STATS 201/8 Assignment 1

Chunkai Meng / 857058772

Due Date: 3pm, Thursday 4th August 2016

## Loading required package: s20x

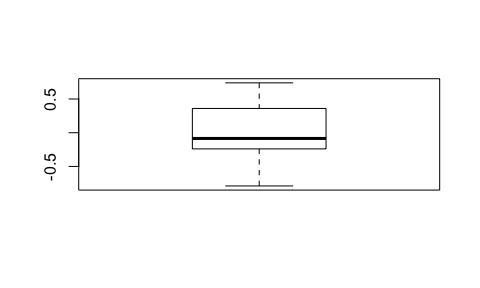
# Question 1

## Question of interest/goal of the study

We are interested in seeing if a fuel treatment improves the fuel consumption of vehicles.

## Read in and inspect the data:

Fuel.df=read.table("FuelConsumption.txt", header=T)  
Fuel.diff=Fuel.df$modified-Fuel.df$standard  
boxplot(Fuel.diff)



**This boxplot looks reasonably symmetric. We can do the fit**

## Using the t.test function

t.test(Fuel.diff,var.equal=TRUE)

##   
## One Sample t-test  
##   
## data: Fuel.diff  
## t = 0.36177, df = 19, p-value = 0.7215  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.1674911 0.2374911  
## sample estimates:  
## mean of x   
## 0.035

# Why 2-sample t-test is not appropriate

Because the independence assumption is not valid. The vehicles having high standard fuel consuption could have high modified fuel consumption and vice-versa. So these two measurements are not independent of each other. We should use their differences as single measurement in one-sample test.

# Manual calculation in R

Formulae: , and 95% confidence interval

NOTES: The R code mean(y) calculates , sd(y) calculates the standard deviation of , and qt(0.975,19) gives the multiplier. The standard error, is calculated by the standard deviation of divided by the square root of of the sample size, .

# sample mean  
mn\_diff=mean(Fuel.diff)  
mn\_diff

## [1] 0.035

# sample sd  
sd\_diff=sd(Fuel.diff)  
sd\_diff

## [1] 0.4326601

# sample size  
n\_diff=length(Fuel.diff)  
n\_diff

## [1] 20

# t-multiplier  
tmult\_diff=qt(1-.05/2, df=n\_diff-1)  
tmult\_diff

## [1] 2.093024

# 95% CI  
CI\_diff=mn\_diff+tmult\_diff\*c(-1,1)\*sd\_diff/sqrt(n\_diff)  
CI\_diff

## [1] -0.1674911 0.2374911

# standard error  
se\_diff=sd\_diff/sqrt(n\_diff)  
se\_diff

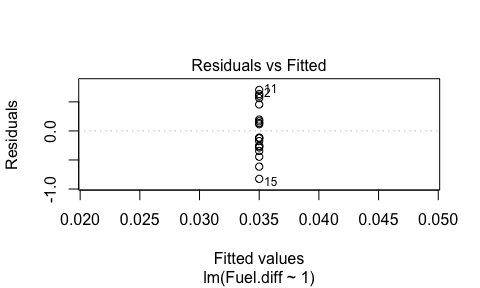
## [1] 0.09674573

#t\_stat\_diff=(mn\_diff)  
t\_stat\_diff=(mn\_diff-0)/se\_diff  
t\_stat\_diff

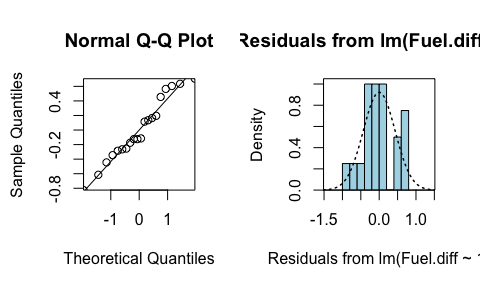
## [1] 0.3617731

# Using the lm function

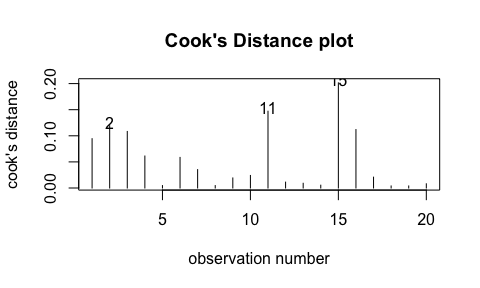
Fuel.lm=lm(Fuel.diff~1)  
plot(Fuel.lm,which=1)



normcheck(Fuel.lm)



cooks20x(Fuel.lm)



summary(Fuel.lm)

##   
## Call:  
## lm(formula = Fuel.diff ~ 1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.825 -0.270 -0.120 0.260 0.705   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.03500 0.09675 0.362 0.722  
##   
## Residual standard error: 0.4327 on 19 degrees of freedom

confint(Fuel.lm)

## 2.5 % 97.5 %  
## (Intercept) -0.1674911 0.2374911

## Note:

You should get exactly the same results from the manual calculations and using the linear model function as the t.test function. Doing this was to giving you practice using some R code.

# Method and Assumption Checks

As the data has two measurements on each vehicle, we have applied a paired sample t-test (i.e., a one-sample t-test applied to the differences within each vehicle.) There is no reason to suspect lack of independence between vehicles and no problem with residuals or any overly influential observations.

The estimated coefficient for the true fuel difference was not significantly different from zero (p-value 0.72). So, our preferred model is: where

# Executive Summary

**We cannot reject the Null hypothesis that fuel treatment did not improves the fuel consumption of vehicles (P-value 0.72).**

**We estimate that consumption of fuel would increase or decrease by less than 0.24 liters and 0.17 liters respectively after the treatment.**

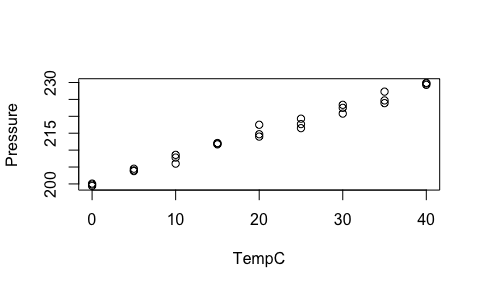
# Question 2

# Question of interest/goal of the study

**We want to quantify the relashionship between tank pressure changed and increasing temperature, and estmate the expected tank pressure at 50 degree C.**

# Read in and inspect the data:

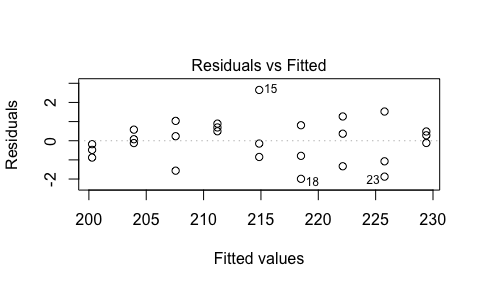
Air.df=read.table("Pressure.txt", header=T)  
plot(Pressure~TempC,data=Air.df)



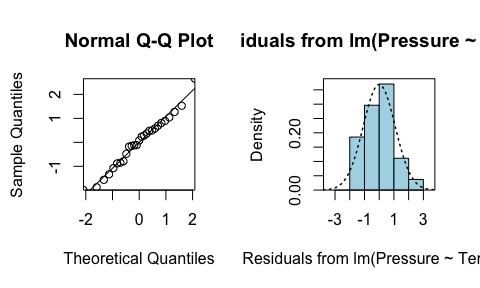
**The underlying trend here looks like a straight line. The variability looks constant.**

# Fit model and do checks

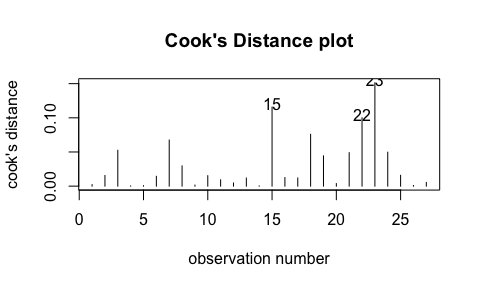
# working model  
presstemp.fit=lm(Pressure~TempC, data=Air.df)  
  
# assumption checks  
eovcheck(presstemp.fit)



normcheck(presstemp.fit)



cooks20x(presstemp.fit)



summary(presstemp.fit)

##   
## Call:  
## lm(formula = Pressure ~ TempC, data = Air.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.99037 -0.81926 0.07852 0.63630 2.65185   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 200.27926 0.39097 512.26 <2e-16 \*\*\*  
## TempC 0.72844 0.01642 44.35 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.102 on 25 degrees of freedom  
## Multiple R-squared: 0.9875, Adjusted R-squared: 0.9869   
## F-statistic: 1967 on 1 and 25 DF, p-value: < 2.2e-16

confint(presstemp.fit)

## 2.5 % 97.5 %  
## (Intercept) 199.4740441 201.0844744  
## TempC 0.6946186 0.7622702

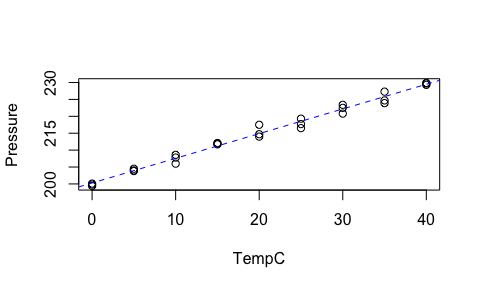
# Estimate the expected pressure at 50 degrees C

predPress.df=data.frame(TempC=50)  
predict(presstemp.fit, predPress.df, interval="confidence")

## fit lwr upr  
## 1 236.7015 235.5967 237.8062

# Create plot with superimposed lines

plot(Pressure~TempC, data=Air.df)  
abline(200.28,0.73, lty=2, col="blue")



# Method and Assumption Checks

**The scatter plot of pressure vs temperature showed a linear relationship with approximately constant scatter and so a linear model was fitted. This working model is where .**

**All model assumptions seem to be satisfied - a slight trend of increase in the residual plot was observed, but does not appear to be of major concern.**

# Executive Summary

**The relationship between Pressure and Temperature was modelled as simple linear and our model explains about 99% of the variation in Pressure.**

**We estimate that every 1 degree C increase in temperature would increase the pressure of the tank by 0.69 to 0.76 bar on average.**

**For temperature 50 degree C, we predict the pressure between 236.6 bar to 237.8 bar.**

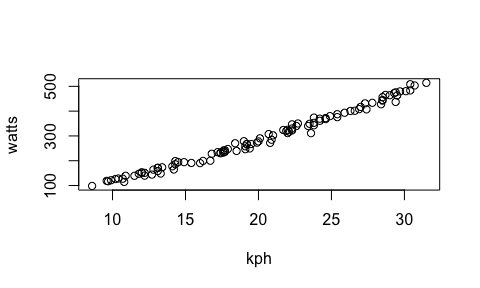
# Question 3

# Question of interest/goal of the study

**We wish to quantify the relationship between power consumption and speed, especially for purpose of being able to predict power consumption from a given speed. We may also interest in the relationship among maximun distance, speed and power consumption according to a given 400 watt-hour battery.(Bonus question)**

# Read in and inspect the data:

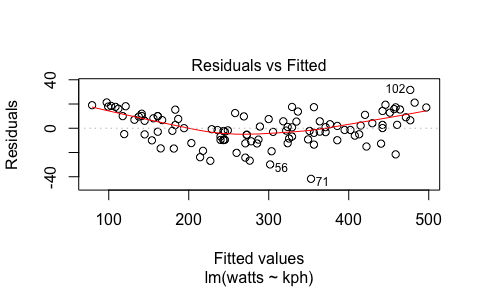
Ebike.df=read.table("CyclePower.txt", header=T)  
plot(watts~kph,data=Ebike.df)



**It looks pretty much like a straight line, let's fit one.**

# Model fitting and checks.

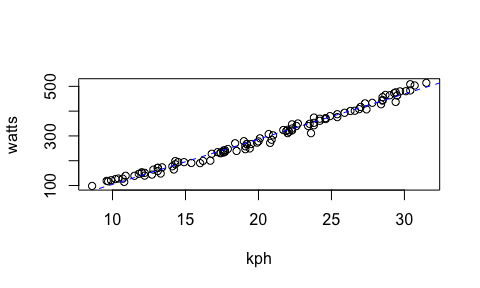
# Working model  
wattskph.fit=lm(watts~kph, data=Ebike.df)  
  
# Assumption checks  
plot(wattskph.fit, which=1)



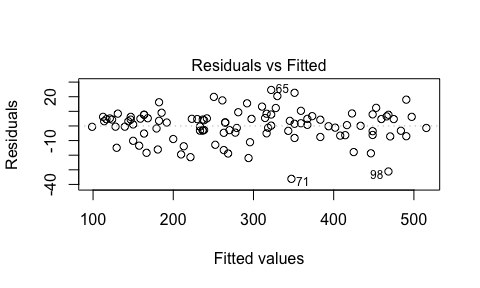
# Not good, the residuals and fitted values have significant 'curvy' relationship. let's doubule confirm by ploting with a superimposed lines.  
summary(wattskph.fit)$coef

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -77.71811 4.3908028 -17.70020 5.208055e-33  
## kph 18.24417 0.2079432 87.73633 1.299094e-98

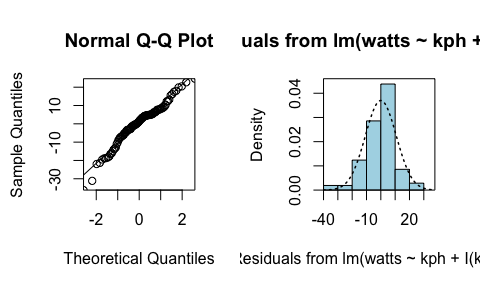
plot(watts~kph,data=Ebike.df)  
abline(-77.72,18.24, lty=2, col="Blue")



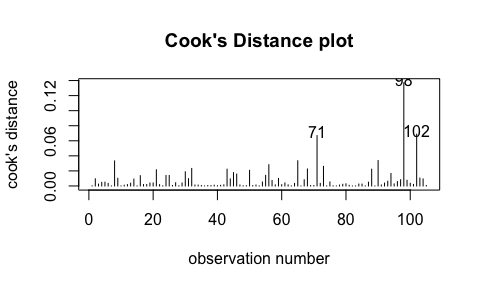
# Yep, it looks a little like a curve.  
  
# Add a squared term for kph  
wattskph2.fit=lm(watts~kph + I(kph^2), data=Ebike.df)  
eovcheck(wattskph2.fit)



normcheck(wattskph2.fit)



cooks20x(wattskph2.fit)



# Assumptions all look satisfied.  
  
summary(wattskph2.fit)

##   
## Call:  
## lm(formula = watts ~ kph + I(kph^2), data = Ebike.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.182 -5.054 1.478 6.219 24.628   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.80175 11.23558 -0.160 0.873   
## kph 9.92299 1.17888 8.417 2.52e-13 \*\*\*  
## I(kph^2) 0.20630 0.02892 7.134 1.45e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.89 on 102 degrees of freedom  
## Multiple R-squared: 0.9912, Adjusted R-squared: 0.991   
## F-statistic: 5739 on 2 and 102 DF, p-value: < 2.2e-16

confint(wattskph2.fit)

## 2.5 % 97.5 %  
## (Intercept) -24.087469 20.4839735  
## kph 7.584689 12.2612886  
## I(kph^2) 0.148940 0.2636619

# Prediction of power consumption at 20 kph

predwatts.df=data.frame(kph=20)  
predict(wattskph2.fit, predwatts.df, interval="prediction")

## fit lwr upr  
## 1 279.1784 257.3702 300.9866

# Method and Assumption Checks

\*\* Scatter plot didn't show a significant 'curvy', while residuals plot from the fit of simple linear model suggested slight curvature in the relationship. So a quadratic term was added. This model coorespond to all assumptions. The final model is where

# Executive Summary

The relationship between power consumption and speed was modelled as quadratic and our model explains about 99% of variation in power consumption. At the beginning for a one km/s increase in speed, the increase in expected power consumption was smaller as speed increased. However, after the speed growed up to a specific value, the increase in expected power consumption became greater as the speed increased. For speed of 20km/s the estimated expectd power consumption of their model of bike are between 257.4w to 300.1w.

# Bonus question:

# The relationship of distance, speed and power consumption can be explain as follow equation: S = V \* T; T = 400 / W; So S = 400 \* V / W  
v\_w = Ebike.df$kph / Ebike.df$watts  
summary(v\_w)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.05978 0.06613 0.07191 0.07226 0.07748 0.09432

# The Max of v\_w was about 0.094  
# The Max of S = 400 \* 0.094 :  
S\_max = 400 \* 0.094  
S\_max

## [1] 37.6

# Executive Summary

**This set of equations based on the assumption that the watts output of the battery would not decline as the power consumed by the bike. This was not going to happen. Since we do not have the measurement about the output changes, the maximum distance 37.6 km still could be a prusue goal.**