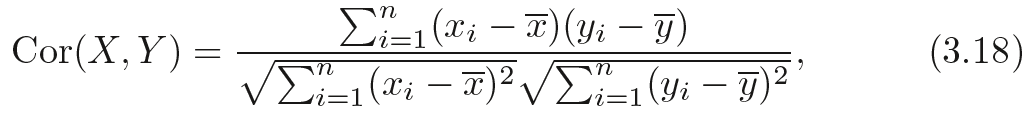
1. ***Using Equation 3.4, argue that in the case of simple linear regression, the least squares line always passes through the point (,).***



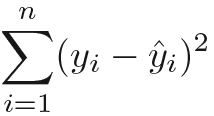
A simple linear regression model will always pass through (mean-x, mean-y) as it is a measure of central tendency that attempts to minimize the distance between each observation and itself.

1. ***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. For simplicity, you may assume that = = 0.***





TSS = Total Sum of Squares = 

RSS = Residual Sum of Squares = 

# 8

setwd('/Users/DamanisMacBook/Documents/Supplementary\_Materials/ISLR/Labs/Lab02')

# NOTE: The AUTO dataset has missing values that must be removed before any analyses can be made on the data

Auto = read.csv("Auto.csv",header=TRUE,na.strings="?")

Auto=na.omit(Auto)

attach(Auto)

# Applying a SIMPLE LINEAR MODEL (Regressing 'mpg' on 'horsepower')

mpgLM=lm(mpg~horsepower)

summary(mpgLM)

# Call:

# lm(formula = mpg ~ horsepower)

# Residuals:

# Min 1Q Median 3Q Max

#-13.5710 -3.2592 -0.3435 2.7630 16.9240

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*

# horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 4.906 on 390 degrees of freedom

# Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049

# F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

plot(horsepower,mpg)

abline(mpgLM,col="red",lwd=2)

# Attemping a PREDICTION

predict(mpgLM,data.frame(horsepower=c(98)))

# 95% Confidence Interval

predict(mpgLM,data.frame(horsepower=c(98)),interval="confidence")

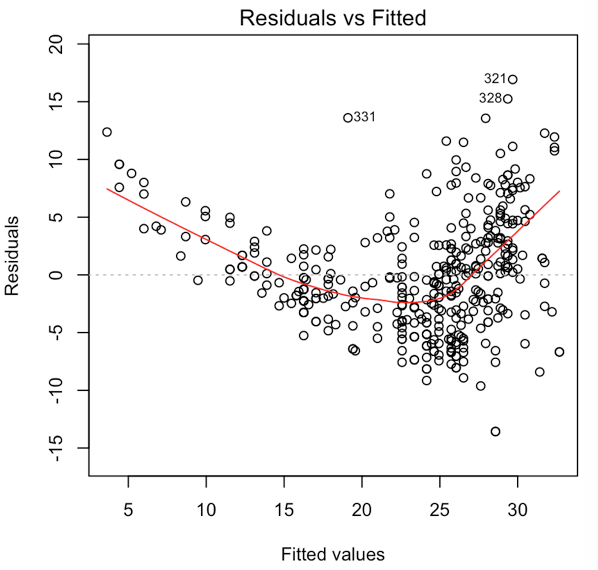
# 95% Prediction Interval

predict(mpgLM,data.frame(horsepower=c(98)),interval="prediction")

# Plotting DIAGNOSTIC PLOTS

par(mfrow=c(2,2))

plot(mpgLM)



# NOTE: A strong trend in the residual plot vs the fitted values suggest

# non-linearity of the data. The lack of a trend in the residuals

# suggest linear data

# 9

# Produce a scatterplot matrix which includes all the variables in the

# dataset

pairs(Auto)

# Comput the MATRIX OF CORRELATIONS (omit the "name" variable,

# since it is qualitative).This can be accomplished using

# the subset function.

cor(subset(Auto,select=-name))

# Create a MULTIPLE LINEAR REGRESSION omitting the "name" variable

mpgMLR = lm(mpg~.-name,data=Auto)

summary(mpgMLR)

# i - There is a relationship between the predictors and the response, with some of the relationships being

# statistically significant

# ii - The most significant relationships with "mpg" are with "displacement", "weight", and "year".

#

# iii - The coefficient for "year" (0.751) suggests that year of manufacture has the strongest relationship with

# fuel economy. For each year that passes, miles per gallon (on average) increases by 3/4 of a mile.

# D

par(mfrow=c(2,2))

plot(mpgMLR)

# There appears to be some trend in the residual plot suggesting that the data are in fact non-linear

# There also appears to be some outliers near fitted values near the higher end of the spectrum

# There appears to be one observation with considerably high leverage (0.5)

# E

# The following interaction terms appear to be statistically significant

# displacement\*horsepower

summary(lm(mpg~displacement\*horsepower,data=Auto))

# displacement\*weight

summary(lm(mpg~displacement\*weight,data=Auto))

# horsepower\*weight

summary(lm(mpg~horsepower\*weight,data=Auto))

# horsepower\*acceleration

summary(lm(mpg~horsepower\*acceleration,data=Auto))

# F

# Trying a few data transformations:

# LOG TRANSOFRMATION

summary(lm(mpg~log(displacement)\*log(horsepower),data=Auto))

# SQUARE ROOT TRANSFORMATION

summary(lm(mpg~sqrt(horsepower)\*sqrt(weight),data=Auto))

# POLYNOMIAL TRANSFORMATION

summary(lm(mpg~poly(horsepower,2),data=Auto))

# 10

# A

library(ISLR)

fix(Carseats)

attach(Carseats)

# Fit a MLR model to regress Sales on Price, Urban, and US:

salesMLR=lm(Sales~Price+Urban+US,data=Carseats)

summary(salesMLR)

# B

# Coefficient Summary

# Price (-0.054)

# Negative relationship

# Urban (-0.022)

# No relationship

# US (1.200)

# Strong positive relationship

# C

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

# Sales = 13.043 - 0.022Urban + 1.20US + E

# Urban 0: 13.043 + E

# Urban 1: 13.021 + E

# US 0: 13.043 + E

# US 1: 14.243 + E

# Urban 1 & US 1: 14.221 - 0.054Price + E

# D

# For which predictors can we reject the null hypothesis?

# Only for Price can the null hypotheses be rejected

# E

salesMLR1=lm(Sales~Price + US,data=Carseats)

summary(salesMLR1)

# F

# Model E is slightly better

# G

confint(salesMLR1)

# H

plot(predict(salesMLR1),rstudent(salesMLR1))

par(mfrow=c(2,2))

plot(salesMLR1)

#----------------------------------------------------------------

# 11 - Linear Regression with randomly generated values

#----------------------------------------------------------------

# In this problem we will investigate the t-statistic for the null hypothesis in a 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)

noINTyx=lm(y~x+0)

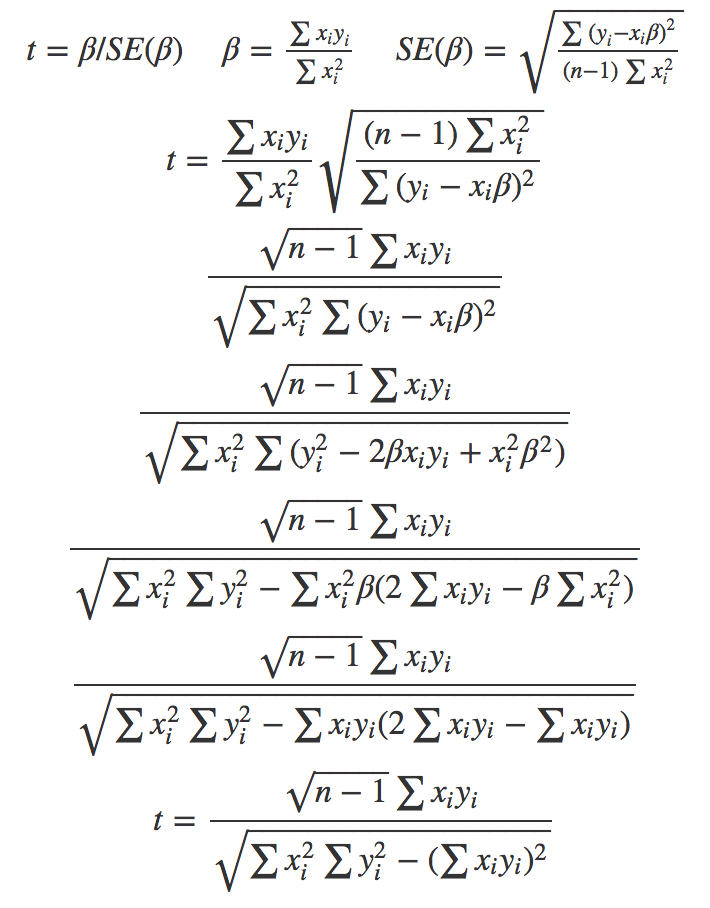
summary(noINTyx)

noINTxy=lm(x~y+0)

summary(noINTxy)

# The relationship of y regressed onto x and x regressed onto y is that they show the same relationship/line

# D



(sqrt(length(x)-1) \* sum(x\*y)) / (sqrt(sum(x\*x) \* sum(y\*y) - (sum(x\*y))\*(sum(x\*y))))

# E

# If you swap t(x,y) for t(y,x) the result will be the same [ t(x,y) = t(y,x) ]

# F - SHow that the t-statistics for two LM (x~y and y~x) are the SAME when performed WITH an intercept

yesINTyx=lm(y~x)

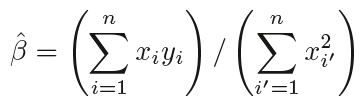
summary(yesINTyx)

yesINTxy=lm(x~y)

summary(yesINTxy)

# Under what circumstance is the COEFFICIENT estimate for the regression of Y onto X the same ase the coefficient

# estimate for X onto Y?

 Eqn. 3.38

# The coefficient estimates will be the same when the sum x^2 = sum y^2

# 12

# A

# Under what circumstance is the COEFFICIENT estimate for the regression of Y onto X the same ase the coefficient

# estimate for X onto Y?

# This is the case when the sum x^2 = sum y^2

# B

yesINTyx=lm(y~x)

summary(yesINTyx)

yesINTxy=lm(x~y)

summary(yesINTxy)

# C

set.seed(1)

x2 = rnorm(100)

y2 = -sample(x2, 100)

# Check

sum(x2^2)

sum(y2^2)

yesINTyx2=lm(y2~x2)

summary(yesINTyx2)

yesINTxy2=lm(x2~y2)

summary(yesINTxy2)

# The regression coefficients are the same for each linear regression. So long as sum sum(x^2) = sum(y^2) the

# condition in 12a. will be satisfied. Here we have simply taken all the xi in a different order and made them

# negative.