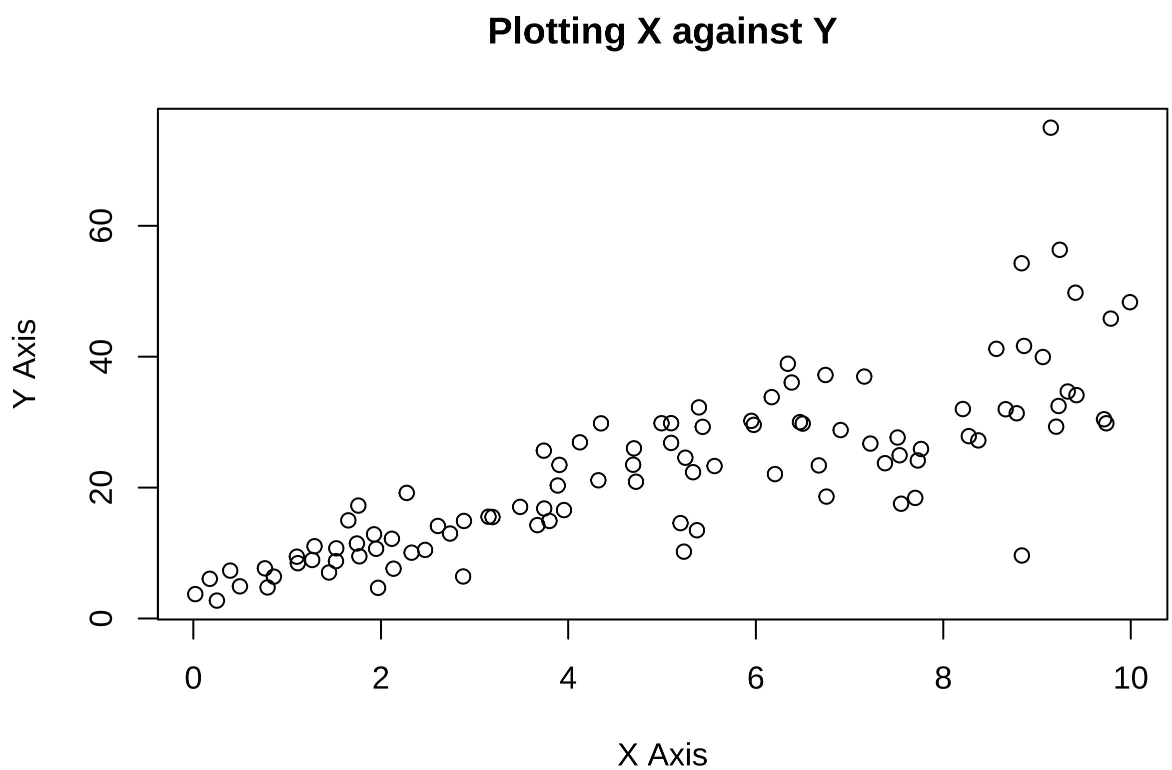
Assignment 2, Regression

By Jeremy Glasgow

jglasgo2@kent.edu


```r
# 1 a plot
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
plot(X, Y, main="Plotting X against Y", xlab="X Axis", ylab="Y Axis")
```




```r
# 1 a explanation
cat("Based on the plot above, there seems to be a moderate positive correlation \nbetween X and Y.")
```

```
## Based on the plot above, there seems to be a moderate positive correlation
## between X and Y.
```

```r
# 1 b
linearModel = lm(Y ~ X)
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
```

```r
# 1 b equation
rSquared <- summary(linearModel)$r.squared
rSquaredFormatted <- paste(round(rSquared, 4))

print(paste("Based on the summary results above and an R2 of", rSquaredFormatted, ", there is a \nlinea
```

```
## [1] "Based on the summary results above and an R2 of 0.6517 , there is a \nlinear relationship that w
```

```r
cat("The Linear Equation for this model is: Y = 4.4655 + (3.6108 * X) ")
```

```
## The Linear Equation for this model is: Y = 4.4655 + (3.6108 * X)
```

```r
cat()
```

```r
cc = cor(X, Y)
cat("Correlation Coefficient: ", cc)
```

```
## Correlation Coefficient:  0.807291
```

```r
cat("\nR Squared: ", rSquaredFormatted)
```

```
##
## R Squared:  0.6517
```

```r
cat("\nThe values of the correlation coefficient and the R Squared values \nboth showing a strong linea
```

```
##
## The values of the correlation coefficient and the R Squared values
## both showing a strong linear relationship.  These two metrics
## suggest the regression model fits the data very well.
```

```r
# 2 a
data(mtcars)
cat("Before constructing the linear model to find the best predictor of horsepower, \nI believe the bes
```

```
## Before constructing the linear model to find the best predictor of horsepower,
## I believe the best predictor would be MPG.  There are many other factors
## such as driving speed, type of engine, AWD vs RWD vs FWD, and person
## behind the wheel.  The higher the horsepower of the car, the harder it has to work, which
## means the more fuel consumption it would need to function.
```

```r
weight <- lm(hp ~ wt, data = mtcars)
mileage <- lm(hp ~ mpg, data = mtcars)
weightR2 <- summary(weight)$r.squared
mileageR2 <- summary(weight)$r.squared

cat("\n\nWeight as a predictor to HP has an R2 of:", weightR2)
```

```
##
##
## Weight as a predictor to HP has an R2 of: 0.4339488
```

```r
cat("\n\nMPG as a predictor to HP has an R2 of:", mileageR2)
```

```
##
##
## MPG as a predictor to HP has an R2 of: 0.4339488
```

```r
cat("\n\nBased on the findings above or MPG and car weight, their R2 scores are the exact same.")
```

```
##
##
## Based on the findings above or MPG and car weight, their R2 scores are the exact same.
```

```r
weightCC <- cor(mtcars$hp, mtcars$wt)
mileageCC <- cor(mtcars$hp, mtcars$mpg)

cat("\n\nWeight correlation coefficient:", weightCC)
```

```
##
##
## Weight correlation coefficient: 0.6587479
```

```r
cat("\n\nMPG correlation coefficient:", mileageCC)
```

```
##
##
## MPG correlation coefficient: -0.7761684
```

```r
cat("\n\nBased on the correlation coefficient, weight is the better predictor.  \nBoth R2 and CC indica
```

```
##
##
## Based on the correlation coefficient, weight is the better predictor.
## Both R2 and CC indicate weight is the better predictor for horsepower.
```

```r
# 2 b
data(mtcars)
newModel <- lm(hp ~ cyl + mpg, data = mtcars)
predData <- data.frame(cyl = 4, mpg = 22)
pred <- predict.lm(newModel, newdata = predData)
cat("The estimated horsepower for a car with a 4 cylinder engine that gets 22 miles \nper gallon is :",
```

```
## The estimated horsepower for a car with a 4 cylinder engine that gets 22 miles
## per gallon is : 88.93618
```

```r
# 3
# install.packages('mlbench')
library(mlbench)
data(BostonHousing)
```

```r
# 3 a
housingModel <- lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)
housingR2 <- summary(housingModel)$r.squared
cat("Boston Housing R2: ", housingR2)
```

```
## Boston Housing R2:  0.359859
```

```r
cat("\nBased on the R2 of the linear model, this falls in the moderately accurate \nclassification.  Th
```

```
##
## Based on the R2 of the linear model, this falls in the moderately accurate
## classification.  The cutoff value between weak and moderate is 0.33.  This model is
## barely in the moderate category and would likely not be a great model to make
## predictions from.
```

```r
# 3 b 1
chasCC <- housingModel$coefficients['chas1']
explanation <- paste("The only value in the linear model equation that would \nchange would be the chas
cat(explanation)
```

```
## The only value in the linear model equation that would
## change would be the chas1 value. The coefficient of chas1 is 4.58392590994697
## which would be the price difference between the two home values.
## The home with the tract bounds to the river would be more expensive by 4.58392590994697
```

```r
# 3 b 2
ptratioCC <- as.numeric(housingModel$coefficients['ptratio'])
house1 <- as.numeric(ptratioCC * 15)
house2 <- as.numeric(ptratioCC * 18)
cat("House 1 with a ratio of 15: ", house1)
```

```
## House 1 with a ratio of 15:  -22.40509
```

```r
cat("\nHouse 2 with a ratio of 18: ", house2)
```

```
##
## House 2 with a ratio of 18:  -26.88611
```

```r
cat("\nDifference: ", as.numeric(house1 - house2))
```

```
##
## Difference:  4.481018
```

```r
cat("\nBased on the ptratios between House 1 and House 2, the lower the \nptratio the higher the price 
```

```
##
## Based on the ptratios between House 1 and House 2, the lower the
## ptratio the higher the price will be.  This means House 1 is $ 4.481018
## more expensive than House 2.
```

```r
# 3 c
pValues <- summary(housingModel)$coefficients[, 'Pr(>|t|)']
print(pValues)
```

```
##  (Intercept)         crim           zn      ptratio        chas1
## 3.111877e-44 2.199766e-10 6.137848e-06 4.357372e-17 5.136898e-04
```

```r
cat("\nBased on the threshold of 0.05, all four of the variables are significant.")
```

```
##
## Based on the threshold of 0.05, all four of the variables are significant.
```

```r
# 3 d
# anova using p values
pAnova <- anova(housingModel)[, "Pr(>F)"]
pAnovaDF <- data.frame(Variables = rownames(anova(housingModel)), P_Values = pAnova)
print((pAnovaDF[order(pAnovaDF$P_Values), ]))
```

```
##   Variables     P_Values
## 1      crim 7.902220e-25
## 3   ptratio 4.738745e-19
## 2        zn 5.252886e-15
## 4      chas 5.136898e-04
## 5 Residuals           NA
```

```r
# 3 d
# using f score for anova
fAnova <- anova(housingModel)
impOrder <- order(-fAnova$`F value`)
orderedNames <- rownames(fAnova)[impOrder]
print(orderedNames)
```

```
## [1] "crim"      "ptratio"   "zn"        "chas"      "Residuals"
```