Q1

Joohyeok

2023-10-04

```
library(ggplot2)

df = read.table("IQ.txt",header=TRUE)

lm_model <- lm(IQ ~ birthweight, df)

result <- summary(lm_model)
result
```

```
##
## Call:
## lm(formula = IQ ~ birthweight, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.705  -8.993  -0.635   8.046  33.411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.015672   3.988953  21.814   <2e-16 ***
## birthweight  0.008974   0.003927   2.286   0.0238 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.59 on 136 degrees of freedom
## Multiple R-squared:  0.03699,    Adjusted R-squared:  0.02991
## F-statistic: 5.224 on 1 and 136 DF,  p-value: 0.02383
```

a. i. Let beta1 be the effect of birthweight in the model. Then H0: beta1=0 vs Ha: beta1>0.

ii. If H0 is true, then t is a random draw from a t136 distribution

iii. t= 2.286 (from the summary output)

```
1-pt(2.286,136)
```

```
## [1] 0.01190008
```

iv. p-value = 0.0119

v. Since p-value>0.01, we do not reject the null hypothesis. We have no evidence at the 1% level that children with higher birthweight (within the population of children born with very low birthweights) have higher IQs.

b. y = 87.0157 + 0.008974x

c. It means that the IQ at the age of 5 is lower than other children of the same birthweight. In short, the IQ at 5 of child #127 is smaller than predicted value.

```
predict(lm_model,newdata=data.frame(birthweight=750),interval="prediction",level=0.99)
```

```
##        fit      lwr      upr
## 1 93.74631 60.64212 126.8505
```

d. a 99% point prediction of the IQ of a baby with a birthweight of 750g is 93.75

e. a 99% prediction interval for the IQ of a baby with a birthweight of 750g is (60.64,126.85).

f. The CI is narrower because we have more information about the mean IQ at 5 of babies with birthweights of 750g (a parameter) than about the IQ at 5 of a baby with birthweight of 750g (a random variable). The latter is inherently more variable than an average of such random variables..

g. The width of the prediction interval is mainly determined by standard error of the least squares estimator of beta1. Referring to Q4(c) of assignment1, the variance (and therefore likely the estimated variance of the least squares estimator of beta1) decrease with increasing the sample size and increasing the values of the xi’s in the sample. Therefore, to achieve a narrower prediction interval, he could use a bigger sample and a wider range of birthweight.

```
mean(df$IQ)
```

```
## [1] 95.7971
```

h. The estimated mean IQ of a baby with a birthweight of 750g is 95.8

Q2

Joohyeok

2023-10-10

Q2) The residuals vs. predictor plot should have points based on y=0 but this plot does not. Since this plot is an increasing graph ignoring the y=0 line, this plot could never result from fitting a simple linear regression model using the method of least squares.

Q3

Joohyeok

2023-10-10

a. Longitudinal Data: In this scenario, the data involves collecting information on epilepsy patients over a 1-year period. Patients are recruited from multiple medical centers, and the key variable of interest is the number of seizures they experience over time. Longitudinal data involves repeated measurements on the same subjects (in this case, patients) over a period of time. Researchers are interested in observing changes within individuals (seizure frequency) over the course of the study, which is a characteristic of longitudinal data.

b. Time series Data: In this scenario, the researcher is studying the relationship between the weekly number of malaria patients and the average weekly level of mosquitoes. The data is collected at regular intervals (weekly) over time, making it a time series. Time series data focuses on observing changes in a variable over sequential time periods.

c. Clustered Data: The data involves SFU students participating in an E-reader study, where each student uses one of four types of light filters. The data is clustered because students are grouped based on the type of filter they use. This grouping or clustering can introduce dependence among observations within the same cluster, which needs to be considered in the analysis.

d. Clustered Data: Starbucks monitors daily sales of pumpkin spice lattes at 120 different locations, and each location displays one of three different promotional posters. The data is clustered because sales are grouped by location, and different promotional posters are applied within each cluster. Similar to scenario (c), this clustering introduces potential dependence among sales within the same location.

Q4

Joohyeok

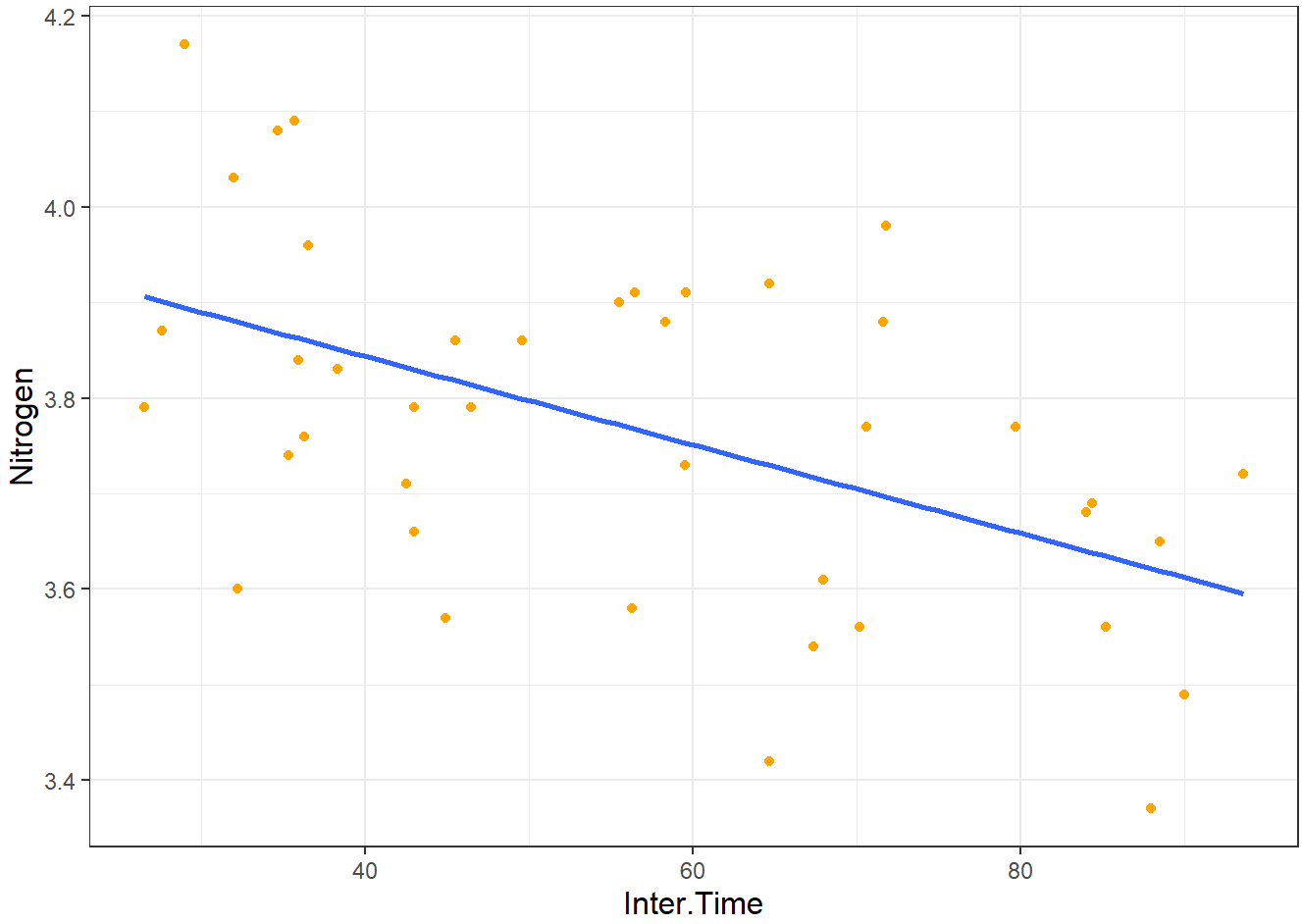
2023-10-09

```
library(ggplot2)

df=read.csv("nitro.csv", header=TRUE)

fit=lm(Nitrogen~Inter.Time,df)
summary(fit)
```

```
##
## Call:
## lm(formula = Nitrogen ~ Inter.Time, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30934 -0.12202 -0.02145  0.12432  0.28350
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.028589   0.074150  54.331  < 2e-16 ***
## Inter.Time  -0.004625   0.001244  -3.717 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1585 on 39 degrees of freedom
## Multiple R-squared:  0.2616, Adjusted R-squared:  0.2427
## F-statistic: 13.82 on 1 and 39 DF,  p-value: 0.0006315
```

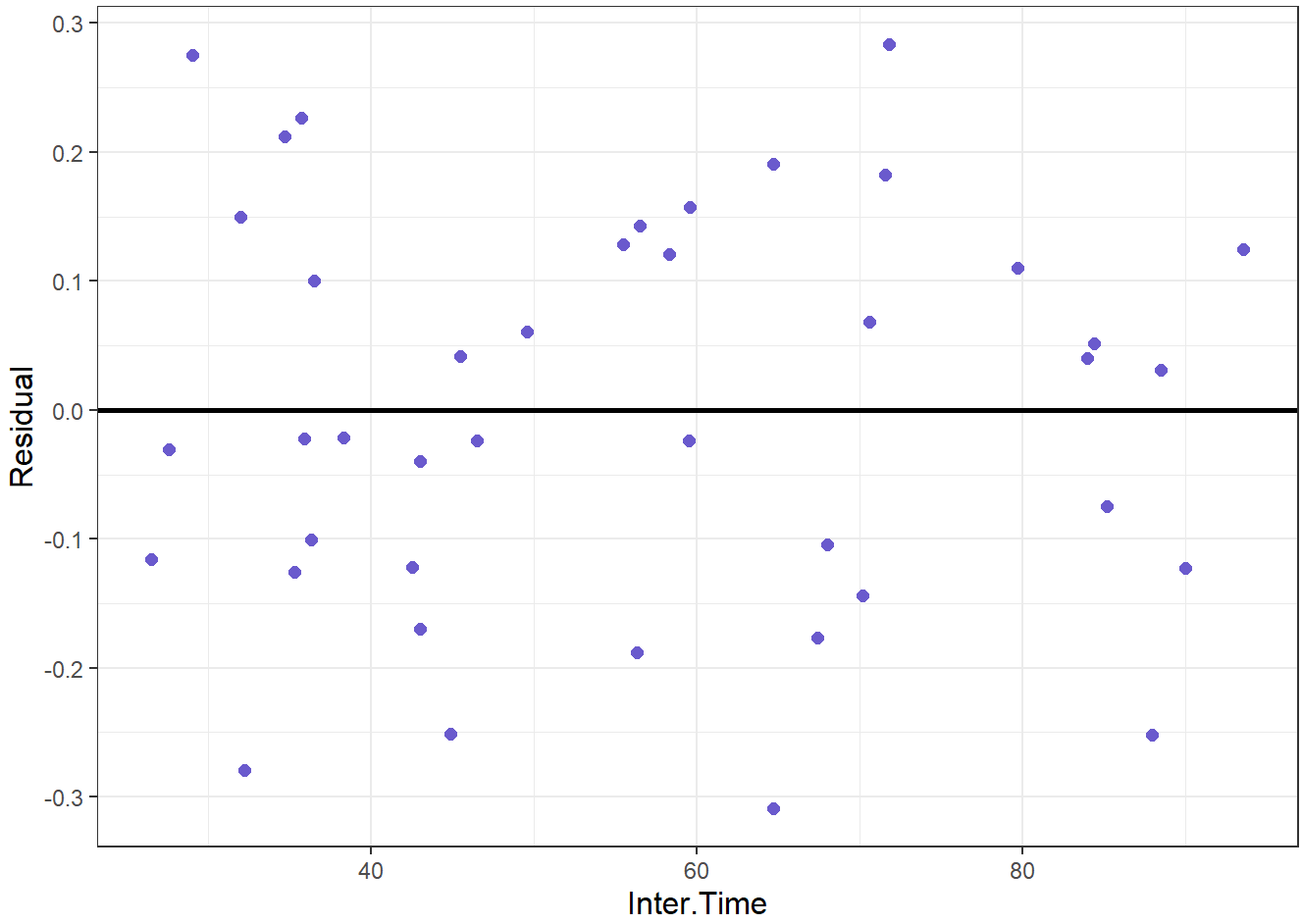
```
ggplot(df,aes(y=Nitrogen,x=Inter.Time))+
  geom_point(colour="orange")+
  geom_smooth(method = 'lm', se = F)+
  theme_bw()+
  theme(axis.title=element_text(size=12))+
  labs(x="Inter.Time", y="Nitrogen")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

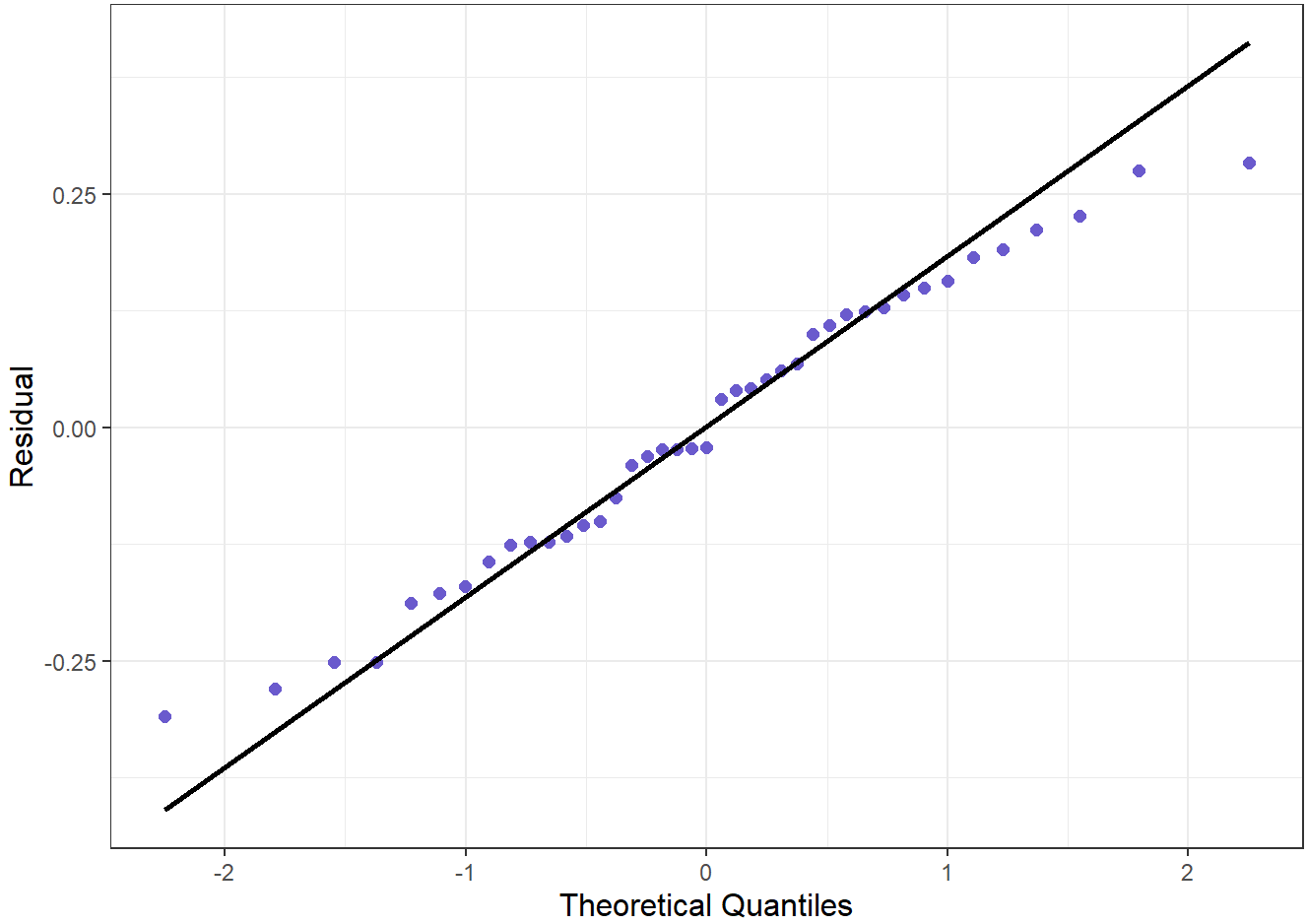



```
df$Residual=resid(fit)

ggplot(df,aes(x=Inter.Time,y=Residual))+
  geom_point(colour="slateblue",size=2)+
  labs(y="Residual",x="Inter.Time")+
  theme_bw()+
  theme(axis.title=element_text(size=12))+
  geom_hline(yintercept=0,size=1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
ggplot(df,aes(sample=Residual))+
  stat_qq(size=2,colour="slateblue")+
  stat_qq_line(size=1)+
  labs(y="Residual",x="Theoretical Quantiles")+
  theme_bw()+
  theme(axis.title=element_text(size=12))
```



i. The residuals vs. fitted values plot shows no suggestion of a trend; mean 0 assumption. Therefore, seems reasonable.

ii. The assumption that the errors have common variance seems reasonable here because from the plot of residual vs inter.time, the vertical spread of the residual is relatively constant.

iii. The Q-Q plot shows no systematic deviations from a straight line, suggesting that the normality assumption is reasonable.