Assignment-2

#### Business Analytics

### Regression Models

* define X and Y variables

## 1 a)

++plot Y against X - in terms of the two variables plot we can easily recognise that a fit line ot we can say that a linear moder to explain y based on x can be employed here . ++fit an abline o the plot of the two variables ++simple linear model of Y based on X ++Summary of model

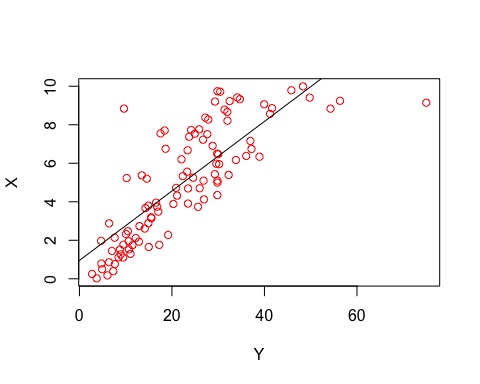
set.seed(2017)  
X=runif(100)\*10  
Y=X\*4+3.45  
Y=rnorm(100)\*0.29\*Y+Y  
library("dplyr")

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library("stats")  
plot(Y,X,col= "RED")  
abline(lsfit(Y,X))

 ##1 b) ++a simple linear model of Y based on X ++equation that explains Y based on X.

linearmodel = lm(Y~X)  
Y\_cap = 4.465 + 3.611\*X  
Y\_cap

## [1] 37.839401 23.862440 21.407655 14.887291 32.272884 32.369678 5.884930  
## [8] 20.169441 21.514928 14.353114 28.815110 4.537970 5.371127 20.067329  
## [15] 22.498042 18.500305 18.742003 30.309192 38.444506 34.336245 27.362839  
## [22] 18.182013 38.483293 31.727822 12.684064 37.493590 26.869270 15.987866  
## [29] 7.219789 13.883908 36.380622 5.097965 15.835274 12.111763 37.188790  
## [36] 27.936636 9.133070 8.451723 10.861225 23.715680 24.540363 26.050357  
## [43] 11.580205 7.323825 39.546068 22.870088 6.258124 28.852481 32.495768  
## [50] 31.669262 23.416770 36.369507 17.053904 27.512090 31.107993 14.858339  
## [57] 11.425415 39.805927 23.940030 37.792876 13.393137 17.980940 25.959734  
## [64] 31.592154 30.544155 21.436567 10.821220 8.488287 26.742813 28.557112  
## [71] 40.544212 29.397613 9.956767 36.180199 39.633234 17.964017 18.566786  
## [78] 23.360218 19.353767 38.146566 9.689243 35.396789 22.870504 37.701072  
## [85] 12.872185 11.506054 9.971260 10.435411 17.717494 7.565804 35.756168  
## [92] 27.831331 23.229779 24.082941 34.701972 9.044349 34.106968 12.176735  
## [99] 36.464708 10.764252

print(linearmodel)

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

summary(linearmodel)

##   
## 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

## 1 c) R2, of the model above is related to the correlation coefficient of X and Y

+R^2 the 0.6517 which means the linear model explains 65.17% variability the target (response) variable i.e. X is a good predictor of the variable Y +Accuracy of the model is R-square= 0.6517 (65.17%)

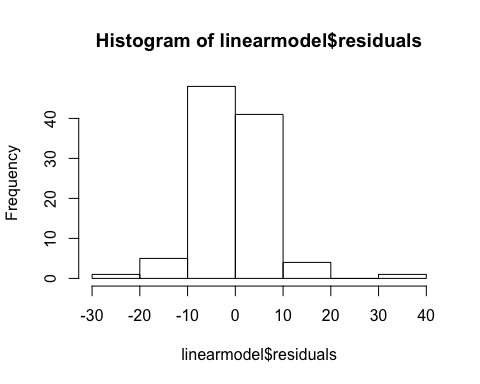
## 1 d)Checking appropriateness of using linear regression for this case using indicator :

++histogram of the residuals ++normal qualtile plot

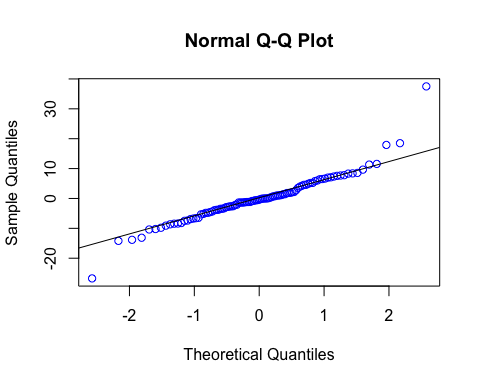
linearmodel$residuals

## 1 2 3 4 5   
## 18.501672740 -10.379893918 2.084271099 0.013010464 -13.847100704   
## 6 7 8 9 10   
## -8.225488097 1.427548815 9.644139813 -0.620050605 -1.373258547   
## 11 12 13 14 15   
## 8.390746639 -0.816255630 -2.636410429 1.049321497 7.334166086   
## 16 17 18 19 20   
## 1.817246622 -2.176461453 6.656261968 11.335591983 -6.477820283   
## 21 22 23 24 25   
## 11.560990445 -3.279977032 -4.354782969 -14.192855481 6.502406011   
## 26 27 28 29 30   
## 37.502853950 -4.801805886 -0.485606393 0.444300190 0.249659569   
## 31 32 33 34 35   
## -26.754610210 0.961227632 -0.287494263 0.066627160 2.760061830   
## 36 37 38 39 40   
## 1.838982603 1.903543601 0.989566313 -1.363442590 -1.366859122   
## 41 42 43 44 45   
## -1.246826840 3.531580051 -6.899581054 -2.582740335 -9.103827757   
## 46 47 48 49 50   
## 3.959974087 -1.347382909 -10.227239486 -6.594210723 -6.727843381   
## 51 52 53 54 55   
## 1.151109841 17.912915275 -0.005031902 8.544366336 -7.380155555   
## 56 57 58 59 60   
## -8.442274611 1.433269899 6.017695549 8.314649390 -5.321731848   
## 61 62 63 64 65   
## -2.917466414 -1.179955144 4.238903662 -3.940420416 -3.814942694   
## 66 67 68 69 70   
## 4.560104227 6.431697631 -0.038076288 7.080415550 -5.167372757   
## 71 72 73 74 75   
## 7.788699257 -0.604496952 -1.192257342 -4.823220149 -9.805052797   
## 76 77 78 79 80   
## 7.679195987 4.901407727 -13.148527433 7.562189281 -3.463935631   
## 81 82 83 84 85   
## -2.664037825 5.805422778 6.973314078 -8.389999515 -2.815175872   
## 86 87 88 89 90   
## -0.855837721 0.745579488 4.554990380 -3.450621203 -1.175658053   
## 91 92 93 94 95   
## -3.781148131 2.172320914 -8.659352437 5.200075930 -7.487632429   
## 96 97 98 99 100   
## -0.107320813 -2.101663882 -4.568427985 5.182045734 0.695499813

hist(linearmodel$residuals)



qqnorm(linearmodel$residuals,col="blue")  
qqline(linearmodel$residuals)

 ##results Interpretation By plotting the residuals histogram and qqplot line, we can see the that residual errors are normally distributed which satisfies one assumption of linear regression for normality, Therefore Linear regression approach is appropiate. the above two plots of the residuals confirm that the distribution is normal thus the regression model is fit for this data and R^2 is a good estimator of the model accuracy here.

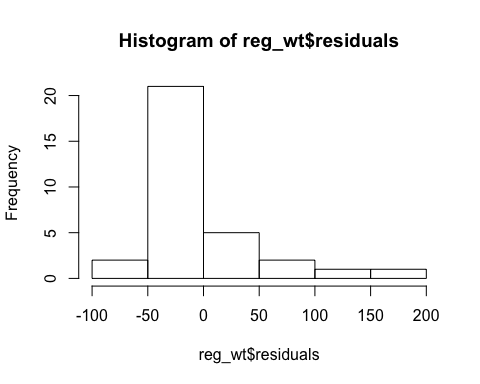
## 2a) Regression model to test relation between :

++weight of a car (wt) as an estimator of cars the Horse Power Vs. ++fuel consumption expressed in Mile Per Gallon (mpg), as an estimator of the (hp)

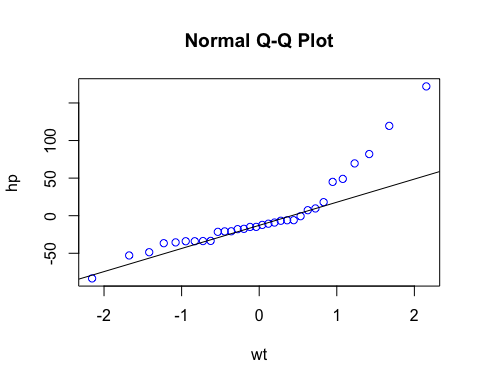
#linear model with weight as an estimator of hp  
reg\_wt <-lm(hp ~ wt, data = mtcars)  
summary(reg\_wt)

##   
## Call:  
## lm(formula = hp ~ wt, data = mtcars)  
##   
## 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   
## 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

hist(reg\_wt$residuals)



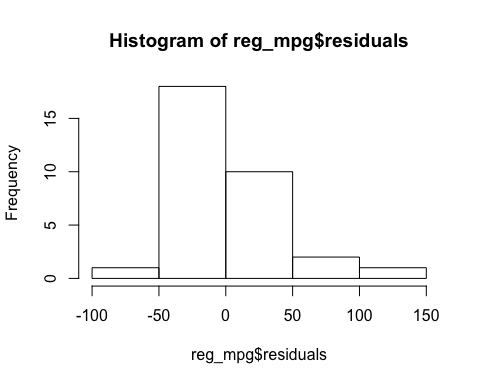
qqnorm(reg\_wt$residuals, xlab = "wt", ylab ="hp",col="blue")  
qqline(reg\_wt$residuals)



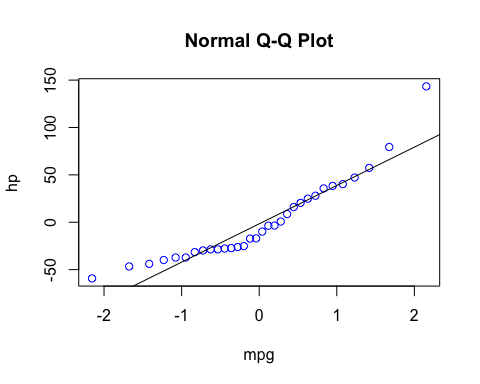
#linear model with mpg as an estimator of hp  
reg\_mpg <-lm(hp ~ mpg, data = mtcars)  
summary(reg\_mpg)

##   
## Call:  
## lm(formula = hp ~ mpg, data = mtcars)  
##   
## 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 \*\*\*  
## 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

hist(reg\_mpg$residuals)



qqnorm(reg\_mpg$residuals,xlab = "mpg", ylab ="hp",col="blue")  
qqline(reg\_mpg$residuals)

 ##Results interpretation : ++the R^2 of linear model of HP ~ mpg and the histogram of the residuals of this model and the qq plot of the residuals of this model all three of them indicate that mpg is a better estimator of the car’s Horsepower ++R square of the hp~mpg model is less than that of hp~wt model and it assures the model accuracy of 60% while the other r square only predicts the model accuracy as 40% ++car weight does not contribute much towards predicted horse power of the car, As car weight is not statistically significant based on the pvalue. However fuel consumption in mile per gallon contributes in better predictor of car horse power.therefore chris assumption is correct. ## thus chris is right !

## 2 b)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).

## 

reg\_cylmpg = lm(hp ~ cyl + mpg, data = mtcars)  
summary(reg\_cylmpg)

##   
## Call:  
## lm(formula = hp ~ cyl + mpg, data = mtcars)  
##   
## 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   
## cyl 23.979 7.346 3.264 0.00281 \*\*  
## 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

## estimated Horse Power of a car with 4 cylinder and mpg of 22

hp\_pediction <- predict(reg\_cylmpg,data.frame(cyl = c(4),mpg=c(22)))  
hp\_pediction

## 1   
## 88.93618

## 85% confidence interval

predict(reg\_cylmpg,data.frame(cyl = c(4),mpg=c(22)),interval = "prediction",level=0.85)

## fit lwr upr  
## 1 88.93618 28.53849 149.3339

## 3 lets get the data set for use in this part - Bostonhousing and relavnt library

library("mlbench")  
data(BostonHousing)

## 3 a) 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 weather the whether the tract bounds Chas River(chas).

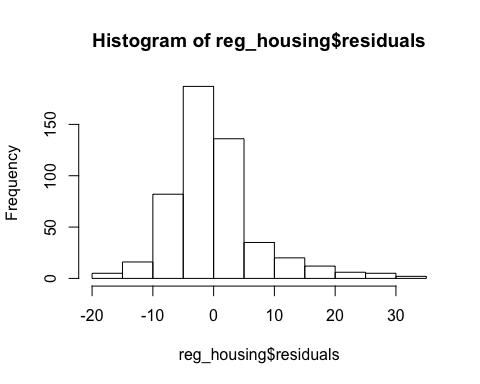
reg\_housing = lm (medv ~ crim + zn + ptratio + chas, data = BostonHousing)  
summary(reg\_housing)

##   
## 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

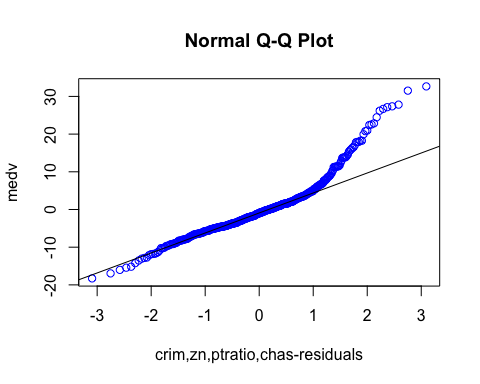
## test- accuracy of model

* the model accuracy is only 35.99% which indicates the model is not very strong but average.

hist(reg\_housing$residuals)



qqnorm(reg\_housing$residuals,xlab= "crim,zn,ptratio,chas-residuals", ylab = "medv",col="blue")  
qqline(reg\_housing$residuals)

 +however the histograms of the residuals and the qq plot shows a normal distribution for the residual , the model is relevant to be used here .

## Further evidence for variables to be a good estimator of the medv ?

reg\_housing1 = lm (medv ~ crim + zn + ptratio, data = BostonHousing)  
summary(reg\_housing1)

##   
## Call:  
## lm(formula = medv ~ crim + zn + ptratio, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.135 -4.584 -1.006 2.727 32.447   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 51.85321 3.22273 16.090 < 2e-16 \*\*\*  
## crim -0.26464 0.04058 -6.522 1.70e-10 \*\*\*  
## zn 0.06523 0.01557 4.191 3.29e-05 \*\*\*  
## ptratio -1.57706 0.17166 -9.187 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.47 on 502 degrees of freedom  
## Multiple R-squared: 0.3442, Adjusted R-squared: 0.3403   
## F-statistic: 87.84 on 3 and 502 DF, p-value: < 2.2e-16

reg\_housing2 = lm (medv ~ crim + zn, data = BostonHousing)  
summary(reg\_housing2)

##   
## Call:  
## lm(formula = medv ~ crim + zn, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.421 -5.060 -1.558 2.121 30.765   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.48563 0.44173 50.904 < 2e-16 \*\*\*  
## crim -0.35208 0.04259 -8.267 1.24e-15 \*\*\*  
## zn 0.11611 0.01571 7.392 6.09e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.065 on 503 degrees of freedom  
## Multiple R-squared: 0.234, Adjusted R-squared: 0.2309   
## F-statistic: 76.82 on 2 and 503 DF, p-value: < 2.2e-16

reg\_housing3 = lm (medv ~ crim, data = BostonHousing)  
summary(reg\_housing3)

##   
## Call:  
## lm(formula = medv ~ crim, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.957 -5.449 -2.007 2.512 29.800   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 24.03311 0.40914 58.74 <2e-16 \*\*\*  
## crim -0.41519 0.04389 -9.46 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.484 on 504 degrees of freedom  
## Multiple R-squared: 0.1508, Adjusted R-squared: 0.1491   
## F-statistic: 89.49 on 1 and 504 DF, p-value: < 2.2e-16

## Results Interpretation :

* we added the variable is an increasing fashion and every time the R square value kept increasing it was the least in reg\_housing3 only one variable of crime ,R square = 0.1508 and the most in the model with all four variables R square= 0.3599 thus ,adding other three improves the model accuracy.

## 3 b)price om basis of house being in the bounds the Chas River or not

* chas variable is binary thus in the model the values of 0 i.e. not near chas is ingnored
* from thr model summary: chas1 4.58393 1.31108 3.496 0.000514 \*\*\* House bounds to chas river would be more expensive than that which does not, because based on linear regression model chas is statistically significant in estimating the median house price. also model indicates with medv = 49.91868 + 4.58393\*chas1

## 3 b)price if the pupil-teacher ratio is 15 and in the other one is 18

houseprice1 = 49.91868 + (-1.49367\*15)  
houseprice1

## [1] 27.51363

houseprice2 = 49.91868 + (-1.49367\*18)  
houseprice2

## [1] 23.03262

## Result interpretation: as the coefficeint of ptration is in negative thsu means if the ratio value is higher the house price would be less and if its less the houseprice would be more.

## 3 c) statistically important variable :

+most statistically important variable are crime rate and pratio - they are explained as lower valuesin them will lead to greater house price which makes sense in genral too- terms of the crime rate ratio and the ptratio is represenative of only this dataset . ##3 d) Anova Analysis to determine the order of importance of these four variables.

anova(reg\_housing)

## Analysis of Variance Table  
##   
## Response: medv  
## Df Sum Sq Mean Sq F value Pr(>F)   
## crim 1 6440.8 6440.8 118.007 < 2.2e-16 \*\*\*  
## zn 1 3554.3 3554.3 65.122 5.253e-15 \*\*\*  
## ptratio 1 4709.5 4709.5 86.287 < 2.2e-16 \*\*\*  
## chas 1 667.2 667.2 12.224 0.0005137 \*\*\*  
## Residuals 501 27344.5 54.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Results Interpretation:

We can see that the variability (sum squared) explained by the crime variable is significantly higher than that of zn or chas. We could guess this as adding the crime,ptratio, and zn significantly improved the model.Also we can see the least indicative variable is chas which is the location in river bounds and then the zone. Still we can see that a large portion of the variability is unexplained, that is shown by residuals 27344.5 so by further adding more relevant variables we might be able to get a better prediction model.