**Exercise 8.2 - Model Diagnostics:** Abdul Wasay

install.packages("psych")

install.packages("nloptr")

install.packages("car")

library(psych)

library(car)

#Obtain the estimate regression function

TBPY <- read.csv("TBCasesinUKUpdated.csv", stringsAsFactors=FALSE)

# Build linear model

RegTB<- lm(NoOfBTBCases ~ Year, data=TBPY

summary(RegTB)

Residuals:

Min 1Q Median 3Q Max

-7.515 -3.529 1.157 3.135 6.814

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3312.552 457.663 -7.238 2.89e-06 \*\*\*

Year 1.664 0.228 7.298 2.62e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.606 on 15 degrees of freedom

Multiple R-squared: 0.7803, Adjusted R-squared: 0.7656

F-statistic: 53.26 on 1 and 15 DF, p-value: 2.617e-06

#Interpret the results – p-value for predictor, p-value for F-test and Adj.R-squared

#The small p-value of 2.617e-06 is a lot below 0.05 which suggests a good co-relation between the two variables. The Ad. R^2 value of 0.7656 suggests that more than 76% of the data is able to support the model and hence the model can be trusted. The F-test of 53.26 suggests that the data on the whole lacks errors.

plot(RegTB, which=1)

#check the plot for constant variance, i.e. there is no specific pattern to the plot

res = resid(RegTB)

summary(lm(abs(res)~fitted(RegTB))) #dirty test for a linear relationship

# -------------------------------------------------------------------

Call:

lm(formula = abs(res) ~ fitted(RegTB))

Residuals:

Min 1Q Median 3Q Max

-3.9886 -1.6421 -0.2226 1.5479 3.4066

Coefficients:

Estimate Std. Error t value Pr(>|t|)

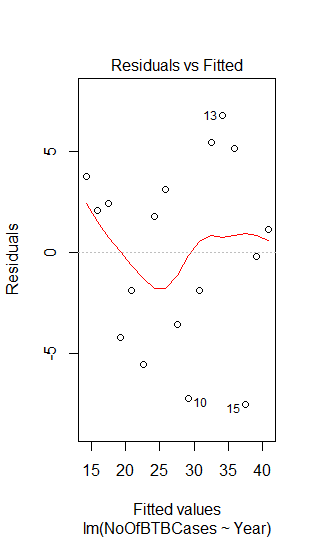
(Intercept) 2.77022 1.94111 1.427 0.174

fitted(RegTB) 0.03566 0.06761 0.528 0.606

Residual standard error: 2.273 on 15 degrees of freedom

Multiple R-squared: 0.01821, Adjusted R-squared: -0.04724

F-statistic: 0.2783 on 1 and 15 DF, p-value: 0.6056



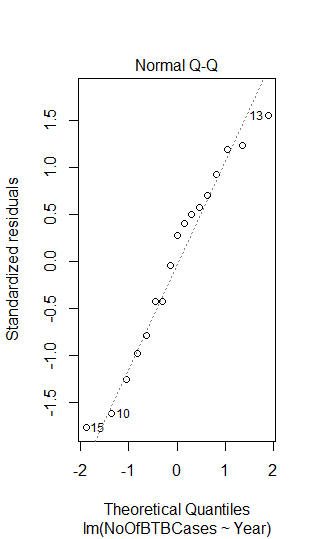
summary(res) #Look at how the residues are distributed and it gives an indication of normality

plot(RegTB,which=2)

#check normal plot, explain the nature of the plot – should be close to the normal line

Min. 1st Qu. Median Mean 3rd Qu. Max.

-7.515 -3.529 1.157 0.000 3.135 6.814



shapiro.test(res)

#do a test for normality – high p-value means you cannot reject the null hypothesis that the residues are normally distributed

Shapiro-Wilk normality test

data: res

W = 0.95389, p-value = 0.5208

durbinWatsonTest(RegTB)

#do a Durbin Watson Test - high p-value indicates you cannot reject the null hypothesis that the residuals are not correlated

# -------------------------------------------------------------------

lag Autocorrelation D-W Statistic p-value

1 0.2174113 1.51597 0.186

Alternative hypothesis: rho != 0

Q1. Regression Model

TBCasesInUK = -3312.552 + 1.664\*Year

Q2. 5% Significance Level

H0: the number of cases of TB in the UK do not have a relation with the years that have gone by

H1: the number of cases of TB in the UK are affected by the number of years that have gone by

p-value: 2.617e-06 is less than test statistic 0.05, thus we can safely reject the null hypothesis up to a 95% significance level.

Q3. 4 Basic Assumptions of Linear Regression

***Assumption 1****: Linearity, Multicollinearity, significance*

Residuals:

Min 1Q Median 3Q Max

-7.515 -3.529 1.157 3.135 6.814

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3312.552 457.663 -7.238 2.89e-06 \*\*\*

Year 1.664 0.228 7.298 2.62e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.606 on 15 degrees of freedom

Multiple R-squared: 0.7803, Adjusted R-squared: 0.7656

F-statistic: 53.26 on 1 and 15 DF, p-value: 2.617e-06

The p value (2.617e-06) is less than the value of alpha (0.05) so there is a correlation between no of tb cases and the years that have passed.

***Assumption 2****: Variance of error are constant for predicted values*

Residuals:

Min 1Q Median 3Q Max

-3.9886 -1.6421 -0.2226 1.5479 3.4066

Coefficients:

Estimate Std. Error t value Pr(>|t|)

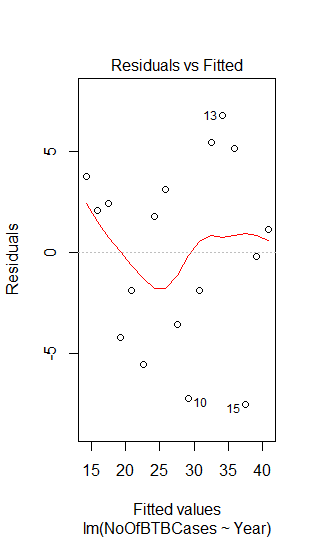
(Intercept) 2.77022 1.94111 1.427 0.174

fitted(RegTB) 0.03566 0.06761 0.528 0.606

Residual standard error: 2.273 on 15 degrees of freedom

Multiple R-squared: 0.01821, Adjusted R-squared: -0.04724

F-statistic: 0.2783 on 1 and 15 DF, p-value: 0.6056



There doesn’tseem to be a pattern in between the residuals and fitted values. The high p-value and low R^2 value and F-statistic suggest that there isn’t a correlation in the model.

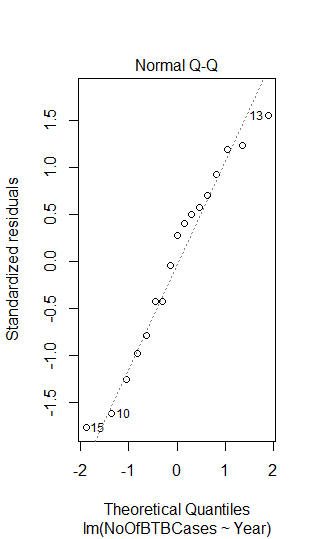
***Assumption 3****: the error terms are normally distributed*

residues are normally distributed

Shapiro-Wilk normality test

data: res

W = 0.95389, p-value = 0.5208



The p-value of 0.5208 is a lot higher than the test statistic of 0.05, hence we will reject the alternate hypothesis. There is a greater chance of residuals being normally distributed.

***Assumption 4****: Observations are independent (residuals are not correlated)*

lag Autocorrelation D-W Statistic p-value

1 0.2174113 1.51597 0.186

Alternative hypothesis: rho != 0

A positive co-relation is suggested with a D-W value of less than 2. Also, since the p-value of 1.86 is very high than the test statistic of 0.05, we reject the alternate hypothesis and accept the null hypothesis.