Assignment 4

* Aditya Parey

Q1)

1. The regression model can be prepared by using the lm function in R and then using the summary function. (Displayed below)

The equation would be

Price = -1.536e+05 + (3.040e+01\*HouseSize) + (1.869e+01\*LotSize) + (4.012e+03\*PoolYes)

Since the scientific notation e represents very small values, for practical purposes we can rewrite the equation as:

Price ≈ -$153,600 + ($30.40 \* HouseSize) + ($18.69 \* LotSize) + ($4012 \* PoolYes)

A screenshot of a computer

Description automatically generated

1. We take three assumptions into factor: normality, constant error variance, mean of residuals is zero, and independence.

Normality:

The Shapiro-Wilk test (as shown in the code) and the Q-Q plot are used to check the normality of residuals. The Shapiro-Wilk test's p-value (0.8185) and the Q-Q plot suggest that the residuals are approximately normally distributed, which is a good sign.

A graph of a normal and theoretical quantities

Description automatically generated with medium confidence

Constant error variance and mean of residuals:

The variability of the residuals needs to be constant.

This assumption can be elevated using a plot of the residuals versus the predicted values.

We have a residuals vs fitted plot for this.

This assumption can be evaluated using a plot of the residuals versus the predicted values.

If the residuals are centred around the line at 0 that is evidence, the assumption is valid.

However, here we can see that we have a slightly U-shaped curve, which contradicts the above written assumption, The U-shaped pattern in the residuals suggests that the spread of residuals is not constant across all levels of the fitted values. The U-shaped curve also implies that the relationship between the predictor variables and the response is not strictly linear.

A graph with a red line

Description automatically generated

To fix this, if we look at the graph with transformed values where the residuals have been standardized and square-rooted, we can confirm the assumption.  
By this graph we can see that there are points exactly at zero and near zero, which states that there’s a constant error variance.

Graph:

A graph with red line and dots

Description automatically generated

A graph with a red line

Description automatically generated

Independence:

The residuals need to be independent of each other. One way to evaluate independency that occurs related to the order the data was collected.

A graph with black dots and a red line

Description automatically generated

1. HouseSize   
   t-test value = 8.42664e+02 ~ 842.67  
   p-value = 5.380e-98 ~ 0.00000

Since p-value < 0.05, we can conclude that HouseSize is statistically significant and states strong evidence that size of the house has a definitive impact on the price of the house.  
  
  
LotSize:

t-test value: 1.149147e+03 ~ 1149.15   
p-value: 3.419e-104 ~ 0.00000  
Since p-value < 0.05, we can conclude that LotSize is statistically significant and states strong evidence that size of the house has a definitive impact on the price of the house.

PoolYes:  
t-test value: 2.193594e+02 ~ 219.36  
p-value: 4.065e-71 ~ 0.00000

Since p-value < 0.05, we can conclude that PoolYes is statistically significant and states strong evidence that size of the house has a definitive impact on the price of the house.

A computer code with blue text

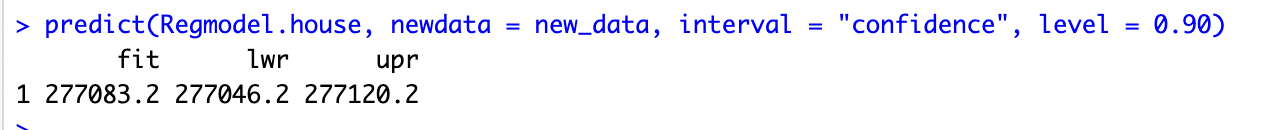
Description automatically generated

1. The estimated price of a house that has a pool which is 2,075 square feet in size, and which is on a lot that is 19,450 square feet is approximately $277,083.2.

A close-up of a white background

Description automatically generated

1. The values in the output represent that we can say with 90% confidence that the population average price of houses with a pool, 2,075 square feet in size, and on a 19,450 square feet lot is between approximately $277,046.2 and $277,120.2.



1. HouseSize:

Intervals: [ 30.34, 30.46]  
We are 90% confident that effect of one unit of increase in house size on the price of house will range between $30.34 and $30.46.

LotSize

Intervals: [ 18.67, 18.72]

We are 90% confident that effect of one unit of increase in lot size on the price of house will range between $18.67 and $18.72.

PoolYes

Intervals: [ 3980.93, 4042.33]

We are 90% confident that the effect of having a pool (PoolYes) vs not having a pool on the price of the house will range between $3980.93 and $4042.33

1. Interpretation of regression coefficients:

HouseSize  
Keeping all other factors equal, the price of a house is predicted to rise by between $30.34 and $30.46 for every unit increase in square feet. This indicates that, assuming all other variables remain constant, larger homes typically will have higher prices as both the variables are directly proportional.

LotSize

Keeping all other factors equal, the house price is predicted to rise by approx. $18.67 to $18.72 for every unit increase in lot size (square feet). This implies that, while other things stay the same, larger lots are likely to have higher prices.

PoolYes:

The price of a house is predicted to rise by roughly $3,980.93 to $4,042.33 if it has a pool (PoolYes = 1) as opposed to not having a pool (PoolYes = 0), holding all other factors constant. This suggests that, generally speaking, a pool raises the price of a house.

Q2)

1. Y ​= −1.35226+ 0.12786\*x1​+ 0.50990\*x2 − 0.01140\*x3 − 0.08007\*x4 + 0.09307\*x5

A screenshot of a computer

Description automatically generated

1. Assumptions:   
   Normality:

The Shapiro-Wilk test (as shown in the code) and the Q-Q plot are used to check the normality of residuals. The Shapiro-Wilk test's p-value (0.4417) and the Q-Q plot suggest that the residuals are approximately normally distributed, which is a good sign.

The p-value is 0.4417, which is greater than the common significance level of 0.05. This means that there is not enough evidence to conclude that the residuals significantly depart from a normal distribution.

A graph of a normal q-q

Description automatically generated

Constant error variance and mean of residuals:

The variability of the residuals needs to be constant.

This assumption can be elevated using a plot of the residuals versus the predicted values.

We have a residuals vs fitted plot for this.

This assumption can be evaluated using a plot of the residuals versus the predicted values.

If the residuals are centred around the line at 0 that is evidence, the assumption is valid.

A graph with a red line and a line

Description automatically generated

Independence:

The residuals need to be independent of each other. One way to evaluate independency that occurs related to the order the data was collected.

A graph with black dots and a red line

Description automatically generated

1. Hypothesis Test

Null Hypothesis(Ho)

Ho = B1 = B2 = B3 = B4 = B5 = 0

( here B is beta )

The population regression coefficient for all predictor variables is equal to 0, which states that there is no significant linear relationship between predictor variable and response variable.

Alternative Hypothesis(Ha)

Ha : at least one of ( B1,B2,B3,B4,B5) =/= 0  
(here, B is beta and =/= represents “is not equal to “)

At least one of the population regression coefficient is not equal to 0, which states that there is a significant linear relationship between predictor variables and response variable.

Test Stat value:

6.897923 ~ 6.898

P-value:

The p-value is 7.868894e-05 ~ 0.000.

Conclusion:

Since the p-value (0.000) is less than alpha(0.05), we reject the null hypothesis and conclude that there is a significant relationship between the predictor and the response variable. The regression model as whole is statistically significant.

1. Forward Selection:

The regression equation would be:   
  
y = -3.4412 + 0.1189\*x1 + 0.5029\*x2 + 0.0879\*x3

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A collage of graphs and charts

Description automatically generated

1. Backward Selection:   
     
   y = -3.4412 + 0.1189\*x1 + 0.5029\*x2 + 0.0879\*x3

A screenshot of a computer

Description automatically generated

A white background with blue text

Description automatically generated

1. Stepwise selection:

The regression equation would be:   
  
y = -1.573 + 0.5703\*x2 + 0.1170\*x5

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A collage of graphs and charts

Description automatically generated

A white background with blue text

Description automatically generated

Full Code:

getwd()

house.data <- read.table("HW4HouseData.txt", header = TRUE)

names(house.data)

## least-squares regression model ##

Regmodel.house <- lm(Price ~ HouseSize + LotSize + Pool, data = house.data)

summary(Regmodel.house)

shapiro.test(Regmodel.house$residuals)

qqnorm(resid(Regmodel.house))

qqline(resid(Regmodel.house))

par(mfrow = c(2,2))

plot(Regmodel.house)

residuals <- resid(Regmodel.house)

# Create a sequence of numbers representing observation order

observation\_order <- 1:length(residuals)

# Create a scatterplot of residuals vs observation order

plot(observation\_order, residuals, xlab = "Observation Order", ylab = "Residuals", main = "Residuals vs Observation Order")

abline(h = 0, col = "red")

summary(Regmodel.house)$coefficients["HouseSize", c("t value", "Pr(>|t|)")]

summary(Regmodel.house)$coefficients["LotSize", c("t value", "Pr(>|t|)")]

summary(Regmodel.house)$coefficients["PoolYes", c("t value", "Pr(>|t|)")]

new\_data <- data.frame(HouseSize = 2075, LotSize = 19450, Pool = "Yes")

predicted\_price <- predict(Regmodel.house, newdata = new\_data)

predicted\_price

predict(Regmodel.house, newdata = new\_data, interval = "confidence", level = 0.90)

intervals <- confint(Regmodel.house, level = 0.90)

intervals

## Q2 ##

q2.data <- read.table("HW4Q2Data.txt", header = TRUE)

names(q2.data)

model <- lm(y ~ x1 + x2 + x3 + x4 + x5, data = q2.data)

summary(model)

Regmodel.house <- lm(y~ x1 + x2 - x3 - x4 + x5, data = q2.data)

summary(model)

shapiro.test(model$residuals)

qqnorm(resid(model))

qqline(resid(model))

par(mfrow = c(2,2))

plot(model)

residuals <- resid(model)

# Create a sequence of numbers representing observation order

observation\_order <- 1:length(residuals)

# Create a scatterplot of residuals vs observation order

plot(observation\_order, residuals, xlab = "Observation Order", ylab = "Residuals", main = "Residuals vs Observation Order")

abline(h = 0, col = "red")

# full regression model

model <- lm(y ~ x1 + x2 + x3 + x4 + x5, data = q2.data)

# summary of the model

model\_summary <- summary(model)

f\_statistic <- model\_summary$fstatistic[1]

p\_value <- pf(f\_statistic, df1 = model\_summary$fstatistic[2], df2 = model\_summary$fstatistic[3], lower.tail = FALSE)

f\_statistic

p\_value

##library(MASS)

# Start with an intercept-only model

MultiReg.empty <- lm(y ~ 1, data = q2.data)

stepForward1 = add1(MultiReg.empty, scope = q2.data, test = "F", trace=TRUE)

stepForward1

#x1 has the highest significant F-value

MultiReg.empty2 <- lm(y ~ x1, data = q2.data)

stepForward2 = add1(MultiReg.empty2, scope = q2.data, test = "F", trace=TRUE)

stepForward2

#x2 has the highest significant F-value

MultiReg.empty3 <- lm(y ~ x1+x2, data = q2.data)

stepForward3 = add1(MultiReg.empty3, scope = q2.data, test = "F", trace=TRUE)

stepForward3

#x5 has the highest significant F-value

MultiReg.empty4 <- lm(y ~ x1+x2+x5, data = q2.data)

stepForward4 = add1(MultiReg.empty4, scope = q2.data, test = "F", trace=TRUE)

stepForward4

shapiro.test(MultiReg.empty4$residuals)

qqnorm(resid(MultiReg.empty4))

qqline(resid(MultiReg.empty4))

plot(MultiReg.empty4$fitted.values,rstandard(MultiReg.empty4))

par(mfrow=c(2,2))

plot(MultiReg.empty4)

par(mfrow=c(1,1))

##library(car)

vif(MultiReg.empty4)

coefficients <- coef(MultiReg.empty4)

coefficients

#backward elimination

# Fit a full regression model with all predictors

MultiReg.full <- lm(y ~ x1 + x2 + x3 + x4 + x5, data = q2.data)

# Perform backward selection to drop the least significant predictors (x3 and x4)

stepBack1 <- drop1(MultiReg.full, scope = ~ x1 + x2 + x3 + x4 + x5, test = "F", trace = TRUE)

stepBack1

# Remove the least significant predictors (x3 and x4) from the linear model

MultiReg.full2 <- lm(y ~ x1 + x2 + x5, data = q2.data)

stepBack2 <- drop1(MultiReg.full2, scope = ~ x1 + x2 + x5, test = "F", trace = TRUE)

stepBack2

vif(MultiReg.full2)

coefficients2 <- coef(MultiReg.full2)

coefficients2

# this is essentially the same model as the forward selection so we don't

# need to verify the assumptions

#stepwise elimination

##chooseCRANmirror(graphics = FALSE)

##install.packages("Matrix", type = "binary")

##install.packages("MatrixModels", type = "binary")

##install.packages("rms")

##library(rms)

##

MultiReg.stepwise=ols(y~x1+x2+x3+x4+x5,data=q2.data)

stepStepwise=fastbw(MultiReg.stepwise,rule="p")

stepStepwise

MultiReg.stepwise2=lm(y~x2+x5,data=q2.data)

summary(MultiReg.stepwise2)

shapiro.test(resid(MultiReg.stepwise2))

par(mfrow=c(2,2))

plot(MultiReg.stepwise2)

par(mfrow=c(1,1))

##library(car)

vif(MultiReg.stepwise2)

coefficients3 <- coef(MultiReg.stepwise2)

coefficients3