西南大学 2022: STATS 201 Assignment 1a

Your Name and ID Number here

Wed 27th April before class at 15:50

# Background

Students in a class were asked to conduct an experiment to determine their average stride length (步幅). That is, the distance traveled with every step.

Each student was instructed to find a flat location where they could walk for 5 to 15 minutes at their natural gait, unimpeded by traffic or other pedestrians. They used a smart device to record the number of steps, and the distance walked (in metres). They were asked to repeat this task about 30 times.

The code below automatically generates the data for a randomly chosen student. The data are in the dataframe **Stride.df**. Variable **steps** is the explanatory variable, and **distance** is the response variable.

# Question of interest

Make inference about the average stride length of the student using a simple linear model.

You need to conduct the analysis using R, complete the Methods and Assumptions Checks, and write the Executive Summary.

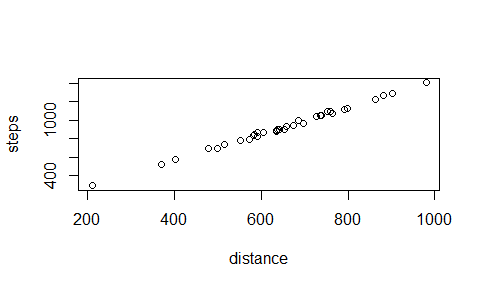
## Warning: package 's20x' was built under R version 4.0.5

# Enter your name here

# Replace "Enter your name here" with your name in quotes,   
# E.g., myname="Ruoxi Xu"  
myname="runze liao"

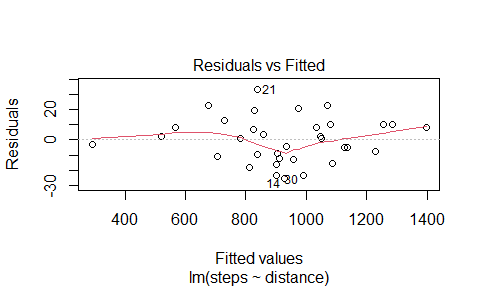
# Scatter plot of steps vs distance

# Add R code below to draw the scatter plot  
plot(steps~distance, data = Stride.df)

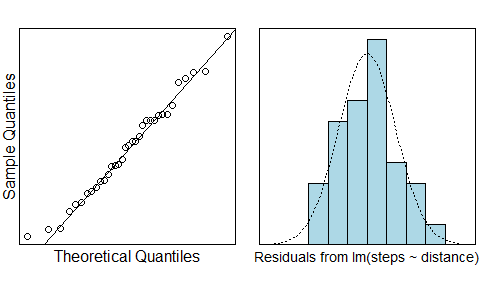


# Fit a simple linear model and do assumption checks

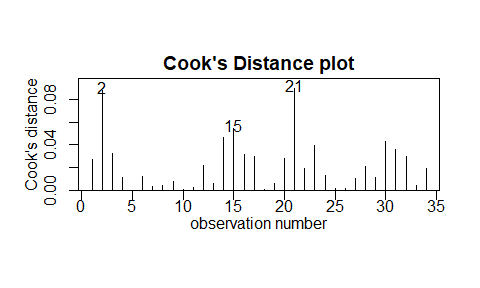
# Add R code below  
# NOTE: If any assumptions look questionable, mention this in the Methods and Assumption Checks,  
# but do not do any alterations to the model or data   
Stride.fit = lm(steps~distance,data = Stride.df)  
plot(Stride.fit,which=1)#test whether same distribution 有问题哦，数据可能不满足同一分布的假设



normcheck(Stride.fit)#normal check looks great, satisfy the normal Distribution



cooks20x(Stride.fit)#observe the Cook's Distance plot, find the distance > 0.4

 # Inference, i.e, check for significance and calculate confidence intervals

# Add R code below  
summary(Stride.fit) #R^2 is 0.9958.

##   
## Call:  
## lm(formula = steps ~ distance, data = Stride.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -25.716 -10.879 0.793 9.553 32.917   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12.14246 11.02571 -1.101 0.279   
## distance 1.43862 0.01646 87.400 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.84 on 32 degrees of freedom  
## Multiple R-squared: 0.9958, Adjusted R-squared: 0.9957   
## F-statistic: 7639 on 1 and 32 DF, p-value: < 2.2e-16

pred.df = data.frame(Test = c(0,10,20))  
predict(Stride.fit,pred.df)

## Warning: 'newdata' had 3 rows but variables found have 34 rows

## 1 2 3 4 5 6 7 8   
## 1085.5254 675.5184 829.4509 911.4523 783.4150 956.0495 1135.8772 825.1350   
## 9 10 11 12 13 14 15 16   
## 836.6440 1050.9985 518.7087 727.3088 1033.7351 928.7157 1071.1392 973.3130   
## 17 18 19 20 21 22 23 24   
## 1284.0552 1046.6827 904.2592 1256.7214 838.0826 901.3819 901.3819 1229.3876   
## 25 26 27 28 29 30 31 32   
## 934.4702 858.2233 289.9680 567.6218 1079.7710 990.5764 1397.7062 812.1874   
## 33 34   
## 1127.2455 704.2909

predict(Stride.fit,pred.df,interval = "confidence")

## Warning: 'newdata' had 3 rows but variables found have 34 rows

## fit lwr upr  
## 1 1085.5254 1079.1405 1091.9104  
## 2 675.5184 667.7210 683.3159  
## 3 829.4509 823.8056 835.0962  
## 4 911.4523 906.2595 916.6451  
## 5 783.4150 777.2649 789.5651  
## 6 956.0495 950.8184 961.2807  
## 7 1135.8772 1128.7429 1143.0114  
## 8 825.1350 819.4491 830.8209  
## 9 836.6440 831.0631 842.2249  
## 10 1050.9985 1045.0477 1056.9494  
## 11 518.7087 507.9041 529.5133  
## 12 727.3088 720.3666 734.2509  
## 13 1033.7351 1027.9713 1039.4989  
## 14 928.7157 923.5327 933.8988  
## 15 1071.1392 1064.9441 1077.3343  
## 16 973.3130 968.0121 978.6139  
## 17 1284.0552 1274.2222 1293.8881  
## 18 1046.6827 1040.7806 1052.5848  
## 19 904.2592 899.0531 909.4653  
## 20 1256.7214 1247.4237 1266.0190  
## 21 838.0826 832.5140 843.6512  
## 22 901.3819 896.1691 906.5948  
## 23 901.3819 896.1691 906.5948  
## 24 1229.3876 1220.6116 1238.1635  
## 25 934.4702 929.2835 939.6569  
## 26 858.2233 852.8088 863.6378  
## 27 289.9680 274.2759 305.6600  
## 28 567.6218 557.8023 577.4414  
## 29 1079.7710 1073.4634 1086.0785  
## 30 990.5764 985.1766 995.9763  
## 31 1397.7062 1385.5419 1409.8706  
## 32 812.1874 806.3709 818.0040  
## 33 1127.2455 1120.2479 1134.2430  
## 34 704.2909 696.9808 711.6009

# Method and Assumption Checks

**The relationship between number of steps and distance walked looks very linear, so a simple linear regression model was fitted. However, when we do the Residual Analysis, the Residual plot seems not the same, which indicates that maybe the residual does not follow the same distribution.**

The fitted model is

where ~ .

**Our model explained 99.58% of the variability in the students’ final exam marks**

# Executive Summary

**We are interested in building a model to estimate the steps with distance** **The relation between the steps and distance is quite linear. when we do the Residual Analysis, the Residual plot seems not the same, which indicates that maybe the residual does not follow the same distribution.** **the average stride length is 1.43862, For instance, if we walk 1 steps, the distance is about 1.43862 meters**