R Assignment 5

Jacey Davies

11/23/2020

Using the data file “mydata.csv”

1. Create a scatterplot of y vs x

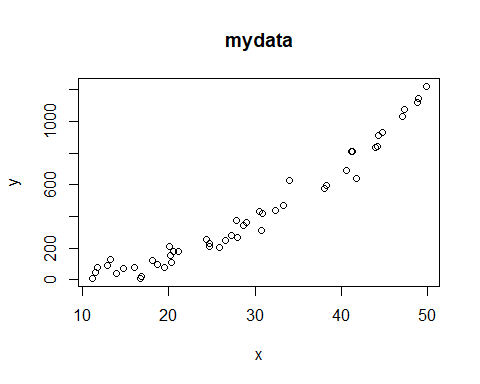
Putting the data into a data frame:

mydata = data.frame(y = c(311.848077,440.942820,41.674404,417.743549,177.364193,639.072713,179.923471,19.649626,1030.218388,211.607839,468.797015,281.96407,360.414913,626.325417,692.87202,840.811552,71.517735,97.756431,251.069676,81.512875,270.344452,1221.872769,110.315154,595.441155,126.218820,11.159988,230.554237,77.302501,1117.462801,122.568383,932.664982,911.059895,255.662452,810.009706,210.474471,9.884425,75.983619,153.659519,578.725421,93.283791,378.110173,203.940822,837.901765,44.456711,1145.790170,1073.485056,431.139416,343.550433,810.066458),  
 x = c(30.77326,32.40036,13.89724,30.82836,21.17247,41.70052,20.52949,16.78782,47.05621,24.73312,33.30568,27.20706,28.98507,33.98696,40.61913,44.14024,14.71966,18.69047,26.53534,19.51529,28.00065,49.81578,20.33470,38.29436,13.26268,16.73084,24.64804,15.99319,48.85320,18.10108,44.75007,44.23208,24.33537,41.18667,20.06741,11.10681,11.67823,20.20392,38.05732,12.89079,27.82776,25.83180,43.87759,11.49288,48.94833,47.30910,30.53461,28.65658,41.25828))  
mydata

## y x  
## 1 311.848077 30.77326  
## 2 440.942820 32.40036  
## 3 41.674404 13.89724  
## 4 417.743549 30.82836  
## 5 177.364193 21.17247  
## 6 639.072713 41.70052  
## 7 179.923471 20.52949  
## 8 19.649626 16.78782  
## 9 1030.218388 47.05621  
## 10 211.607839 24.73312  
## 11 468.797015 33.30568  
## 12 281.964070 27.20706  
## 13 360.414913 28.98507  
## 14 626.325417 33.98696  
## 15 692.872020 40.61913  
## 16 840.811552 44.14024  
## 17 71.517735 14.71966  
## 18 97.756431 18.69047  
## 19 251.069676 26.53534  
## 20 81.512875 19.51529  
## 21 270.344452 28.00065  
## 22 1221.872769 49.81578  
## 23 110.315154 20.33470  
## 24 595.441155 38.29436  
## 25 126.218820 13.26268  
## 26 11.159988 16.73084  
## 27 230.554237 24.64804  
## 28 77.302501 15.99319  
## 29 1117.462801 48.85320  
## 30 122.568383 18.10108  
## 31 932.664982 44.75007  
## 32 911.059895 44.23208  
## 33 255.662452 24.33537  
## 34 810.009706 41.18667  
## 35 210.474471 20.06741  
## 36 9.884425 11.10681  
## 37 75.983619 11.67823  
## 38 153.659519 20.20392  
## 39 578.725421 38.05732  
## 40 93.283791 12.89079  
## 41 378.110173 27.82776  
## 42 203.940822 25.83180  
## 43 837.901765 43.87759  
## 44 44.456711 11.49288  
## 45 1145.790170 48.94833  
## 46 1073.485056 47.30910  
## 47 431.139416 30.53461  
## 48 343.550433 28.65658  
## 49 810.066458 41.25828

Creating a scatterplot:

plot(mydata$x,mydata$y, xlab="x", ylab="y", main="mydata")



1. Fit a simple linear regression model using y as the response and plot the regression line (with the data)

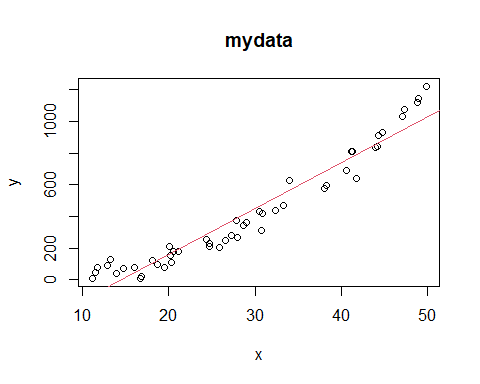
Fitting the model:

reg = lm(y~x, data=mydata)  
summary(reg)

##   
## Call:  
## lm(formula = y ~ x, data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -159.43 -63.69 -15.73 59.37 198.85   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -420.366 34.637 -12.14 4.33e-16 \*\*\*  
## x 28.975 1.111 26.09 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 91.19 on 47 degrees of freedom  
## Multiple R-squared: 0.9354, Adjusted R-squared: 0.934   
## F-statistic: 680.5 on 1 and 47 DF, p-value: < 2.2e-16

Plotting the regression line:

plot(mydata$x,mydata$y, xlab="x", ylab="y", main="mydata")  
abline(-420.366,28.975,col=2)



1. Test whether x is a significant predictor and create a 95% CI around the slope coefficient.

H0: β1 = 0

Ha: β1 ≠ 0

The test has been performed by R and the p-value, as seen in the summary of “reg”, is less than 2 x 10-16. This is very small, thus H0 is rejected and x is a significant predictor.

95% CI:

28.975 + c(-1,1)\*qt(.975,47)\*1.111

## [1] 26.73996 31.21004

95% CI: (26.7400, 31.2100)

1. Report and interpret the coefficient of determination.

R2 = 0.9354

The coefficient of determination is the percentage variation in y explained by x-variables. A good fitting model would have an R2 value as close to 1 as possible, as that means more of the data points land near or on the regression line. This value is very close to 1 and thus the model seems to be close to the actual data points.

1. For x = 20, create a CI for E(Y|X = 20).

Creating a 95% CI for E(Y|X = 20):

newdata = data.frame(x=20)  
predict(reg,newdata,interval="confidence",level=0.95)

## fit lwr upr  
## 1 159.1262 126.2353 192.0172

95% CI: (126.2353, 192.0172)

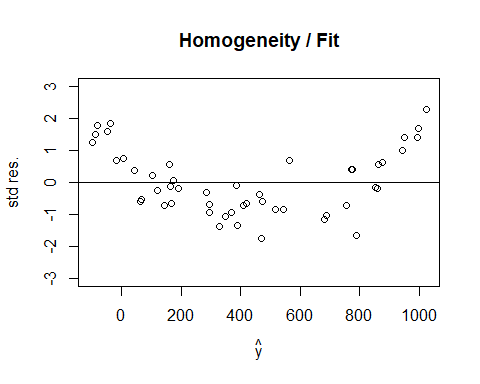
1. For x = 150, can you use the model to estimate E(Y|X = 150)? Discuss.

You cannot use the model to estimate E(Y|X = 150) because x = 150 lies outside the range of the given data. We have no way of knowing how the data will act beyond x = 50 and thus we do not know if the model is a good fit beyond that point.

1. Does the model appear to be linear with respect to x? Discuss, and if not, provide alternative model and repeat steps 1-6.

When checking the model for fit, it is clear that a linear regression model is not valid.

re = rstandard(reg) #standardized residuals  
ylimits=c(-3,3)  
ylimits[1]=ifelse(min(re)<(-3),min(re),-3)  
ylimits[2]=ifelse(max(re)>(3),max(re),3)  
  
# Homogeneity of variance/Model Fit  
plot(re~fitted.values(reg),xlab=expression(hat(y)),ylab="std res.",main="Homogeneity / Fit",ylim=ylimits)  
abline(h=0)



The data points are not evenly distributed above and below the zero line. The points are above the line on the edges and below the line in the middle. Thus one of the assumptions is not met and the model is not linear with respect to x.

This can also be seen with boxcox() and powerTransform():

#Using Box-Cox  
library(MASS)  
bc1=boxcox(reg)

bc1$x[which.max(bc1$y)]

## [1] 0.5858586

#Using Power Transform  
library(car)

bc2=powerTransform(y~x, data=mydata)  
summary(bc2)

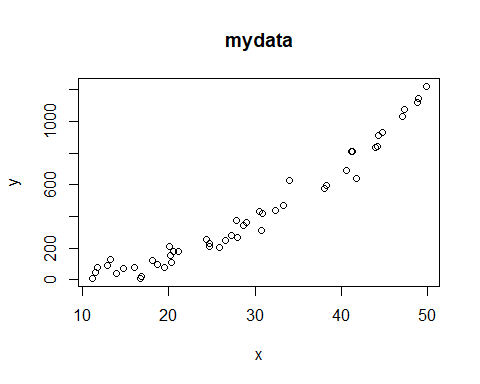
## bcPower Transformation to Normality   
## Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd  
## Y1 0.5816 0.5 0.4827 0.6806  
##   
## Likelihood ratio test that transformation parameter is equal to 0  
## (log transformation)  
## LRT df pval  
## LR test, lambda = (0) 80.22636 1 < 2.22e-16  
##   
## Likelihood ratio test that no transformation is needed  
## LRT df pval  
## LR test, lambda = (1) 44.4888 1 2.5582e-11

With both tests, the estimated lambda ≠ 1 and the p-value for the test if lambda = 1 is very small. This means that a power transformation is necessary and the model needs to be non-linear.

Alternative:

1. Create a scatterplot of y vs x

plot(mydata$x,mydata$y, xlab="x", ylab="y", main="mydata")



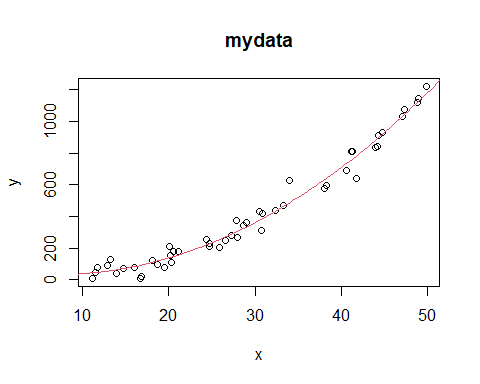
1. Fit a simple linear regression model using y as the response and plot the regression line (with the data)

reg2 = lm(y~x+I(x^2), data = mydata)  
summary(reg2)

##   
## Call:  
## lm(formula = y ~ x + I(x^2), data = mydata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -142.880 -34.391 3.912 29.554 140.530   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.58227 49.46654 1.326 0.1915   
## x -8.85144 3.61118 -2.451 0.0181 \*   
## I(x^2) 0.62422 0.05875 10.624 5.72e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 49.6 on 46 degrees of freedom  
## Multiple R-squared: 0.9813, Adjusted R-squared: 0.9805   
## F-statistic: 1207 on 2 and 46 DF, p-value: < 2.2e-16

Plotting the regression line:

plot(mydata$x,mydata$y, xlab="x", ylab="y", main="mydata")  
curve(65.58227-8.85144\*x+0.62422\*x^2,from=9,to=52,col=2,add=TRUE)



1. Test whether x is a significant predictor and create a 95% CI around the slope coefficient.

H0: β1 = 0

Ha: β1 ≠ 0

The test has been performed by R and the p-value, as seen in the summary of “reg”, is 0.0181. This is smaller than the error of 0.05, thus H0 is rejected and x is a significant predictor.

95% CI:

-8.85144 + c(-1,1)\*qt(.975,46)\*3.61118

## [1] -16.120368 -1.582512

95% CI: (-16.1204, -1.5825)

1. Report and interpret the coefficient of determination.

R2 = 0.9813

This value is even closer to 1 and thus the model seems to be a good fit.

1. For x = 20, create a CI for E(Y|X = 20).

Creating a 95% CI for E(Y|X = 20):

predict(reg2,newdata,interval="confidence",level=0.95)

## fit lwr upr  
## 1 138.2421 119.9104 156.5737

95% CI: (119.9104, 156.5737)

1. For x = 150, can you use the model to estimate E(Y|X = 150)? Discuss.

Once again, you cannot use the model to estimate E(Y|X = 150) because x = 150 lies outside the range of the given data. You can’t extrapolate data with the regression model since the behavior of the data outside the given range is unknown.