Regression Analytics

Dutt Thakkar

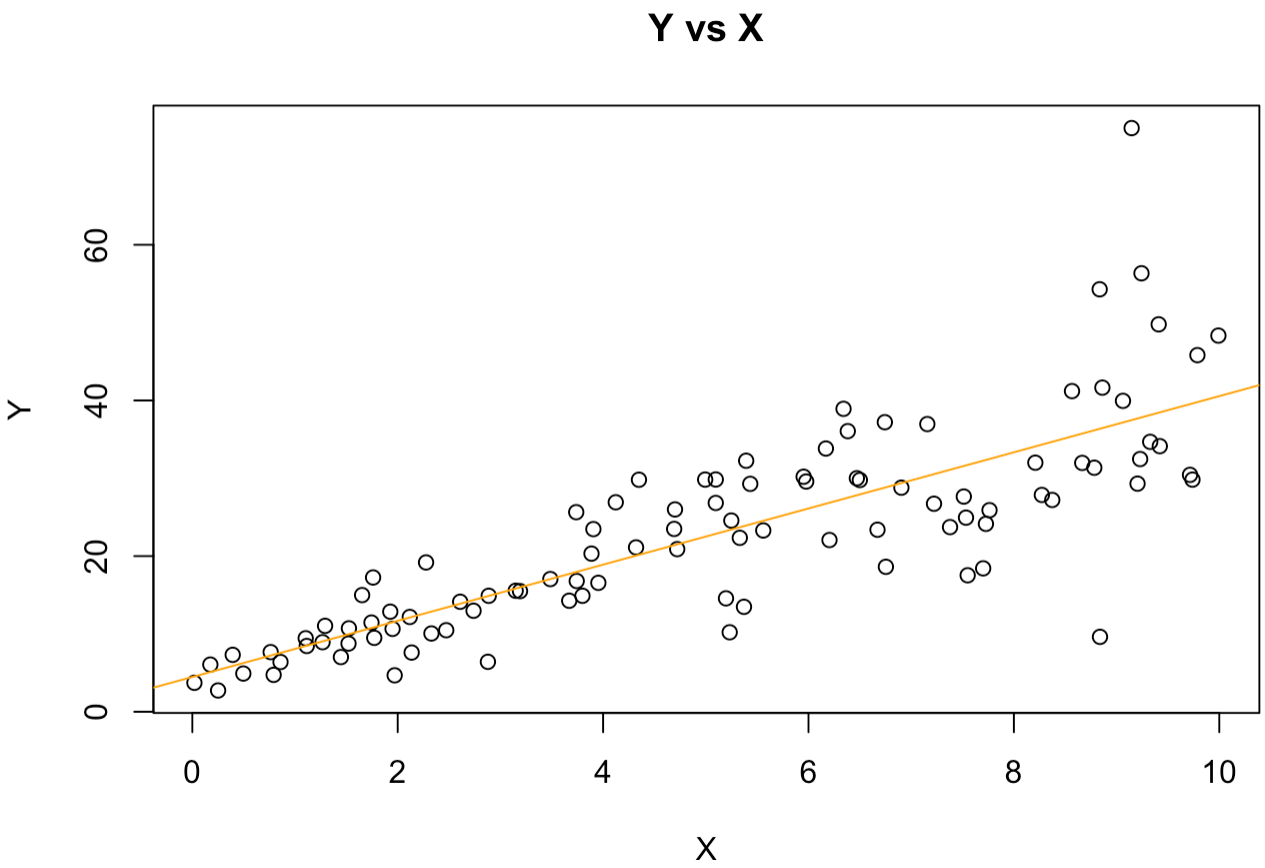
2023-05-15

#Question 1: Run the following code in R-studio to create two variables X and Y.

```
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
```

##a: Plotting Y against X to determine if we can fit a linear model to explain Y based on X

```
graph = plot(X,Y, main = "Y vs X")
abline(lsfit(X, Y), col = "orange")
```




```
#After examining the scatter plot, it can be concluded that we can fit a linear model
to explain Y based on X
```

b: Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model?

```
lm(Y~X)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611
```

```
summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
#based on the results obtained from the summary, the value of X is estimated as 3.610
8, P value is 2e-16 which is less than 5% so the value of X is significant, and the R
squared value is 0.65 i.e. 65%. It can be said that the accuracy of this model is 6
5%.
```

c: How the coefficient of determination R squared of the model above is related to the correlation coefficient of X and Y?

```
Correlation <- cor(X,Y)
Correlation
```

```
## [1] 0.807291
```

```
coefficientofdetermination <- round(Correlation^2,4)
coefficientofdetermination
```

```
## [1] 0.6517
```

```
#Two variables, X and Y, have a linear relationship that can be quantified using the
coefficient of determination (R-squared) and the correlation coefficient (r). The squ
are of the correlation coefficient (r), in basic linear regression, when there is jus
t one independent variable (X) and one dependent variable (Y), equals the coefficient
of determination (R-squared). This means that the square of the correlation coefficie
nt is connected to the R-squared value of the linear regression model between X and
Y. For example, if the correlation between X and Y is r, then the R-squared value of
the X-Y linear regression model is around r^2
```

#Question 2: We will use 'mtcars' dataset for this question.

a: Constructing a simple linear model using mtcars data to predict which factor estimates the horse power at its best; weight of the car or fuel consumption (mpg)

```
#loading the dataset and viewing the summary
data("mtcars")
head(mtcars)
```

	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0

6 rows | 1-10 of 12 columns

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
#constructing two separate linear models using two variable "wt" and "mpg"
model_wt= lm(mtcars$hp ~ mtcars$wt)
model_mpg=lm(mtcars$hp ~ mtcars$mpg)
summary(model_wt)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -83.430 -33.596 -13.587   7.913 172.030 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   -1.821     32.325  -0.056    0.955    
## mtcars$wt     46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151 
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

```
summary(model_mpg)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mtcars$mpg     -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
cat("r squared value for the weight model is 0.43 i.e. 43%, and for the fuel model is
0.60 i.e. 60%. After examining both the r squared values, it can be concluded that fu
el (mpg) is a better predictor of horse power.")
```

```
## r squared value for the weight model is 0.43 i.e. 43%, and for the fuel model is
0.60 i.e. 60%. After examining both the r squared values, it can be concluded that fu
el (mpg) is a better predictor of horse power.
```

##b: Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 cylinder and mpg of 22?

```
#constructing a model for "cyl" and "mpg"
model_hp= lm(mtcars$hp ~ mtcars$cyl + mtcars$mpg)
summary(model_hp)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$cyl + mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## mtcars$cyl    23.979      7.346   3.264  0.00281 **
## mtcars$mpg    -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

```
#using coef() to determine the horse power with 4 cylinder and mpg of 22, as the summ
ary shows the results when the numbers of both cylinders and mpg is zero.
b0 = coef(model_hp)[1]
b1 = coef(model_hp)[2]
b3 = coef(model_hp)[3]
hp_predict = b0 + b1*4 + b3*22
hp_predict
```

```
## (Intercept)
##    88.93618
```

```
#OR

model <- lm(hp ~ cyl + mpg, data = mtcars)
new_data <- data.frame(cyl = 4, mpg = 22)
prediction <- predict(model, newdata = new_data)
prediction
```

```
##        1
## 88.93618
```

```
cat("The estimated Horse Power of a car with 4 cylinder and mpg of 22 is 88.94")
```

```
## The estimated Horse Power of a car with 4 cylinder and mpg of 22 is 88.94
```

#Question 3: we will use boston housing dataset from mlbench package

```
#loading mlbench package and attaching the dataset
library(mlbench)
data("BostonHousing")
head(BostonHousing)
```

	crim	zn	indus	chas	nox	rm	age	dis	rad
	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3
6	0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3

6 rows | 1-10 of 15 columns

```
summary(BostonHousing)
```

```
##       crim                zn             indus       chas         nox
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490
##  Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710
##        rm             age              dis              rad
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000
##  Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000
##       tax           ptratio            b               lstat
##  Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73
##  1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95
##  Median :330.0   Median :19.05   Median :391.44   Median :11.36
##  Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95
##  Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97
##       medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00
```

a: constructing a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and bounds Chas River(chas)

```
house_price= lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)
summary(house_price)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
#All the coefficients of the model are significant as the P values are less than 5%.
The R squared value is 0.3599 i.e. 36%. It can be said that the model only explains 3
6% of the variation in the dependent variable, which means that there maybe other var
iables are not included in the model that are also important in predicting median val
ue of owner-occupied homes.
```

b: Use the estimated coefficient to answer the i and ii

i: Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much?

```
# The linear model equation can be written as medv = 49.91868 - 0.26018 * crim + 0.07
073 * zn - 1.49367 * ptratio + 4.58393 * chas
# The coefficient for the variable "chas" is 4.58393, as seen in the model summary th
at is provided. All other things being equal, this means that, on average, a home tha
t borders the Chas River (chas = 1) is linked to an increase in the median value of o
wner-occupied homes of $4,583.93 compared to a property that does not border the rive
r (chas = 0).
```

ii: Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15

and in the other one is 18. Which one is more expensive and by how much?

```
# The coefficient for the variable "ptratio" is -1.49367, as shown in the model summa
ry that is provided. This means that, when all other circumstances are held constant,
a one-unit rise in the local pupil-teacher ratio is typically accompanied with a $1,4
93.67 fall in the median value of owner-occupied residences.

# Inferring that all other model variables are equivalent for both homes, it follows
that the home with a pupil-teacher ratio of 15 would be more expensive than the home
with a pupil-teacher ratio of 18. There is a 3 point discrepancy between the two hous
es' student-teacher ratios (18-15). As a result, the two homes' estimated median valu
es would differ by: 3 * (-1.49367) = -4.48101

# Hence, if all other factors in the model are equal for both homes, the home with th
e lower pupil-teacher ratio of 15 would cost $4,481.01 more than the home with a high
er pupil-teacher ratio of 18.
```

c: Which of the variables are statistically important (i.e. related to the house price)?

```
# The variables "crim", "zn", "ptratio", and "chas1" all are statistically significan
t in predicting the median value of owner-occupied homes, according to the model summ
ary supplied. This is due to the fact that all of the p-values for the coefficients o
f these variables are less than 0.05

# Specifically, the variables "crim" and "ptratio" have negative coefficients, indica
ting that an increase in these variables is associated with a decrease in median valu
e of owner-occupied homes, while the variable "zn" and "chas1" has a positive coeffic
ient, indicating that houses that bound the Charles River tend to have a higher media
n value of owner-occupied homes.
```

d: Use the anova analysis and determine the order of importance of these four variables.

```
anova(house_price)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
crim	1	6440.7831	6440.78306	118.00683	7.902220e-25
zn	1	3554.3362	3554.33620	65.12189	5.252886e-15
ptratio	1	4709.5358	4709.53584	86.28724	4.738745e-19

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
chas	1	667.1868	667.18681	12.22407	5.136898e-04
Residuals	501	27344.4535	54.57975	*NA*	*NA*

5 rows

```
# Based on the ANOVA table we can determine the order of importance by examining the
F value. The more significant the variable is in explaining the variation in the resp
onse variable, the higher the F value.

#  we can see that the order of importance for the variables is:
# crim = 118.007
# ptratio = 86.287
# zn = 65.122
#chas = 12.224
```