Assignment\_2

Rafsaan Sanvir

2024-07-08

Let’s do a residual analysis on the data in the datatrans.csv file, and process the data in order to prepare it for linear regression analysis.

1a) Exploring the table & Printing the first 6 observations

library(readr)  
setwd("/Users/user/Desktop/STAC67\_summer/Assignment\_2")  
datatrans <- read.csv("datatrans.csv")  
head(datatrans)

## x y  
## 1 8 0.2560425  
## 2 9 0.2078799  
## 3 10 0.1591369  
## 4 11 0.1837352  
## 5 12 0.1243692  
## 6 13 0.1259174

1b) Finding estimaated regression equation

x = datatrans$x  
y = datatrans$y  
fit = lm (y~x)  
summary(fit)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.034418 -0.019274 -0.006799 0.010658 0.153887   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1627957 0.0070070 23.23 <2e-16 \*\*\*  
## x -0.0014121 0.0001162 -12.15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03008 on 91 degrees of freedom  
## Multiple R-squared: 0.6188, Adjusted R-squared: 0.6146   
## F-statistic: 147.7 on 1 and 91 DF, p-value: < 2.2e-16

Estimated regression equation is Yi = 0.1627 -0.0014x + error

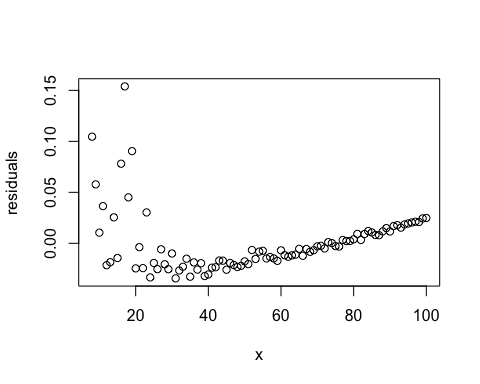
1c) Calculating predicted values and residuals. Creating table with x values, y values, poredicted values and residuals.

ypredicted = fit$coefficients[1] + (fit$coefficients[2]\*x)  
residuals = resid(fit)  
data = data.frame(  
 x\_values = x,  
 y\_values = y,  
 predicted\_values = ypredicted,  
 Residuals = residuals  
)  
head(data)

## x\_values y\_values predicted\_values Residuals  
## 1 8 0.2560425 0.1514990 0.10454344  
## 2 9 0.2078799 0.1500870 0.05779289  
## 3 10 0.1591369 0.1486749 0.01046203  
## 4 11 0.1837352 0.1472628 0.03647234  
## 5 12 0.1243692 0.1458507 -0.02148158  
## 6 13 0.1259174 0.1444387 -0.01852126

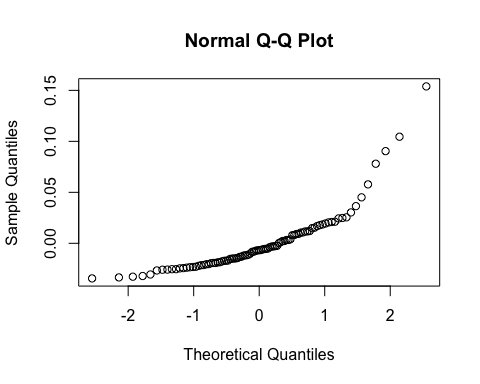
1. Plotting and analyzing residuals:

plot(residuals~x)

 Residuals are violating the constant variance assumption. It is also not clustered around 0. Residuals are showing a pattern, all suggesting the model might not be a good fit. Further tests need to be run in order to analyze residuals

Constructing Normal Quantile plot:

qqnorm(residuals)

 This graph does not follow a straight line, indicating that the residuals might not be normally distributed. However, this is a visual representation, and other statistical tests need to be ran in order to confirm this. Graph also seems to indicate heavy tailed data.

Let’s try a shapiro Wilk test to test whether the residuals came from a normally distributed population, tested at significance level alpha = 0.05 :

shapiro.test(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.77835, p-value = 1.592e-10

since p value << alpha = 0.05, reject null hypothesis. In Shapiro-Wilk test, Null hypotheis is that residuals are normally distributed, so we can conclude from this test that residuals are not normally distributed.

1. Let’s test whether errors have constant variance with the Brown-Forsythe test

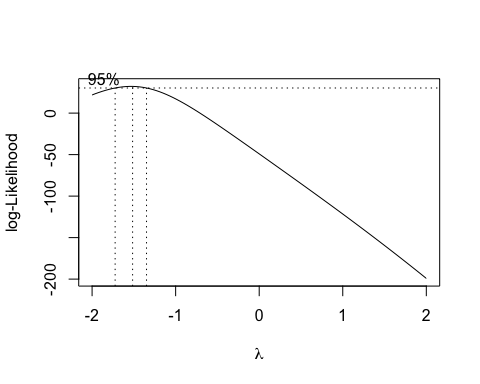
table = cbind(x, residuals)  
orderedbyx = table[order(x),] #reorders rows of table according to sorted order x  
eorderedbyx = orderedbyx[,2]  
n = nrow(table)  
n1 = ceiling(n/2)  
n2 = n - n1  
e1 = eorderedbyx[1:n1]  
e2 = eorderedbyx[(n1+1):n]  
med1 = median(e1) #find median of   
med2 = median(e2)  
  
#find d1, d2 and sample means of d1, d2  
d1 = abs(e1 - med1)   
d2 = abs(e2 - med2)  
dbar1 = mean(d1)  
dbar2 = mean(d2)  
  
#find test statistic and p value  
s\_sq = (sum((d1-dbar1)^2) + sum((d2-dbar2)^2)) / (n-2) #s^2 formula  
tBF = (dbar1-dbar2) / sqrt(s\_sq\*((1/n1)+(1/n2)))  
p\_value = 2\*(1-pt(abs(tBF), n-2))  
p\_value

## [1] 0.03824042

Since p\_value < alpha = 0.05, we reject Null hypothesis of the Brown-Forsythe test which states that variance of erros are equal, concluding that the errors violate the constant variance assumption.

1. Let’s improve the model by using Box-Cox transformation to find a better fit.

#Find MLE estimate of lambda  
library(MASS)  
bc = boxcox(y~x) #graph shows MLE estiamte close to -1.5



lambda = bc$x[which.max(bc$y)] #"give me the value of x that maximizes y"  
lambda

## [1] -1.515152

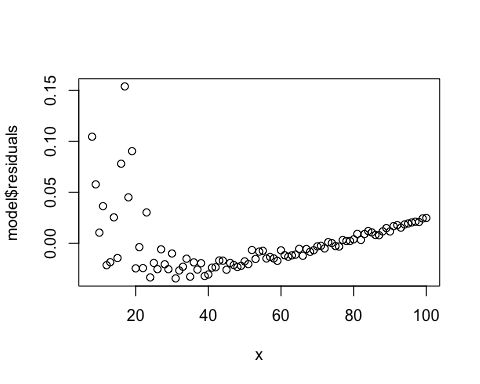
1. We have lambda value, now we transform y

K2 = prod(y)^(1/length(y))  
#new linear regression model according to   
yt = ((y^lambda) - 1) /( lambda \* (K2^(lambda-1)))   
yt[1:6] #printing first 6 values of the transformed y

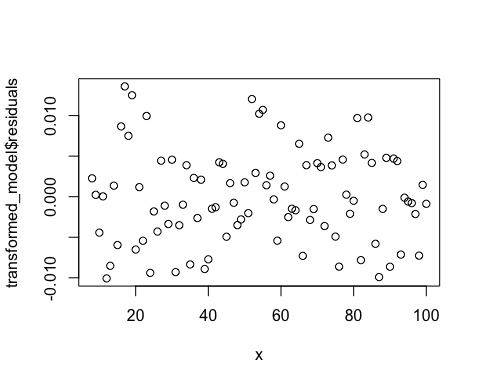
## [1] -0.00733747 -0.01045758 -0.01620888 -0.01282820 -0.02403141 -0.02356532

1. Let’s plot this tranformed y

transformed\_model = lm(yt~x)  
model = fit  
plot(model$residuals~x)

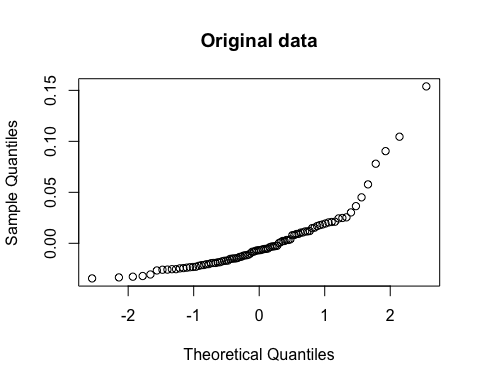


plot(transformed\_model$residuals~x)

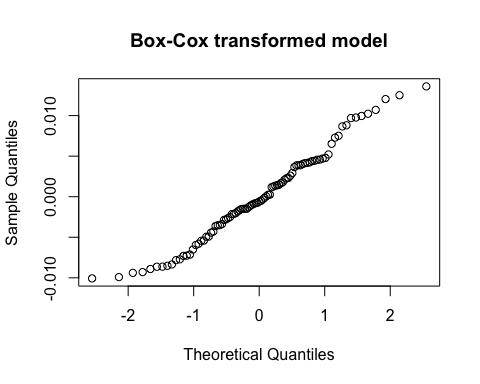
 Transformed model’s data is clustered around zero, demonstrates constant variance and no apparent patterns compared to original model. Indicates that transformed model is a much fit for data than original model for linear regression analysis.

1. Plotting original vs transformed normal qq plot

qqnorm(fit$residuals, main = "Original data")



qqnorm(transformed\_model$residuals, main = "Box-Cox transformed model")

 Box-Cox transformation model’s data is much closer to a straight line than the original model’s data! meaning it is much closer to normal than the original model.

1. Using Shapiro Wilk for normality test of transformed model:

shapiro.test(transformed\_model$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: transformed\_model$residuals  
## W = 0.97732, p-value = 0.1052

Since p\_value > alpha = 0.05, fail to reject null hypothesis that the residuals are normally distributed, indicating that residuals follow normal distribution

1. Brown-Forsythe Test For constant variance of error:

table = cbind(x, transformed\_model$residuals)  
orderedbyx = table[order(x),] #reorders rows of table according to sorted order x  
eorderedbyx = orderedbyx[,2]  
n = nrow(table)  
n1 = ceiling(n/2)  
n2 = n - n1  
e1 = eorderedbyx[1:n1]  
e2 = eorderedbyx[(n1+1):n]  
med1 = median(e1) #find median of   
med2 = median(e2)  
  
#find d1, d2 and sample means of d1, d2  
d1 = abs(e1 - med1)   
d2 = abs(e2 - med2)  
dbar1 = mean(d1)  
dbar2 = mean(d2)  
  
#find test statistic and p value  
s\_sq = (sum((d1-dbar1)^2) + sum((d2-dbar2)^2)) / (n-2) #s^2 formula  
tBF = (dbar1-dbar2) / sqrt(s\_sq\*((1/n1)+(1/n2)))  
p\_value = 2\*(1-pt(abs(tBF), n-2))  
p\_value

## [1] 0.4421956

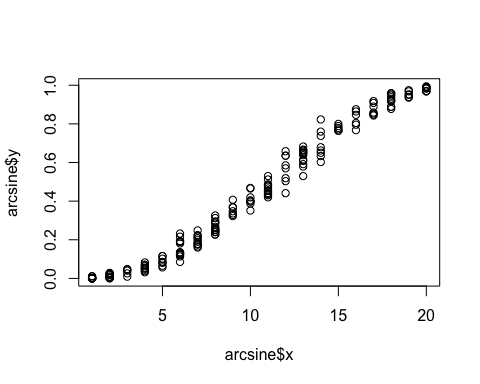
Since p\_value > alpha = 0.05, reject null hypothesis that resiudals violate constant variance assumption of residuals, concluding that residuals have constant variance

We have a transformed model that can now be used for linear regression analysis!

Now, let;s explore data in the arcsine csv file, containing relationship between the dose (x) of an insecticide and the proportion of insects killed (y).

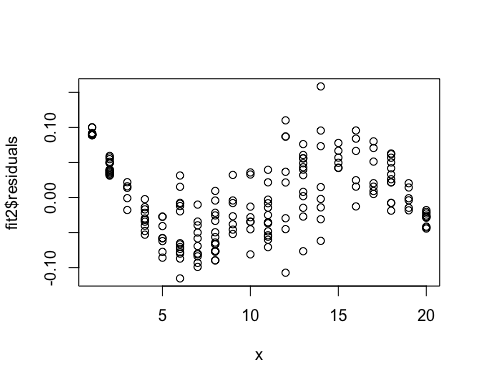
2a) Explore data related to arcsine csv file.

library(readr)  
setwd("/Users/user/Desktop/STAC67\_summer/Assignment\_2")  
arcsine <- read.csv("arcsineTr1.csv")  
x = arcsine$x  
y = arcsine$y  
plot(arcsine$x, arcsine$y)

 There are multiple y values mapping to the given x values. The plot is also a curved line, visually indicating a non-linear relationship.

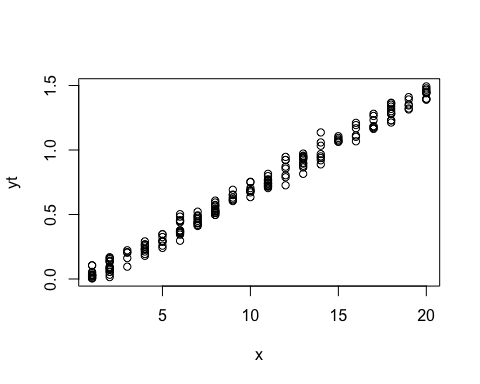
1. fit regression model & create residual plot:

fit2 = lm(y~x)  
plot(x, fit2$residuals)

 The residual plot is clustered around 0, but follows a clear pattern which indicates a bad fit. Also violates constant variance assumptions at certain points of the graph.

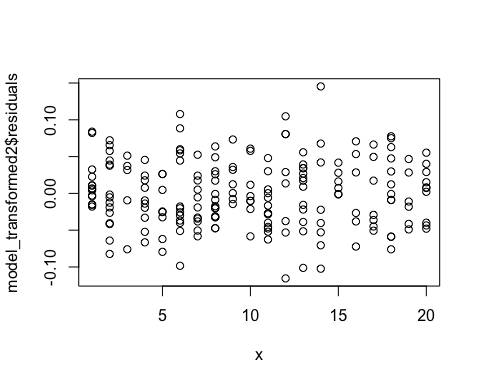
* 1. Let’s use y′ = arcsine(√y) transformation on data

yt = asin(sqrt(y))  
plot(x, yt)

 Still multiple y values mapping to one x, but the graph is no longer curved, indicating a possible linear relationship

1. Let’s check if this transformed model is good to run linear regression.

model\_transformed2 = lm(yt~x)  
plot(x, model\_transformed2$residuals)

 Compared to original residual plot, the residuals are scattered around 0, indicating a mean of 0 which is good. Variance seems to be constant, and no patterns seem to exist in the shape of the data, showing that the new model is a better fit for the data.

3a) The hospital.txt datafile contains relationship between hospital charges (y in dollars) and age (x in years). Let’s fit two separate regression lines Yf = β0f + β1fxf + ϵf and Ym = β0m + β1mxm + ϵm for females and male patients.

setwd("/Users/user/Desktop/STAC67\_summer/Assignment\_2")  
hospital = read.table("hospital.txt", header = TRUE)  
  
sex = hospital[,1]  
charge = hospital[,2]  
age = hospital[,3]  
hospital = cbind(sex, charge, age)  
hospital = as.data.frame(hospital)  
hospital = transform(hospital, charge = as.numeric(charge), age = as.numeric(age))  
female = subset(hospital, sex == "F")  
male = subset(hospital, sex == "M")  
female\_model = lm(charge~age, data = female)  
male\_model = lm(charge~age, data = male)  
summary(male\_model)

##   
## Call:  
## lm(formula = charge ~ age, data = male)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6778 -3915 -1402 3842 8968   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4035.97 4491.50 -0.899 0.381428   
## age 290.74 70.36 4.132 0.000697 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5157 on 17 degrees of freedom  
## Multiple R-squared: 0.5011, Adjusted R-squared: 0.4717   
## F-statistic: 17.07 on 1 and 17 DF, p-value: 0.0006967

summary(female\_model)

##   
## Call:  
## lm(formula = charge ~ age, data = female)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14033 -4293 -2413 3525 44855   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7203.7 12575.8 -0.573 0.5731   
## age 377.7 172.4 2.190 0.0405 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11660 on 20 degrees of freedom  
## Multiple R-squared: 0.1935, Adjusted R-squared: 0.1532   
## F-statistic: 4.798 on 1 and 20 DF, p-value: 0.04051

3bi) The difference in the average (i.e. mean) changes for females and males with the same age x = x0 is E(Yf |xf = x0) − E(Ym|xm = x0). This is estimated by ˆyf−ˆym and ˆyf−ˆym ∼ N(E(Yf |xf = x0)−E(Ym|xm = x0), kσ2). Let’s calculate the value of k when x0 = 60

x0 = 60  
n\_f = nrow(female)  
n\_m = nrow(male)  
  
xbar\_f = mean(female$age)  
xbar\_m = mean(male$age)  
sxx\_f = sum((female$age - xbar\_f)^2)  
sxx\_m = sum((male$age - xbar\_m)^2)  
  
k1 = ((1/n\_f) + (((x0 - xbar\_f)^2) / sxx\_f)) #MSE not included  
k2 = ((1/n\_m) + (((x0 - xbar\_m)^2) / sxx\_m)) #MSE not included   
k = k1 + k2  
k

## [1] 0.1274541

This equation is derived from expanding the equation of variance(yhat\_female -yhat\_male) = variance (yhat\_female) + variance(yhat\_male) = sigma^2\* (k1) + sigma^2 \* (k2) where equation for k1, k2 is in code above. = sigma^2 (k1+k2), so k = k1+k2 according to the question.