Lab 9

Tian Qiu 00265 35063

1.

airline <- read.table(file = "airline\_cleaned.txt", header = TRUE)

attach(airline)

localAirline <- subset(airline, ArrDelay < 60 & (TaxiIn + TaxiOut) < 60 & Distance <= 2704 )

detach(airline)

detach(studynew)

attach(ArrDelay)

#1)

library(lattice)

xyplot(localAirline$ActualElapsedTime ~ Distance,

data = localAirline,

panel = function(x, y){

panel.xyplot(x, y)

panel.lmline(x, y)

})

#2) correlation

cor(Distance, ActualElapsedTime)

#c), d), i) calculate linear regression and get results

local.lm = lm(localAirline$ActualElapsedTime ~ Distance)

summary(local.lm)

#7)

x <- subset(airline, Month == 11 & DayofMonth == 7 & Origin == "MSP" & Dest == "MKE" & DepTime == 1529)

x$Distance

x$ActualElapsedTime

#8) calculate the residuals

local.resid = local.lm$res #Extract residuals obtained in job.lm operation

xyplot(local.resid ~ Distance,

data = local.lm,

main="Residual plot",

ylab = "Residual",

panel = function(x, y){

panel.xyplot(x, y)

panel.abline(h = 0)

})

# f)

# Calculate the histogram and qqplot on the residuals please see previous labs for this

# Note: this is a single sample

# Generate the 2-sided Confidence Interval (CI) for the parameters

histogram(local.resid, type="density",

panel=function(x,...)

{panel.histogram(x,...)

panel.mathdensity(dmath=dnorm,col="blue",lwd=2,

args=list(mean=mean(x, na.rm=T), sd = sd(x,na.rm=T)),...)

panel.densityplot(x,col="red",lwd=2,...)

})

qqmath(local.resid, panel = function(x){

panel.qqmath(x)

panel.qqmathline(x)

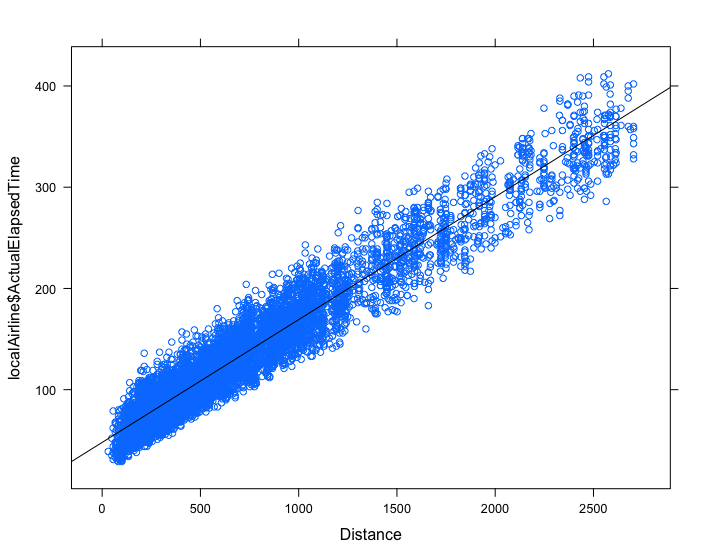
})

#h)

confint(local.lm, level = 0.99)

#NOTE: This can also be done by hand from output of summary(job.lm) > # However, in this lab, you must use the code above.

2.



3.

This scatter plot shows that distance and actual elapsed time are positive and strong linear relationship. I cannot see any obvious outliers. Yes it is linear.

4.

[1] 0.960428

The correlation coefficient between distance and actual elapsed time is 0.960428. This looks like there is strong association between distance and actual elapsed time. Therefore, the strength is tremendous and it is the same as problem 3.

5.

Yes, the correlation is a good numerical summary of the graphical display in the scatterplot. Because correlation is used to describe the linear relationship between two continuous variables. In general, correlation tends to be used when there is no identified response variable.

6.

Call:

lm(formula = localAirline$ActualElapsedTime ~ Distance)

Residuals:

Min 1Q Median 3Q Max

-72.97 -14.98 -1.14 13.42 71.19

Coefficients:

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

(Intercept) 4.793e+01 3.979e-01 120.4 <2e-16 \*\*\*

Distance 1.213e-01 4.263e-04 284.4 <2e-16 \*\*\*

---

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

Residual standard error: 19.61 on 6804 degrees of freedom

Multiple R-squared: 0.9224, Adjusted R-squared: 0.9224

F-statistic: 8.09e+04 on 1 and 6804 DF, p-value: < 2.2e-16

7.

> x$ActualElapsedTime

[1] 78

A = 47.93 + 0.1213D

D = 297

A = 83.9561

Residual = A – 78 = 5.9561

7. bonus

Distance of this time is exactly the same as 297.

> x$ActualElapsedTime

[1] 78

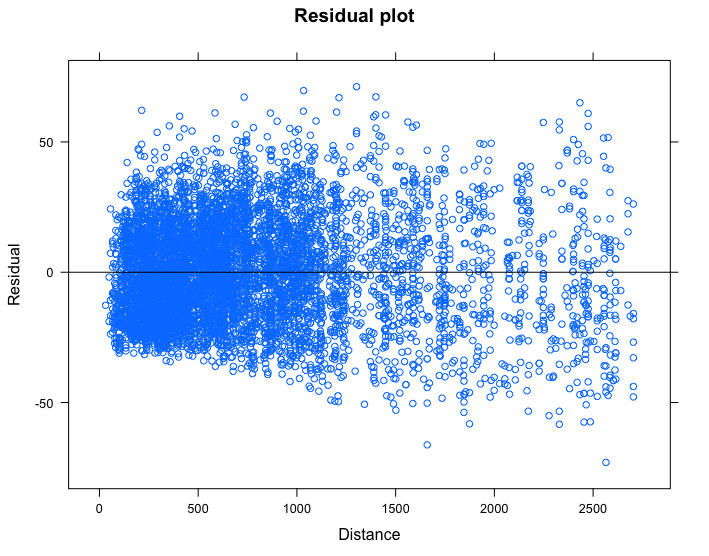
A = 47.93 + 0.1213D

D = 297

A = 83.9561

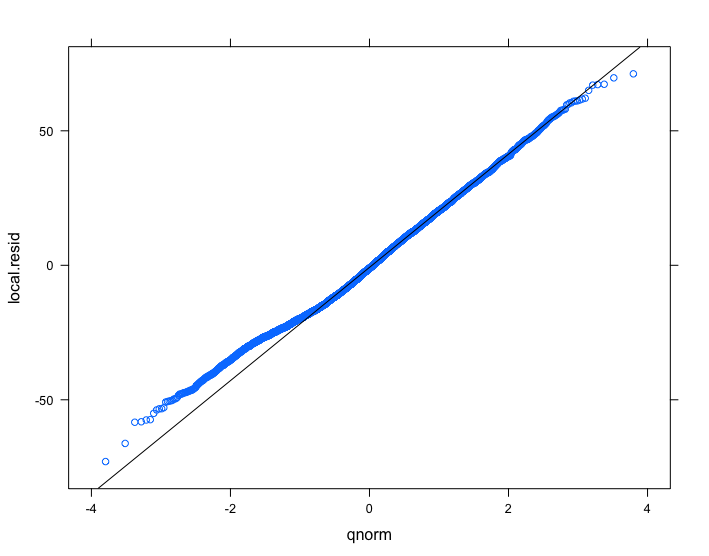
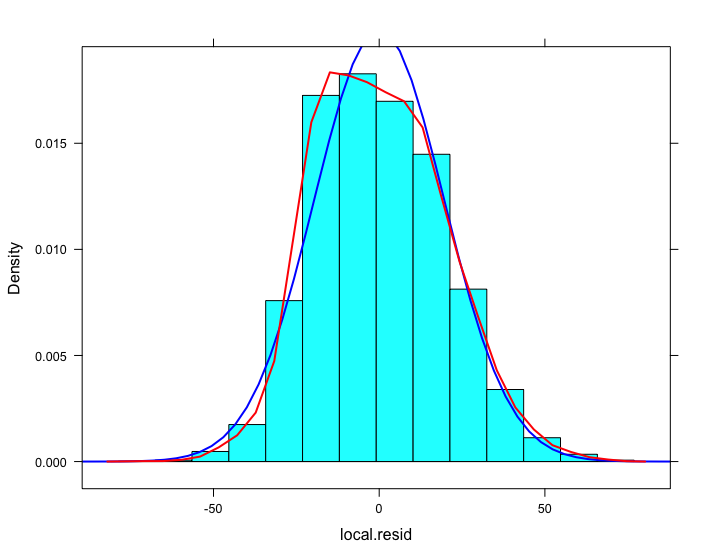
Residual = A – 78 = 5.9561

8.



I see no pattern here so the association seems to be linear. I do not see any outliers. The answer is corresponding with part3 and part 5 which both shows the relationship between x and y is linear.

9.



It looks like the residuals are normal because on the QQ plot the points are close to the line and the blue/red lines on the histogram seems to be close. So I believe residuals appear to be normal.

10.

Assuming that we have an SRS, the three other assumptions are met; linear, constant standard deviation of the residuals and normality of the residuals, therefore linear regression analysis appears to be reasonable.

11.

0.5 % 99.5 %

(Intercept) 46.9011956 48.9517804

Distance 0.1201661 0.1223631

Slope

95% CI (0.1201661, 0.1223631)

We are 95% confident that the population slope of Stress vs. LOC is between0.1201661 and 0.1223631.

Intercept

95% CI (46.9011956 , 48.9517804)

We are 95% confident that the population y-intercept of Stress vs. LOC is between 46.9011956 and 48.9517804.

12.

284.4 <2e-16 \*\*\*

Residuals:

Min 1Q Median 3Q Max

-72.97 -14.98 -1.14 13.42 71.19

Coefficients:

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

(Intercept) 4.793e+01 3.979e-01 120.4 <2e-16 \*\*\*

Distance 1.213e-01 4.263e-04 284.4 <2e-16 \*\*\*

---

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

Residual standard error: 19.61 on 6804 degrees of freedom

Multiple R-squared: 0.9224, Adjusted R-squared: 0.9224

F-statistic: 8.09e+04 on 1 and 6804 DF, p-value: < 2.2e-16

Step 1: Definition of the terms

1 is the population slope

Step2:H0:1 =0 Ha:1 ≠0

Step 3: Find the Test Statistic, p-value, report DF tts = 284.429 DF = 6804 P-value < 2.2e-16

Step 4: Conclusion:

= 0.01 Since P-value < 2.2e-16 ≤ 0.01, we should reject H0 The data provides sufficiently strong evidence (P-value = 0.00156) to the claim that there is an association between job stress and LOC.

13.

They are both indicating that the actual elapsed time and distance are in linear relationship. The different part is that they are using different method to get this solution.

14.

The three other assumptions are met; linear, constant standard deviation of the residuals and normality of the residuals, therefore linear regression analysis appears to be reasonable. So the relationship between distance and actual elapsed time is linear. This situation is good for prediction. Because we can see that if the distance is longer, the time will be longer. We cannot generalize this situation for Nov 2015 because Nov is a rush time which is in a different situation with the whole year. The distances over 2704 miles will have more different situation what we cannot predict so that we need not to include them.