西南大学 2022: STATS 201 Assignment 1a

runze liao

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.

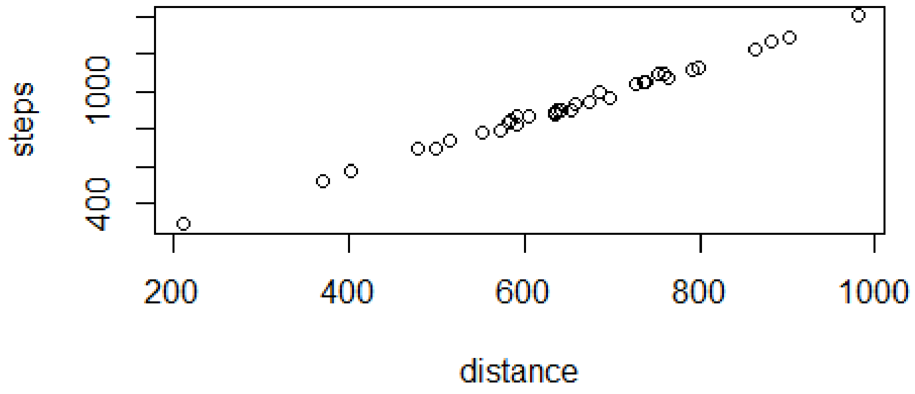
```
## Warning: package 's20x' was built under R version 4.0.5
```

Enter your name here

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

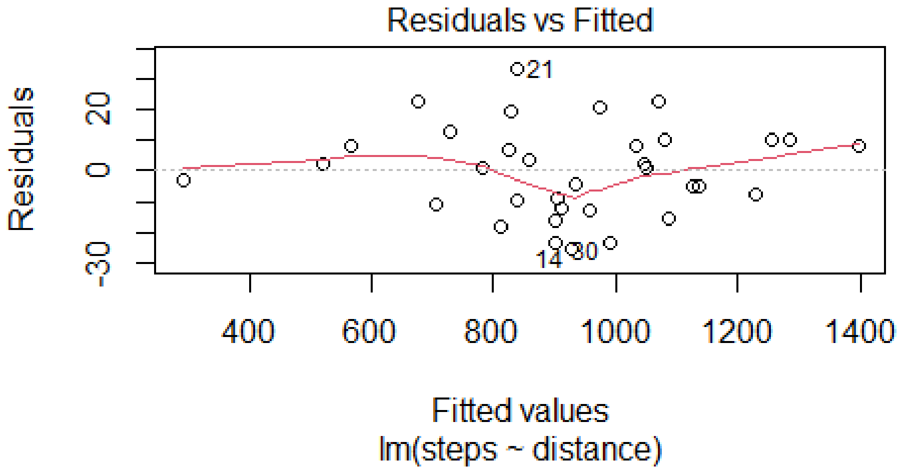
Scatter plot of steps vs distance

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

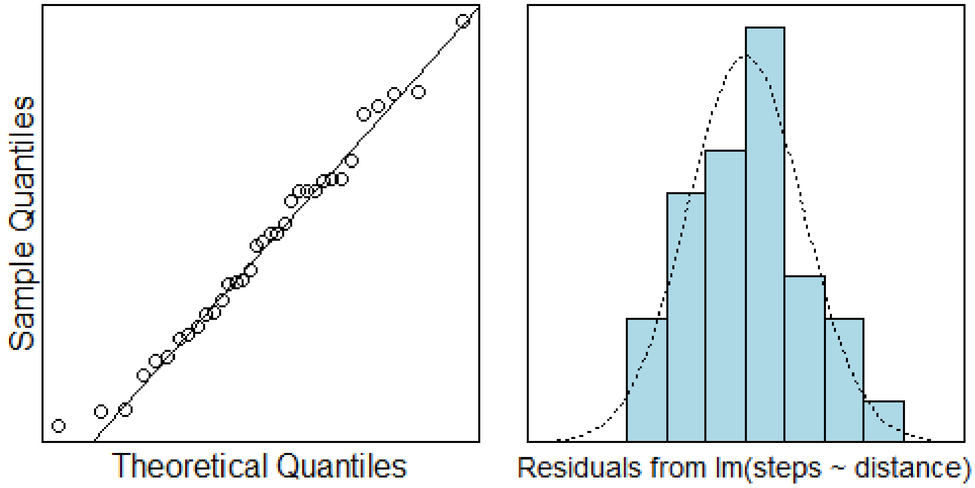


Fit a simple linear model and do assumption checks

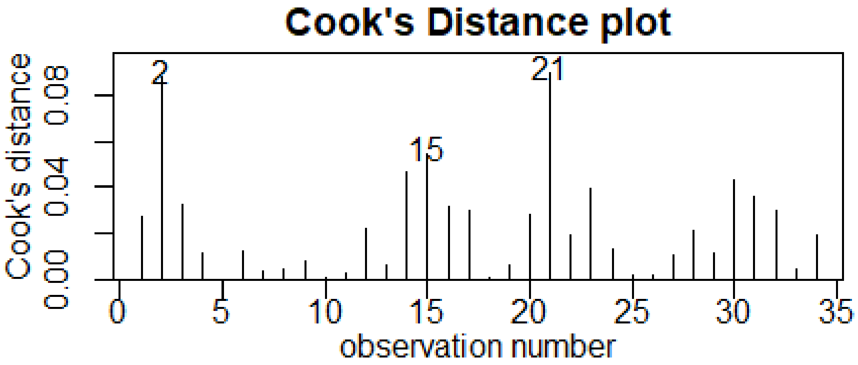
```
# Add R code below
# NOTE: If any assumptions look questionable, mention this in the Metho
ds and Assumption Checks,
# but do not do any alterations to the model or data
Stride.fit = lm(steps~distance,data = Stride.df)
plot(Stride.fit,which=1)#test whether same distribution 有问题哦，数据可
能不满足同一分布的假设
```



```
normcheck(Stride.fit)#normal check looks great, satisfy the normal Dist
ribution
```



```
cooks20x(Stride.fit)#observe the Cook's Distance plot, find the distanc
e > 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
```

```
confint(Stride.fit)
```

```
##                  2.5 %    97.5 %
## (Intercept) -34.601084 10.316171
## distance      1.405093  1.472149
```

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

$$distance_i = \beta_0 + \beta_1 \times steps_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

Our model explained 99.58% of the variability in the students' final exam marks. The confidence interval of stride length is 1.405093 to 1.472149, the intercept's confidence interval is -34.601084 to 10.316171.

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