Date/ Page/

Q1

$$Var[\hat{\beta}] = \sigma^2 (X'X)^{-1}$$

$$\tilde{\mu}_{Y|x} = x^T \beta$$

$$Var[Y] = x^T Var(\beta) x'$$

$$= \sigma^2 x^T (X'X)^{-1} x'$$

a) 95% confidence interval for the mean response

$$\tilde{\mu}_{Y|x} \pm t_{1-\alpha/2, n-p-1} SE_{\hat{\mu}}$$

$$= x^T \beta \pm t_{0.025, n-p-1} \sqrt{\sigma^2 x^T (X'X)^{-1} x'}$$

b) 95% prediction interval for the response

$$\tilde{\mu}_{Y|x} \pm t_{1-\alpha/2, n-p} \sqrt{Var(Y) + \sigma^2}$$

$$= x^T \beta \pm t_{0.025, n-p} \sqrt{\sigma^2 x^T (X'X)^{-1} x' + \sigma^2}$$

$$= x^T \beta \pm t_{0.025, n-p} \sqrt{\sigma^2 (x^T (X'X)^{-1} x' + 1)}$$

Date/ Page/

Q2. $Var[\hat{\beta}] = \sigma^2(X'X)^{-1}$.

If the columns of X are orthogonal, then this variance-covariance matrix is diagonal, indicating that $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent.

A3_Q3

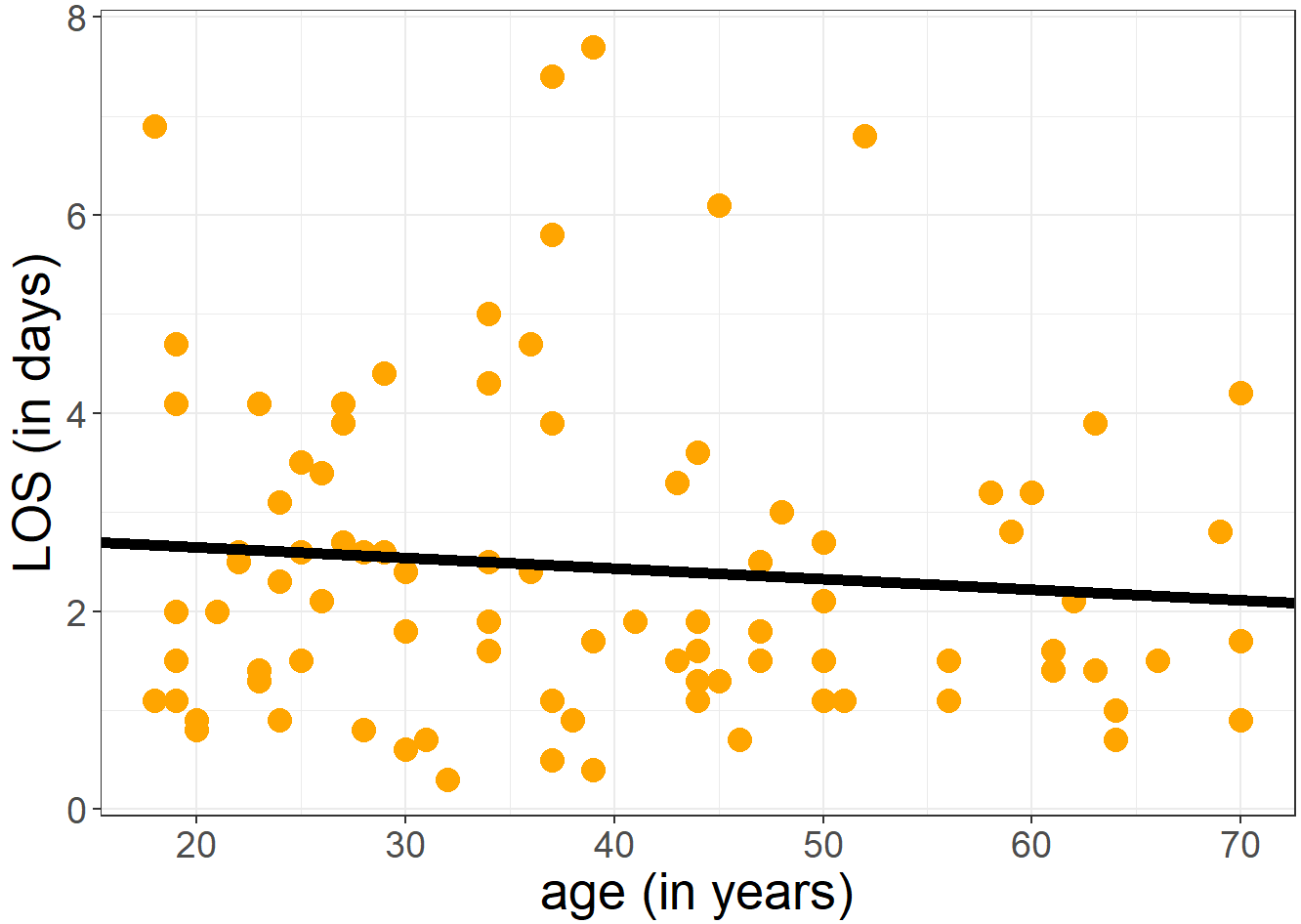
Joohyeok

2023-10-30

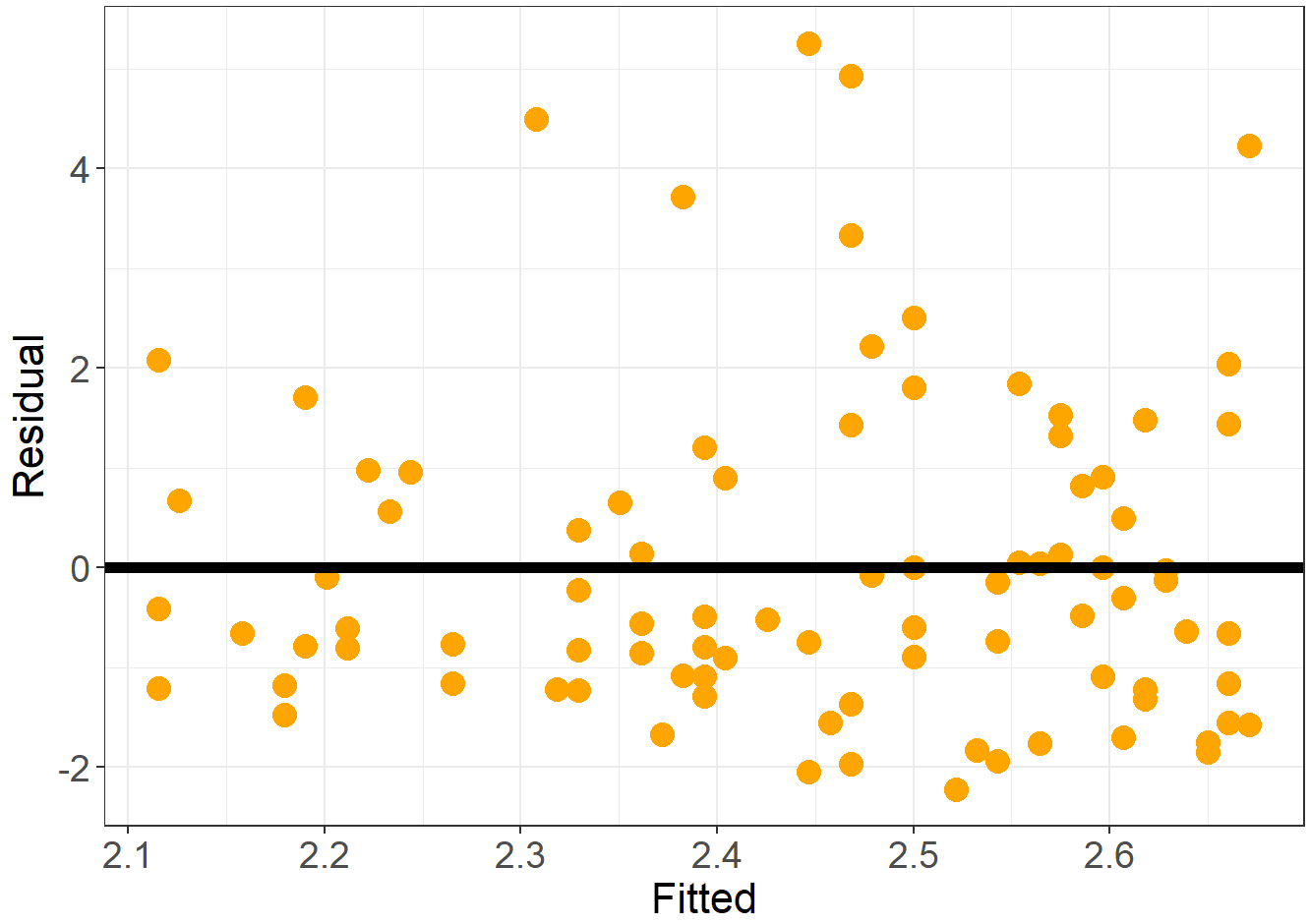
```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

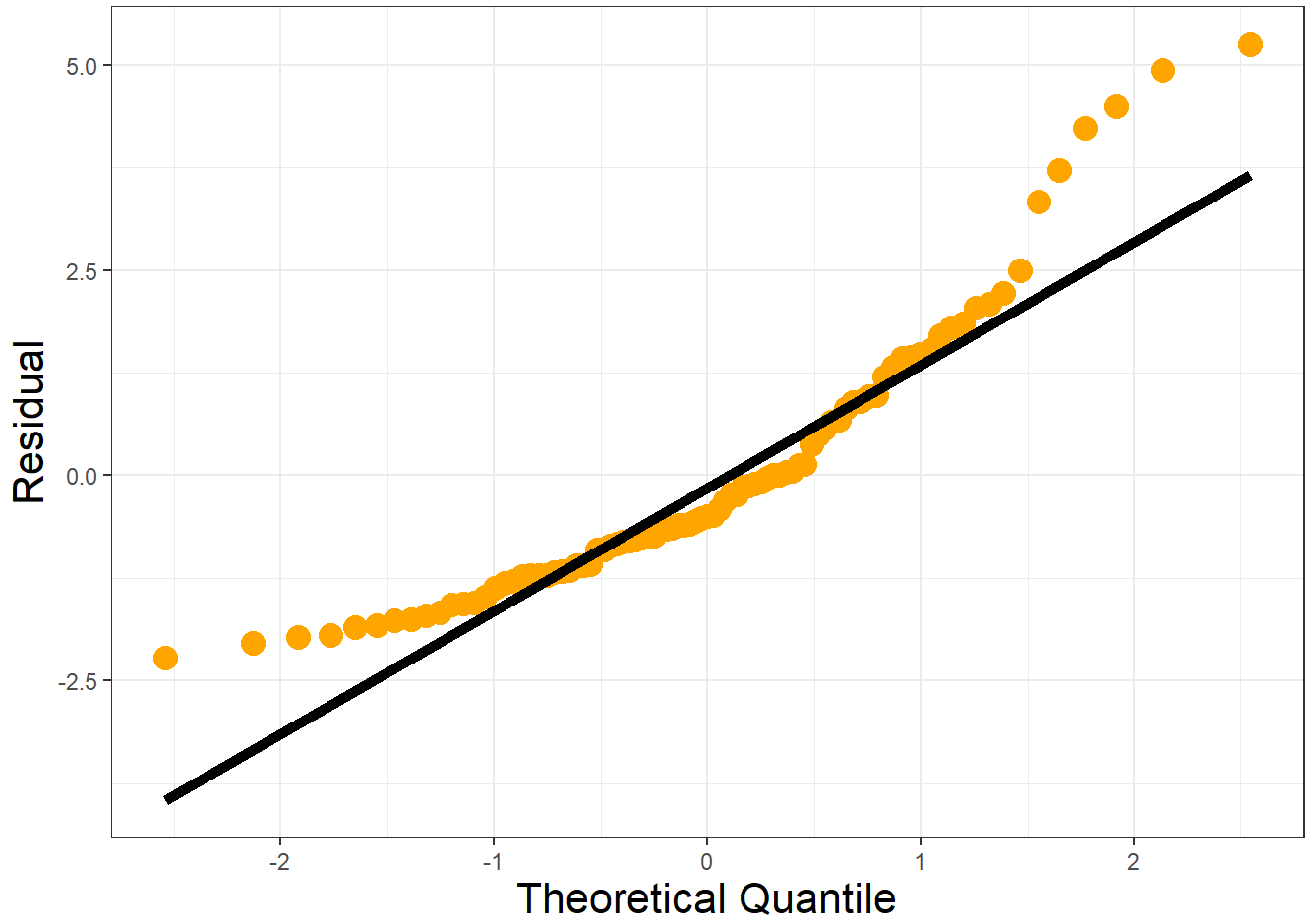
```
df = read.table("LOS.txt",header=TRUE)

fit=lm(LOS~age,df)

df = df %>% mutate(Residual=resid(fit),
                   Fitted=fitted(fit))

ggplot(df,aes(y=LOS,x=age))+
  geom_point(size=4,colour="orange")+
  geom_abline(intercept=coef(fit)[1],slope=coef(fit)[2],
              linewidth=2)+
  theme_bw()+
  theme(axis.title=element_text(size=20),
        axis.text=element_text(size=14))+
  labs(x="age (in years)",y="LOS (in days)")
```




```
ggplot(df,aes(y=Residual,x=Fitted))+
  geom_point(size=4,colour="orange")+
  geom_hline(yintercept=0,linewidth=2)+
  theme_bw()+
  theme(axis.title=element_text(size=16),
        axis.text=element_text(size=14))
```

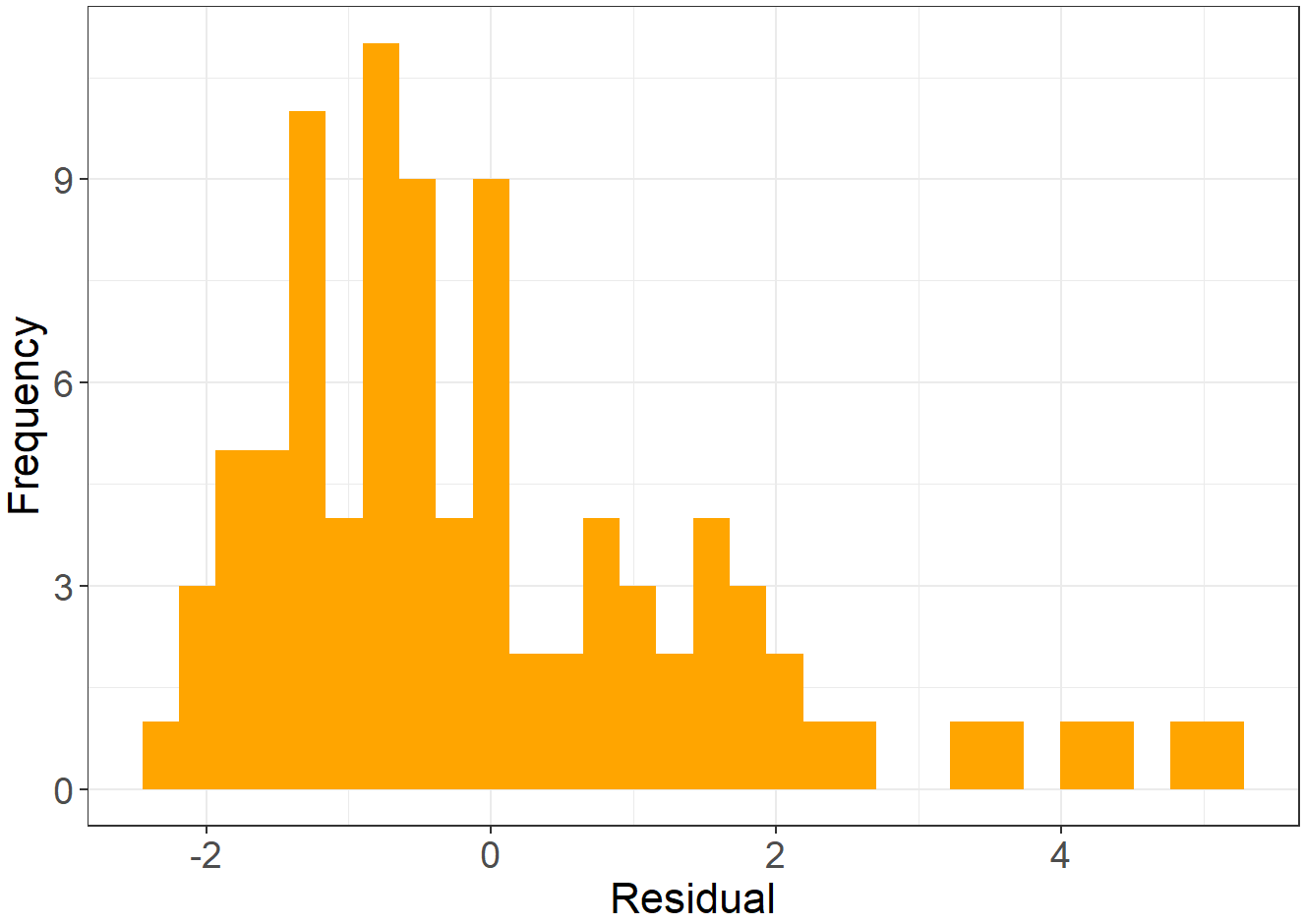



```
ggplot(df,aes(sample=Residual))+
  stat_qq(size=4,colour="orange")+
  stat_qq_line(linewidth=2)+
  labs(y="Residual", x="Theoretical Quantile")+
  theme_bw()+
  theme(axis.title=element_text(size=16))
```



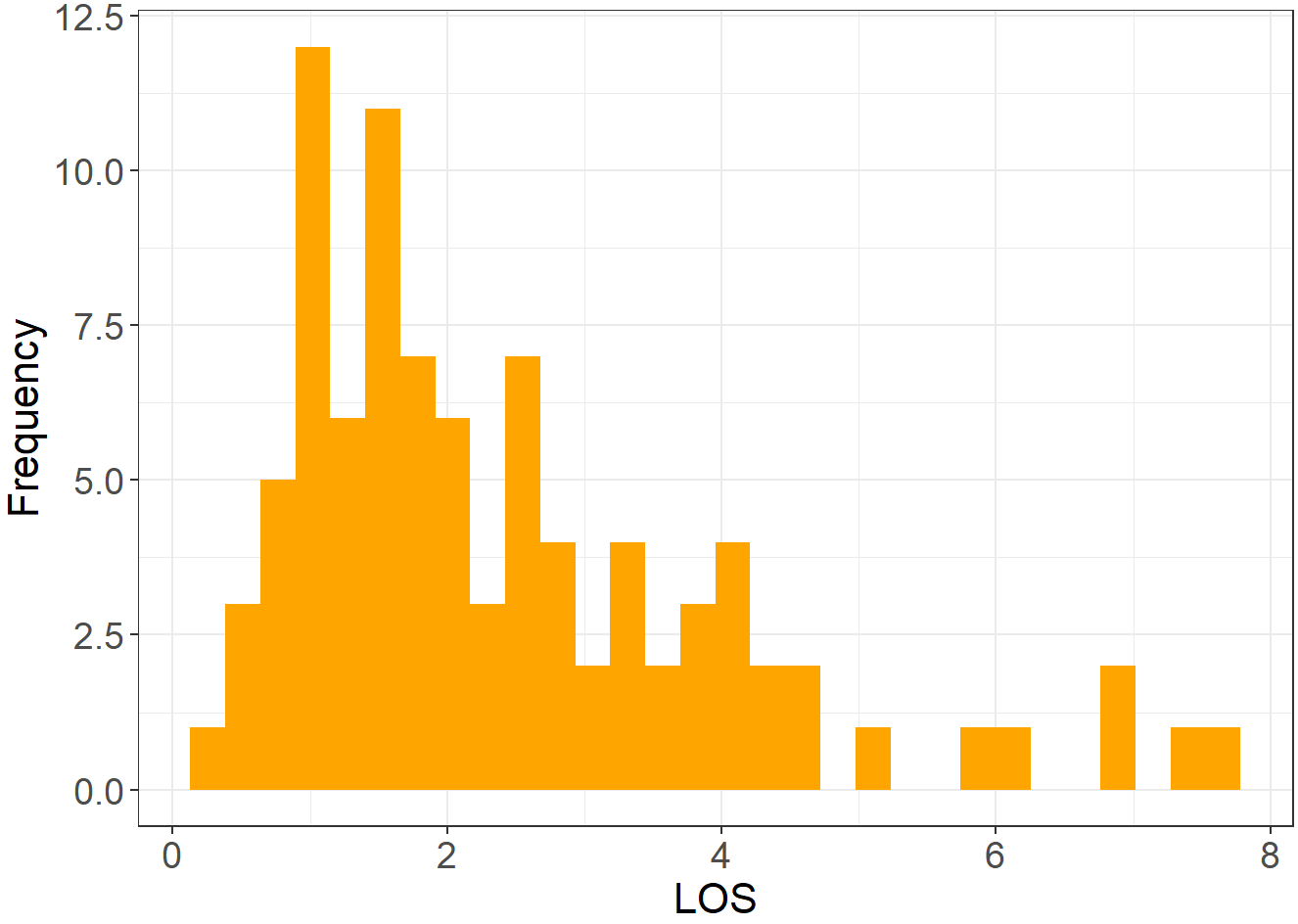

```
ggplot(df,aes(x=Residual))+
  geom_histogram(fill="orange")+
  theme_bw()+
  theme(axis.title=element_text(size=16),
        axis.text=element_text(size=14))+
  labs(y="Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df,aes(x=LOS))+
  geom_histogram(fill="orange")+
  theme_bw()+
  theme(axis.title=element_text(size=16),
        axis.text=element_text(size=14))+
  labs(y="Frequency")
```

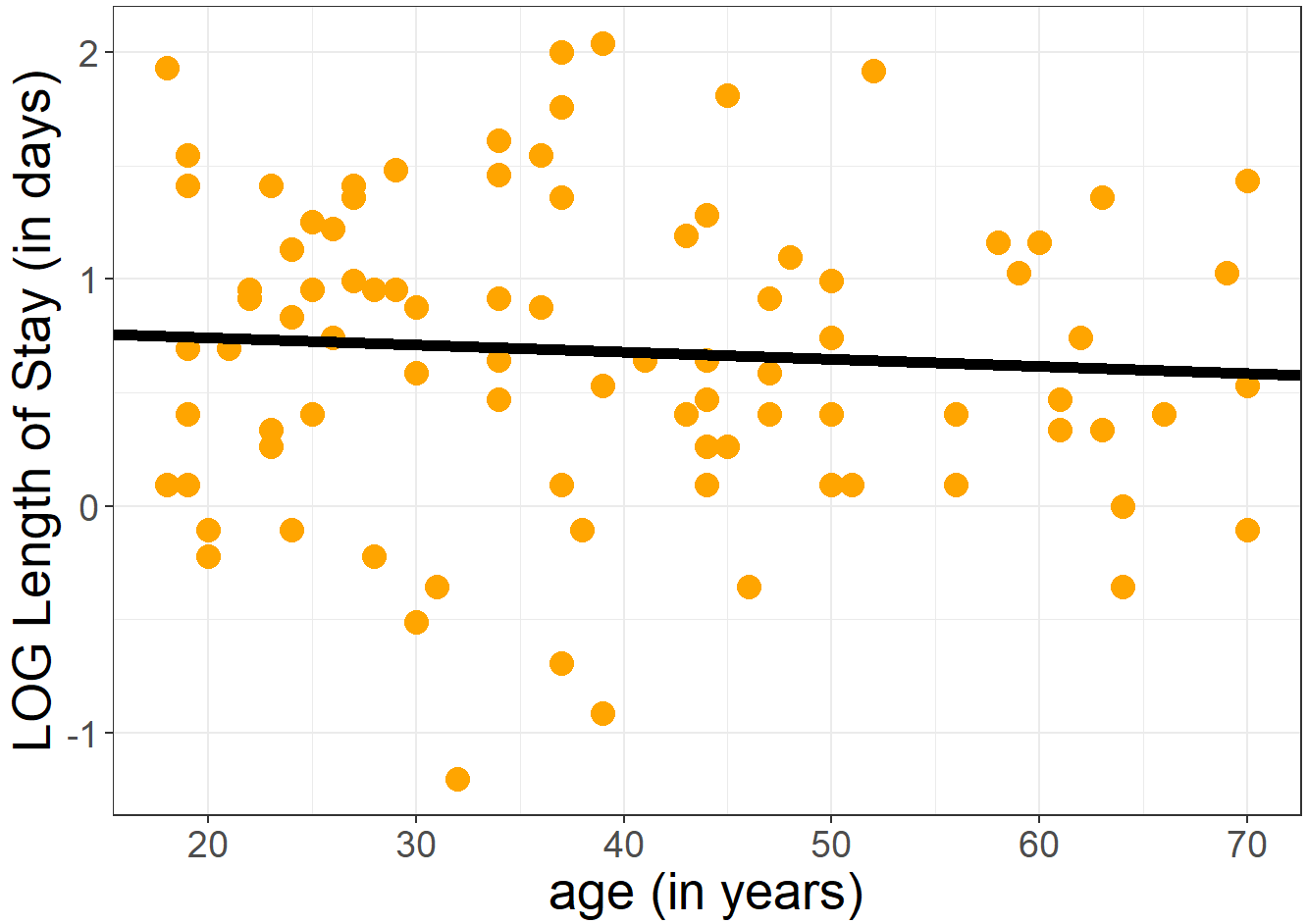
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



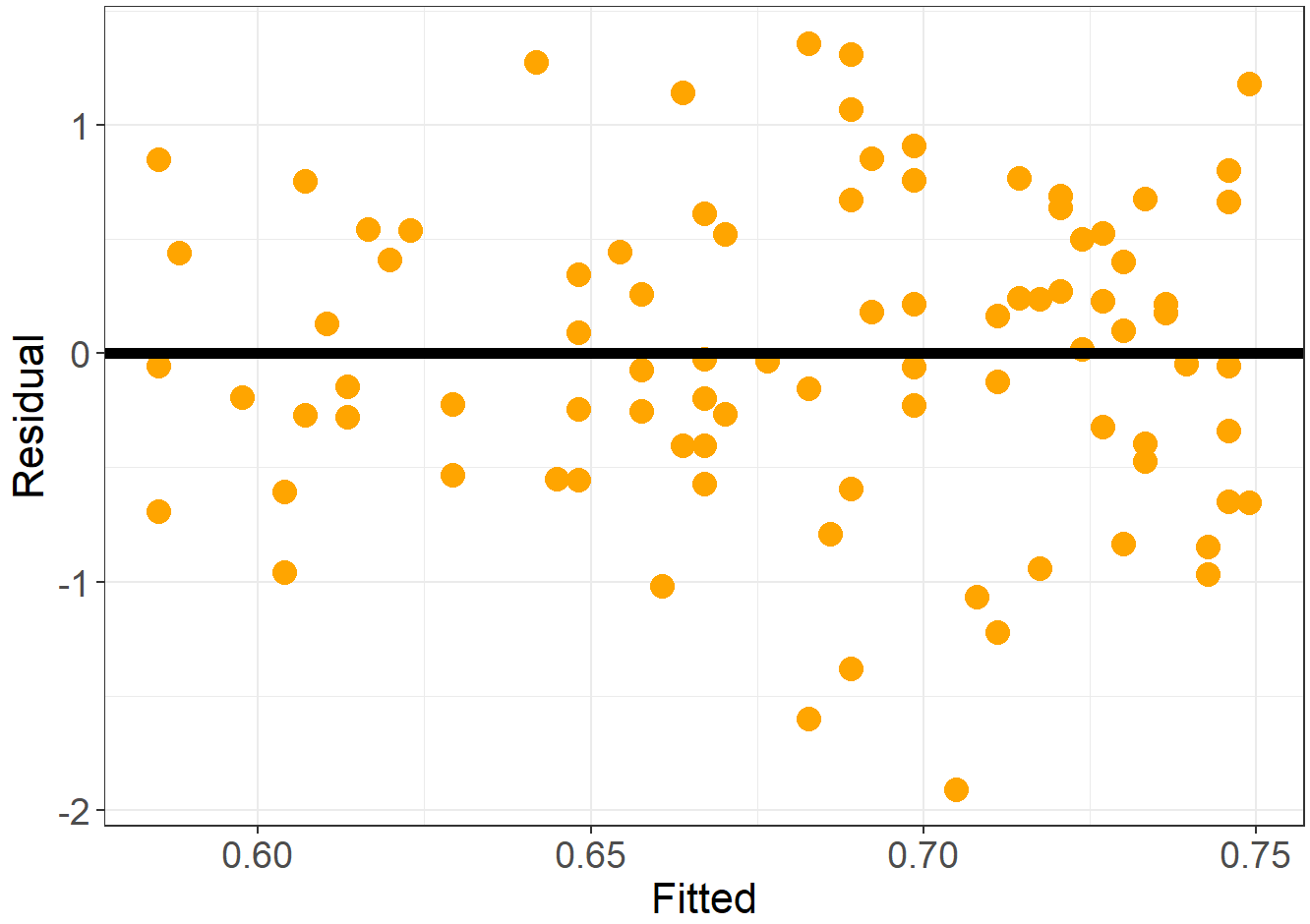
i. Linearity seems reasonable because the residuals vs.fitted values plot does not suggest any trends.
ii. Common SD seems reasonable because for the residuals vs.fitted values plot, vertical spread is approximately constant over the range of the fitted values.
iii. Normality is questionable because the Q-Q plot shows a serious departure from the line.

The residuals are right-skewed. Also, the responses are right-skewed. So, I recommend that the researcher use logarithm of LOS as the response.

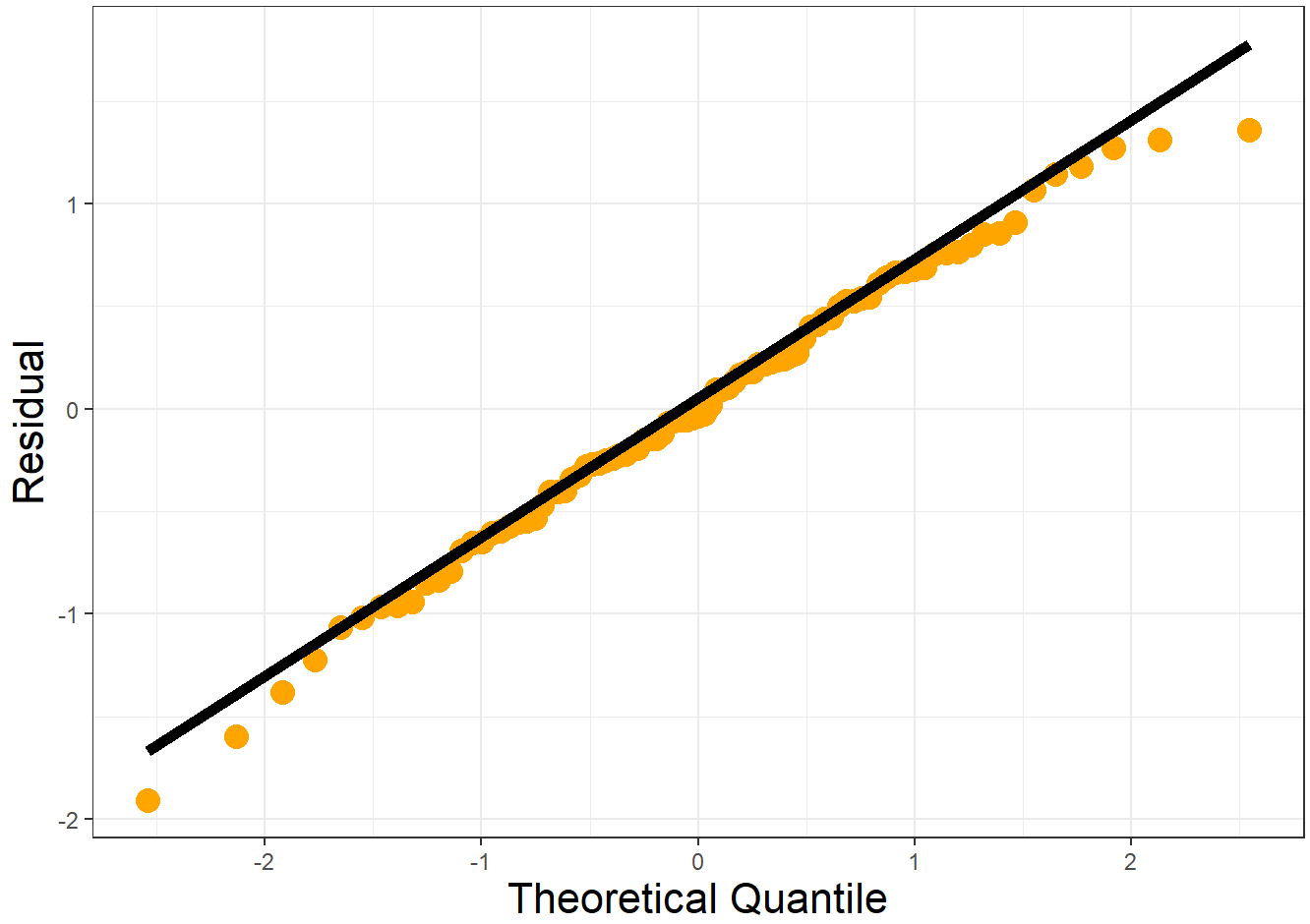
```
df = df %>% mutate(LogLOS=log(LOS))

fit1=lm(LogLOS~age,df)

df = df %>% mutate(Residual.log=resid(fit1),
                   Fitted.log=fitted(fit1))

ggplot(df,aes(y=LogLOS,x=age))+
  geom_point(size=4,colour="orange")+
  geom_abline(intercept=coef(fit1)[1],slope=coef(fit1)[2],
              linewidth=2)+
  theme_bw()+
  theme(axis.title=element_text(size=20),
        axis.text=element_text(size=14))+
  labs(x="age (in years)",y="LOG Length of Stay (in days)")
```




```
ggplot(df,aes(y=Residual.log,x=Fitted.log))+
  geom_point(size=4,colour="orange")+
  geom_hline(yintercept=0,linewidth=2)+
  labs(y="Residual",x="Fitted")+
  theme_bw()+
  theme(axis.title=element_text(size=16),
        axis.text=element_text(size=14))
```




```
#Create the normal quantile plot of the residuals:
ggplot(df,aes(sample=Residual.log))+
  stat_qq(size=4,colour="orange")+
  stat_qq_line(linewidth=2)+
  labs(y="Residual", x="Theoretical Quantile")+
  theme_bw()+
  theme(axis.title=element_text(size=16))
```



i. Linearity seems reasonable because the residuals vs. fitted values plot does not suggest any trends.
ii. Common SD seems reasonable because for the residuals vs. fitted values plot, vertical spread is approximately constant over the range of the fitted values.
iii. Normality is much closer to reasonable because for the Q-Q plot, minor departures from the line are likely not serious.

```
pi=predict(fit1,newdata=data.frame(age=40),interval="prediction",level=0.95)
pi
```

```
##         fit        lwr      upr
## 1 0.6796304 -0.6674053 2.026666
```

```
exp(pi)
```

```
##        fit      lwr      upr
## 1 1.973148 0.513038 7.588744
```

b. Using the model with logarithm of LOS as the response, a 95% prediction interval for the LOS of a 40 year old patient is (0.513, 7.589).

Date/ Page/

Q4 a) Let Y_i, x_{1i}, and x_{2i} be the number of eggs in their carapaces, the lengths (in mm) and weights (in kg) of their carapaces, respectively, of the i^{th} individual

Then,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and the ε_i's are independent.

b) A 1 mm increase in length is associated with a change of β_1 units in mean number of eggs when weight is held constant.

A3_Q4

Joohyeok

2023-10-21

```
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.3     ✓ tibble    3.2.1
## ✓ purrr     1.0.2     ✓ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

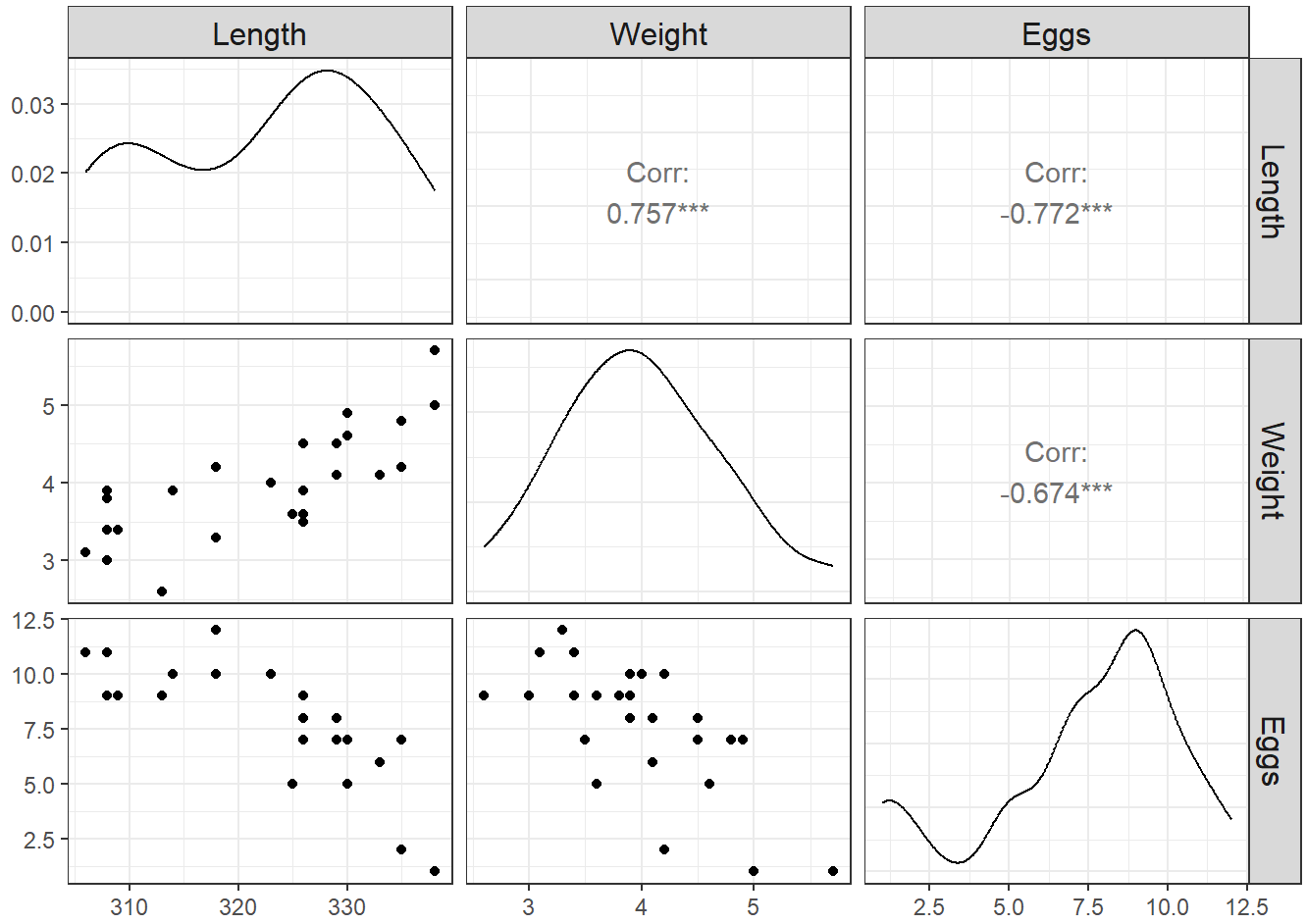
```
library(plotly)
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
df=read.csv("tortoise.csv", header=TRUE)

fit=lm(Eggs~Length+Weight,df)

ggpairs(df,columns=c("Length","Weight","Eggs"))+
  theme_bw()+
  theme(strip.text=element_text(size=12))
```



c. The number of eggs and Length of carapace appear to have a negative correlation. And also, the number of eggs and Wieght of carapace appear to have a negative correlation.

```
confint(fit,parm="Length",level=0.95)
```

```
##              2.5 %      97.5 %
## Length -0.2891413 -0.05380156
```

d. A 95% CI for the effect of length is (-0.2891,-0.0538)

```
summary(fit)
```

```
##
## Call:
## lm(formula = Eggs ~ Length + Weight, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2427 -1.4042  0.3264  1.5152  3.0511
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 66.37758   15.89382   4.176 0.000392 ***
## Length      -0.17147    0.05674  -3.022 0.006265 **
## Weight      -0.87903    0.84682  -1.038 0.310530
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.909 on 22 degrees of freedom
## Multiple R-squared:  0.6145, Adjusted R-squared:  0.5795
## F-statistic: 17.54 on 2 and 22 DF,  p-value: 2.792e-05
```

e. From the summary output, R^2 = 0.6145, meaning that 61.5% of the observed variation in number of eggs is explained by variation in lengths and weights.

```
fit1=lm(Eggs~Length,df)

summary(fit1)
```

```
##
## Call:
## lm(formula = Eggs ~ Length, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2207 -1.7028  0.2972  1.1865  3.4579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77.25258   11.97234   6.453 1.39e-06 ***
## Length      -0.21607    0.03712  -5.821 6.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.912 on 23 degrees of freedom
## Multiple R-squared:  0.5956, Adjusted R-squared:  0.5781
## F-statistic: 33.88 on 1 and 23 DF,  p-value: 6.244e-06
```

f. The test of simple linear regression is about linear relationship between mean number of eggs and length. In contrast, the test of multiple linear regression is the test of whether any additional variation in number of

eggs can be explained by variation in length after having accounted for the variation explained by weight. These tests are therefore of fundamentally different effects (non-adjusted vs. adjusted). The difference in the conclusions is a result of the high correlation between length and weight (they contain similar information about number of eggs)