> # Exercise 1.

> # Plot the fitted y with the prediction interval based on the quadratic regression model: lm.fit3

> library(MASS)

> library(ISLR)

> Boston$lstat2 <- (Boston$lstat)^2

> lm.fit3 <- lm(medv~lstat+lstat2, data=Boston)

> x1 <- seq(min(Boston$lstat),max(Boston$lstat),length=30)

> x2 <- x1^2

> newX <- data.frame(lstat=x1, lstat2=x2)

> yhat <- predict(object=lm.fit3, newdata=newX, interval="prediction")

> plot(Boston$lstat,Boston$medv, pch=20)

> lines(x1, yhat[,1], lwd=3)

> lines(x1, yhat[,2], col="red", lwd=3)

> lines(x1, yhat[,3], col="red", lwd=3)

>

> # Exercise 2

> # (1) Download the housing dataset from https://www.kaggle.com/harlfoxem/housesalesprediction

> # and run a regression to predict housing prices.

>

>

> # (2) Build a model to predict the housing price given characteristics of a house in the dataset.

> # Consider the following predictors:

> #

> # season, sqft\_living, yr\_built, interaction of sqft\_living yr\_built, and waterfront

>

> # Create a variable "season" which equals

> # "Winter" if a house was sold in Jan, Feb, Mar, Dec.

> # "Spring" if it was sold in Apr, May, Jun

> # "Summer" if it was sold in Jul, Aug

> # "Fall" if it was sold in Sep, Oct, Nov.

> # Do you find any seasonality in housing price?

>

>

> # (3) Do you find any nonlinearity or heteroskedasticity?

> # What is the problem if the error term is heteroskedastic?

> # How can you address these problems (if you have here)?

>

>

> # (4) Conduct an F test for the following hypotheses.

> # H0: there is no seasonality on the housing price. H1: H0 is not true.

>

>

> # (5) Predict the housing price when season = spring, sqft\_living=2500, yr\_built=2000, waterfront=0.

>

>

> ### Exercise 2

> ### (1) & (2)

> setwd("D:/Documents (Louis Booth)/R/Big Data/kc\_house\_data.csv")

> house <- read.csv("kc\_house\_data.csv", header=TRUE)

> head(house)

id date price bedrooms bathrooms sqft\_living sqft\_lot floors waterfront view

1 7129300520 20141013T000000 221900 3 1.00 1180 5650 1 0 0

2 6414100192 20141209T000000 538000 3 2.25 2570 7242 2 0 0

3 5631500400 20150225T000000 180000 2 1.00 770 10000 1 0 0

4 2487200875 20141209T000000 604000 4 3.00 1960 5000 1 0 0

5 1954400510 20150218T000000 510000 3 2.00 1680 8080 1 0 0

6 7237550310 20140512T000000 1225000 4 4.50 5420 101930 1 0 0

condition grade sqft\_above sqft\_basement yr\_built yr\_renovated zipcode lat long sqft\_living15

1 3 7 1180 0 1955 0 98178 47.5112 -122.257 1340

2 3 7 2170 400 1951 1991 98125 47.7210 -122.319 1690

3 3 6 770 0 1933 0 98028 47.7379 -122.233 2720

4 5 7 1050 910 1965 0 98136 47.5208 -122.393 1360

5 3 8 1680 0 1987 0 98074 47.6168 -122.045 1800

6 3 11 3890 1530 2001 0 98053 47.6561 -122.005 4760

sqft\_lot15

1 5650

2 7639

3 8062

4 5000

5 7503

6 101930

> summary(house)

id date price bedrooms bathrooms

Min. :1.000e+06 20140623T000000: 142 Min. : 75000 Min. : 0.000 Min. :0.000

1st Qu.:2.123e+09 20140625T000000: 131 1st Qu.: 321950 1st Qu.: 3.000 1st Qu.:1.750

Median :3.905e+09 20140626T000000: 131 Median : 450000 Median : 3.000 Median :2.250

Mean :4.580e+09 20140708T000000: 127 Mean : 540088 Mean : 3.371 Mean :2.115

3rd Qu.:7.309e+09 20150427T000000: 126 3rd Qu.: 645000 3rd Qu.: 4.000 3rd Qu.:2.500

Max. :9.900e+09 20150325T000000: 123 Max. :7700000 Max. :33.000 Max. :8.000

(Other) :20833

sqft\_living sqft\_lot floors waterfront view condition

Min. : 290 Min. : 520 Min. :1.000 Min. :0.000000 Min. :0.0000 Min. :1.000

1st Qu.: 1427 1st Qu.: 5040 1st Qu.:1.000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:3.000

Median : 1910 Median : 7618 Median :1.500 Median :0.000000 Median :0.0000 Median :3.000

Mean : 2080 Mean : 15107 Mean :1.494 Mean :0.007542 Mean :0.2343 Mean :3.409

3rd Qu.: 2550 3rd Qu.: 10688 3rd Qu.:2.000 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:4.000

Max. :13540 Max. :1651359 Max. :3.500 Max. :1.000000 Max. :4.0000 Max. :5.000

grade sqft\_above sqft\_basement yr\_built yr\_renovated zipcode

Min. : 1.000 Min. : 290 Min. : 0.0 Min. :1900 Min. : 0.0 Min. :98001

1st Qu.: 7.000 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951 1st Qu.: 0.0 1st Qu.:98033

Median : 7.000 Median :1560 Median : 0.0 Median :1975 Median : 0.0 Median :98065

Mean : 7.657 Mean :1788 Mean : 291.5 Mean :1971 Mean : 84.4 Mean :98078

3rd Qu.: 8.000 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997 3rd Qu.: 0.0 3rd Qu.:98118

Max. :13.000 Max. :9410 Max. :4820.0 Max. :2015 Max. :2015.0 Max. :98199

lat long sqft\_living15 sqft\_lot15

Min. :47.16 Min. :-122.5 Min. : 399 Min. : 651

1st Qu.:47.47 1st Qu.:-122.3 1st Qu.:1490 1st Qu.: 5100

Median :47.57 Median :-122.2 Median :1840 Median : 7620

Mean :47.56 Mean :-122.2 Mean :1987 Mean : 12768

3rd Qu.:47.68 3rd Qu.:-122.1 3rd Qu.:2360 3rd Qu.: 10083

Max. :47.78 Max. :-121.3 Max. :6210 Max. :871200

> str(house)

'data.frame': 21613 obs. of 21 variables:

$ id : num 7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...

$ date : Factor w/ 372 levels "20140502T000000",..: 165 221 291 221 284 11 57 252 340 306 ...

$ price : num 221900 538000 180000 604000 510000 ...

$ bedrooms : int 3 3 2 4 3 4 3 3 3 3 ...

$ bathrooms : num 1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...

$ sqft\_living : int 1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...

$ sqft\_lot : int 5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...

$ floors : num 1 2 1 1 1 1 2 1 1 2 ...

$ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...

$ view : int 0 0 0 0 0 0 0 0 0 0 ...

$ condition : int 3 3 3 5 3 3 3 3 3 3 ...

$ grade : int 7 7 6 7 8 11 7 7 7 7 ...

$ sqft\_above : int 1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...

$ sqft\_basement: int 0 400 0 910 0 1530 0 0 730 0 ...

$ yr\_built : int 1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...

$ yr\_renovated : int 0 1991 0 0 0 0 0 0 0 0 ...

$ zipcode : int 98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...

$ lat : num 47.5 47.7 47.7 47.5 47.6 ...

$ long : num -122 -122 -122 -122 -122 ...

$ sqft\_living15: int 1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...

$ sqft\_lot15 : int 5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...

>

> house$date <- as.character(house$date)

> house$date <- strsplit(house$date, 'T')

> for (i in 1:nrow(house)) {

+ house$date[[i]] <- house$date[[i]][1]

+ }

> house$date <- as.Date(as.character(house$date), "%Y%m%d")

>

>

> house$season[format.Date(house$date, "%m") == "01" | format.Date(house$date, "%m") == "02" | format.Date(house$date, "%m") == "03" | format.Date(house$date, "%m") == "12"] <- "Winter"

> house$season[format.Date(house$date, "%m") == "04" | format.Date(house$date, "%m") == "05" | format.Date(house$date, "%m") == "06"] <- "Spring"

> house$season[format.Date(house$date, "%m") == "07" | format.Date(house$date, "%m") == "08"] <- "Summer"

> house$season[format.Date(house$date, "%m") == "09" | format.Date(house$date, "%m") == "10" | format.Date(house$date, "%m") == "11"] <- "Fall"

>

> lm.fit1 <- lm(price~season+sqft\_living+yr\_built+sqft\_living\*yr\_built+waterfront, data=house)

> summary(lm.fit1)

Call:

lm(formula = price ~ season + sqft\_living + yr\_built + sqft\_living \*

yr\_built + waterfront, data = house)

Residuals:

Min 1Q Median 3Q Max

-1497733 -133514 -17225 107931 4190203

Coefficients:

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

(Intercept) 2.744e+06 2.686e+05 10.216 < 2e-16 \*\*\*

seasonSpring 2.182e+04 4.519e+03 4.828 1.39e-06 \*\*\*

seasonSummer 5.037e+03 5.102e+03 0.987 0.324

seasonWinter 7.067e+03 4.730e+03 1.494 0.135

sqft\_living 1.055e+03 1.222e+02 8.632 < 2e-16 \*\*\*

yr\_built -1.438e+03 1.364e+02 -10.543 < 2e-16 \*\*\*

waterfront 7.828e+05 1.930e+04 40.556 < 2e-16 \*\*\*

sqft\_living:yr\_built -3.841e-01 6.179e-02 -6.216 5.20e-10 \*\*\*

---

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

Residual standard error: 243600 on 21605 degrees of freedom

Multiple R-squared: 0.5598, Adjusted R-squared: 0.5597

F-statistic: 3925 on 7 and 21605 DF, p-value: < 2.2e-16

>

> # Yes, based on the significance of the seasonSpring, seasonSummer, and seasonWinter coefficients,

> # there is likely some seasonality in housing prices

>

> ### (3)

> par(mfrow=c(2,2))

> plot(lm.fit1)

>

> # Yes there is heteroscedasticity present, based on the Residuals vs Fitted and Scale-Location plots

> # The problem with heteroscedasticty is that the variance of the error terms varies based on Xi

> # I would likely use robust standard errors in the model to correct for this

>

> ### (4)

> lm.fit\_u <- lm(price~season+sqft\_living+yr\_built+sqft\_living\*yr\_built+waterfront, data=house)

> lm.fit\_r <- lm(price~sqft\_living+yr\_built+sqft\_living\*yr\_built+waterfront, data=house)

> ssr\_u <- sum((fitted(lm.fit\_u) - house$price)^2)

> ssr\_r <- sum((fitted(lm.fit\_r) - house$price)^2)

> f <- ((ssr\_r - ssr\_u)/3)/(ssr\_u/(21613-5-1))

> # Reject Ho

>

> ### (5)

> predict(object=lm.fit1,newdata=data.frame(season="Spring", sqft\_living=2500, yr\_built=2000, waterfront=0), interval="prediction")

fit lwr upr

1 605086.3 127535.9 1082637

> predict(object=lm.fit1,newdata=data.frame(season="Spring", sqft\_living=2500, yr\_built=2000, waterfront=0), interval="confidence")

fit lwr upr

1 605086.3 598434.5 611738.1

> predict(object=lm.fit1,newdata=data.frame(season="Spring", sqft\_living=2500, yr\_built=2000, waterfront=0))

1

605086.3