Assignment II - Regression Analysis

Gordon Wall (gwall2)

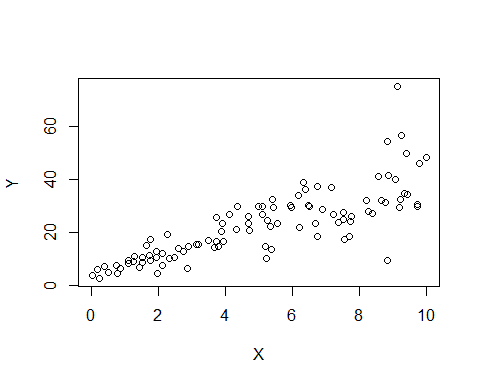
11/9/2019

## 1. Run provided code to create variable X and Y

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

## 1(a) Plot X against Y

plot(X,Y,xlab="X",ylab="Y")



The above plot of X against Y shows that yes, in fact, we can fit a linear regression model to explain the relationship between these two variables. The plotted data suggests that based on Y’s consistent tendency to increase as X increases, some line of “best-fit” should sufficiently line up with, and give insight into, this data subset. The plotted data visually represents a vaguely linear shape, implying some base level of normality (should the residuals be distributed later).

## 1(b) Construct a simple linear model of Y based on X, and write the explanation equation

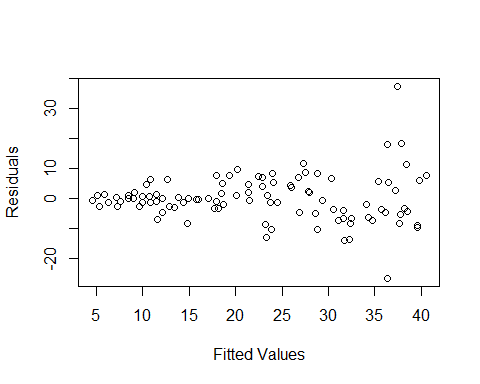
Model=lm(Y ~X)  
print(Model)

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

summary(Model)$sigma #standard error

## [1] 7.756143

plot(Model$fitted.values, Model$residuals, xlab="Fitted Values", ylab="Residuals")



The outputted coefficients reveal that the explanation equation is **Y = 4.465 + 3.611X**.

The initial data looked somewhat linearly shaped however, both the standard error of the model and the plot of residuals vs. fitted values clearly show that the model isn’t ideally accurate. Typically, a low standard error and a random (pattern-less) dataplot of fitted and residual values are ideal in proving accuracy of the model. We may still be able to glean information from this model, but our current approach may not be the best given the above results.

## 1(c) Coefficient of Determination (R^2) in relation to correlation coefficient of X and Y

summary(Model)$r.squared #coefficient of determination

## [1] 0.6517187

cor(X,Y) #correlation coefficient

## [1] 0.807291

Shown above, the correlation coefficient is **0.807291**, meaning that the dataset is reasonably highly correlated (near a value of positive 1). The Coefficient of Determination relates to this value because it is the *square* of the value (0.807291^2), and allows us to determine the certainty of our predictions drawn from this model.

## 2. Load mtcars dataset

mtcars

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160.0 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160.0 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108.0 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258.0 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360.0 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225.0 105 2.76 3.460 20.22 1 0 3 1  
## Duster 360 14.3 8 360.0 245 3.21 3.570 15.84 0 0 3 4  
## Merc 240D 24.4 4 146.7 62 3.69 3.190 20.00 1 0 4 2  
## Merc 230 22.8 4 140.8 95 3.92 3.150 22.90 1 0 4 2  
## Merc 280 19.2 6 167.6 123 3.92 3.440 18.30 1 0 4 4  
## Merc 280C 17.8 6 167.6 123 3.92 3.440 18.90 1 0 4 4  
## Merc 450SE 16.4 8 275.8 180 3.07 4.070 17.40 0 0 3 3  
## Merc 450SL 17.3 8 275.8 180 3.07 3.730 17.60 0 0 3 3  
## Merc 450SLC 15.2 8 275.8 180 3.07 3.780 18.00 0 0 3 3  
## Cadillac Fleetwood 10.4 8 472.0 205 2.93 5.250 17.98 0 0 3 4  
## Lincoln Continental 10.4 8 460.0 215 3.00 5.424 17.82 0 0 3 4  
## Chrysler Imperial 14.7 8 440.0 230 3.23 5.345 17.42 0 0 3 4  
## Fiat 128 32.4 4 78.7 66 4.08 2.200 19.47 1 1 4 1  
## Honda Civic 30.4 4 75.7 52 4.93 1.615 18.52 1 1 4 2  
## Toyota Corolla 33.9 4 71.1 65 4.22 1.835 19.90 1 1 4 1  
## Toyota Corona 21.5 4 120.1 97 3.70 2.465 20.01 1 0 3 1  
## Dodge Challenger 15.5 8 318.0 150 2.76 3.520 16.87 0 0 3 2  
## AMC Javelin 15.2 8 304.0 150 3.15 3.435 17.30 0 0 3 2  
## Camaro Z28 13.3 8 350.0 245 3.73 3.840 15.41 0 0 3 4  
## Pontiac Firebird 19.2 8 400.0 175 3.08 3.845 17.05 0 0 3 2  
## Fiat X1-9 27.3 4 79.0 66 4.08 1.935 18.90 1 1 4 1  
## Porsche 914-2 26.0 4 120.3 91 4.43 2.140 16.70 0 1 5 2  
## Lotus Europa 30.4 4 95.1 113 3.77 1.513 16.90 1 1 5 2  
## Ford Pantera L 15.8 8 351.0 264 4.22 3.170 14.50 0 1 5 4  
## Ferrari Dino 19.7 6 145.0 175 3.62 2.770 15.50 0 1 5 6  
## Maserati Bora 15.0 8 301.0 335 3.54 3.570 14.60 0 1 5 8  
## Volvo 142E 21.4 4 121.0 109 4.11 2.780 18.60 1 1 4 2

## 2(a) Create linear model for weight (wt) vs. horsepower (hp)

wt\_James=mtcars$wt  
hp\_James=mtcars$hp  
Model\_James=lm(hp\_James ~wt\_James)  
summary(Model\_James)

##   
## Call:  
## lm(formula = hp\_James ~ wt\_James)  
##   
## 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\_James 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

## Create linear model for miles/gallon (mpg) vs. horsepower (hp)

mpg\_Chris=mtcars$mpg  
hp\_Chris=mtcars$hp  
Model\_Chris=lm(hp\_Chris ~mpg\_Chris)  
summary(Model\_Chris)

##   
## Call:  
## lm(formula = hp\_Chris ~ mpg\_Chris)  
##   
## 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\_Chris -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

It would appear from the results that Chris’s theory, when modeled, is more accurate a predictor of horsepower given mpg, than James’s theory of horsepower given weight. This can be concluded from the more favorable result of Chris’s r-squared value (0.6024) and lower standard error (43.95). Even still, Chris’s models results are, in themselves, not the most accurate either, yet they do a better job fitting the relationship between X and Y than does James’s model.

## 2(b) Build model to predict horsepower (hp) given number of cylinders (cyl) and miles/gallon (mpg)

cyl\_hpModel=mtcars$cyl  
mpg\_hpModel=mtcars$mpg  
hp\_hpModel=mtcars$hp  
hpModel=lm(hp\_hpModel ~cyl\_hpModel+mpg\_hpModel)  
summary(hpModel)

##   
## Call:  
## lm(formula = hp\_hpModel ~ cyl\_hpModel + mpg\_hpModel)  
##   
## 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\_hpModel 23.979 7.346 3.264 0.00281 \*\*  
## mpg\_hpModel -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

Given the solved model, our line equation is **Y = 54.067 + 23.979X1(cyl) - 2.775X2(mpg)**. So, a car with 4 cylinders and 22 miles/gallon, the horsepower is estimated to be **88.933 hp**.

## 3. Run provided code to load the dataset mlbench

#install.packages('mlbench')  
library(mlbench)  
data(BostonHousing)

## 3(a) Build model to predict median value of owner-occupied homes (medv)

X1\_crim=BostonHousing$crim  
X2\_zn=BostonHousing$zn  
X3\_ptratio=BostonHousing$ptratio  
X4\_chas=BostonHousing$chas  
Y\_medv=BostonHousing$medv  
medvModel=lm(Y\_medv ~X1\_crim+X2\_zn+X3\_ptratio+X4\_chas)  
summary(medvModel)

##   
## Call:  
## lm(formula = Y\_medv ~ X1\_crim + X2\_zn + X3\_ptratio + X4\_chas)  
##   
## 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 \*\*\*  
## X1\_crim -0.26018 0.04015 -6.480 2.20e-10 \*\*\*  
## X2\_zn 0.07073 0.01548 4.570 6.14e-06 \*\*\*  
## X3\_ptratio -1.49367 0.17144 -8.712 < 2e-16 \*\*\*  
## X4\_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

It appears that this model isn’t as accurate as it could be, given the low R^2 values. Only ~36% of the data is described by these four variables. This is likely because one or more of the input variables is statistically insignificant, and needs to be removed to re-formulate this model and produce more accurate prediction capabilities. Or, it could be that we didn’t capture enough of the data because we didn’t pick the correct makeup of variables to use from our dataset BostonHousing.

## 3(bI) What effect does bounding the Chas River produce on the house prices?

The estimated coefficient in our model’s results for the variable X4 pertaining to the Chas River is a positive number, meaning that it positively effects the Y output. Meaning, when houses are near the river the house prices increase, and the value of the estimated coefficient shows that a house by the Chas River would be roughly ~4.58 times more expensive than an identical house that wasn’t bounded by the Chas River. (NOTE: this factor seems quite high for reality, but we must remember that the model itself isn’t ideally accurate for prediction)

## 3(BII)(BONUS) What effect does pupil-teacher ratio have on the house prices?

The estimated coefficient in our model’s results for the variable X3 pertaining to the pupil-teacher ratio is a negative number, meaning that as the ratio grows (more kids per teacher) the housing prices fall proportionally. So, the house with a 15 ptratio would be more expensive than one with an 18 ptratio. That exact value is the difference between them is: (-1.49367 times 15)-(-1.49367 times 18) = 4.481 times more expensive.

## 3(c) Which of the variables are statistically important?

Based on the p-values of our model’s coefficients, it would seem that variable X4 (Chas River bounded) is the largest by far, and therefore the most insignificant when considering changes in the predictor versus changes in the response. However, when using the common 0.05 significance level, it would appear that all four variables’ p-values are less than 0.05, meaning all four are significant (statistically important) enough to keep in the model.

## 3(d) Anova Analysis to determine importance order

anova(medvModel)

## Analysis of Variance Table  
##   
## Response: Y\_medv  
## Df Sum Sq Mean Sq F value Pr(>F)   
## X1\_crim 1 6440.8 6440.8 118.007 < 2.2e-16 \*\*\*  
## X2\_zn 1 3554.3 3554.3 65.122 5.253e-15 \*\*\*  
## X3\_ptratio 1 4709.5 4709.5 86.287 < 2.2e-16 \*\*\*  
## X4\_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

Given the results of the Anova test, we can see the Sum Sq values in the table. Based on this, the order of importance for the modeled variables follows (from most to least): 1. Crime Zone 2. Pupil/Teacher Ratio 3. Zoning 4. Chas River Bounded

### END