R^2, Residual Plots

PSTAT 126

Lab 3

```
library(tidyverse) # Easily Install and Load the 'Tidyverse'
library(palmerpenguins) # Palmer Archipelago (Antarctica) Penguin Data
```

Contents

Dataset: Adelie and Gentoo Penguins

- **Question: Can we predict body mass in grams by a penguins bill length in mm?**

Coefficient of Determination R^2

- A goodness-of-fit measure

$$R^2 = 1 - \frac{RSS}{S_{yy}}$$

$$R^2_{adj} = 1 - \frac{RSS/df}{S_{yy}/(n-1)}$$

```
data("penguins")

penguins_noChinstrap <- penguins %>%
  filter(species != "Chinstrap") %>%
  drop_na(bill_length_mm, body_mass_g)

x <- penguins_noChinstrap$bill_length_mm
y <- penguins_noChinstrap$body_mass_g
x_bar <- mean(x)
y_bar <- mean(y)
n <- length(x)

model <- lm(body_mass_g ~ bill_length_mm , data = penguins_noChinstrap)
summary(model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_length_mm, data = penguins_noChinstrap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -891.91 -272.91   -0.82  282.47 1279.63
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1706.821    201.712  -8.462 1.65e-15 ***
## bill_length_mm   141.088      4.689  30.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 402.5 on 272 degrees of freedom
## Multiple R-squared:  0.769,  Adjusted R-squared:  0.7681
## F-statistic: 905.3 on 1 and 272 DF,  p-value: < 2.2e-16
```

```
b0 <- summary(model)$coef[1,1] # Intercept
b1 <- summary(model)$coef[2,1] # Slope
y_hat <- b0 + b1*x # Fitted values
e <- y - y_hat # Residuals

Syy <- sum((y - y_bar)^2)

r_2 <- 1 - (sum(e^2)/Syy)
r_2
```

```
## [1] 0.7689629
```

```
summary(model)$r.squared
```

```
## [1] 0.7689629
```

```
r <- cor(x,y)
r^2
```

```
## [1] 0.7689629
```

```
adj_r2 <- 1 - (sum(e^2)/(n-2))/(Syy/(n-1))
adj_r2
```

```
## [1] 0.7681135
```

```
summary(model)$adj.r.squared
```

```
## [1] 0.7681135
```

Notes on R^2

- Always between 0 and 1
- Can interpret as $R^2 \times 100$ percent of the variation in Y is explained by the variation in the predictor x.

Simple Linear Regression Model Assumptions

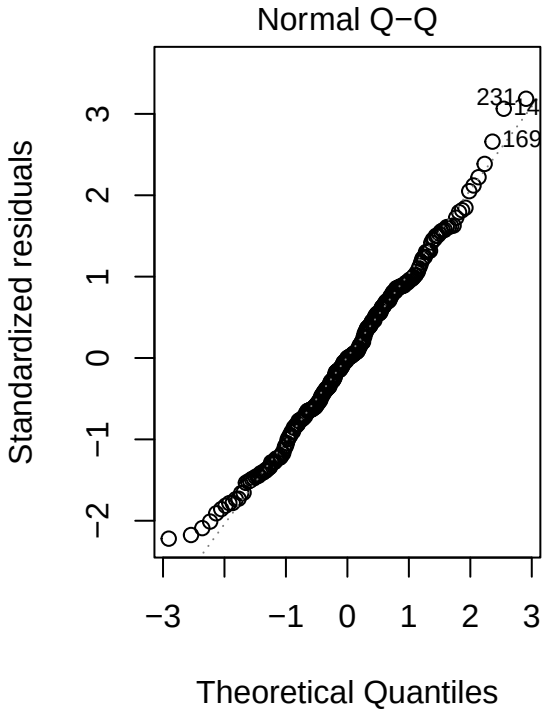
1) The relationship between each Y_n and each x_n, respectively, is linear. **L**inearity
2) Errors have **E**qual variance. $\text{Var}(Y_n) = \sigma^2$ for every n (homoscedasticity)
3) Errors are **N**ormally distributed
4) Errors are **I**ndependent

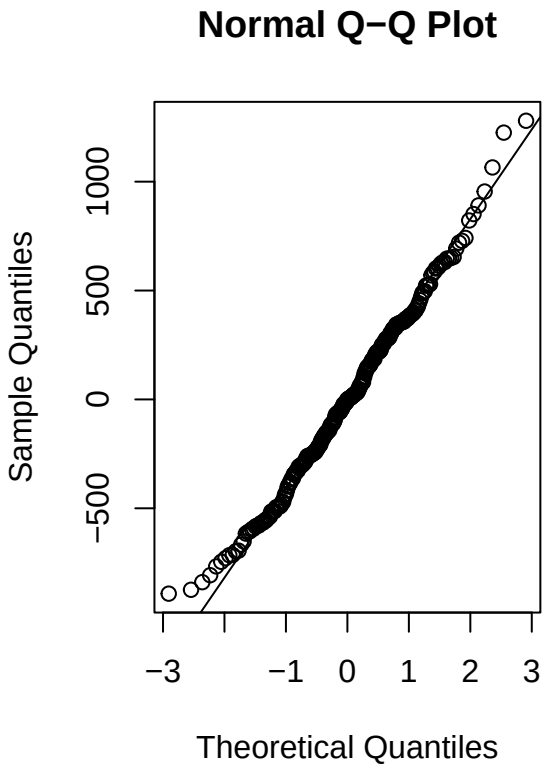
- Can use the acronym **L.I.N.E.** to help you remember.

Graphically checking the normality assumption

QQ - plot

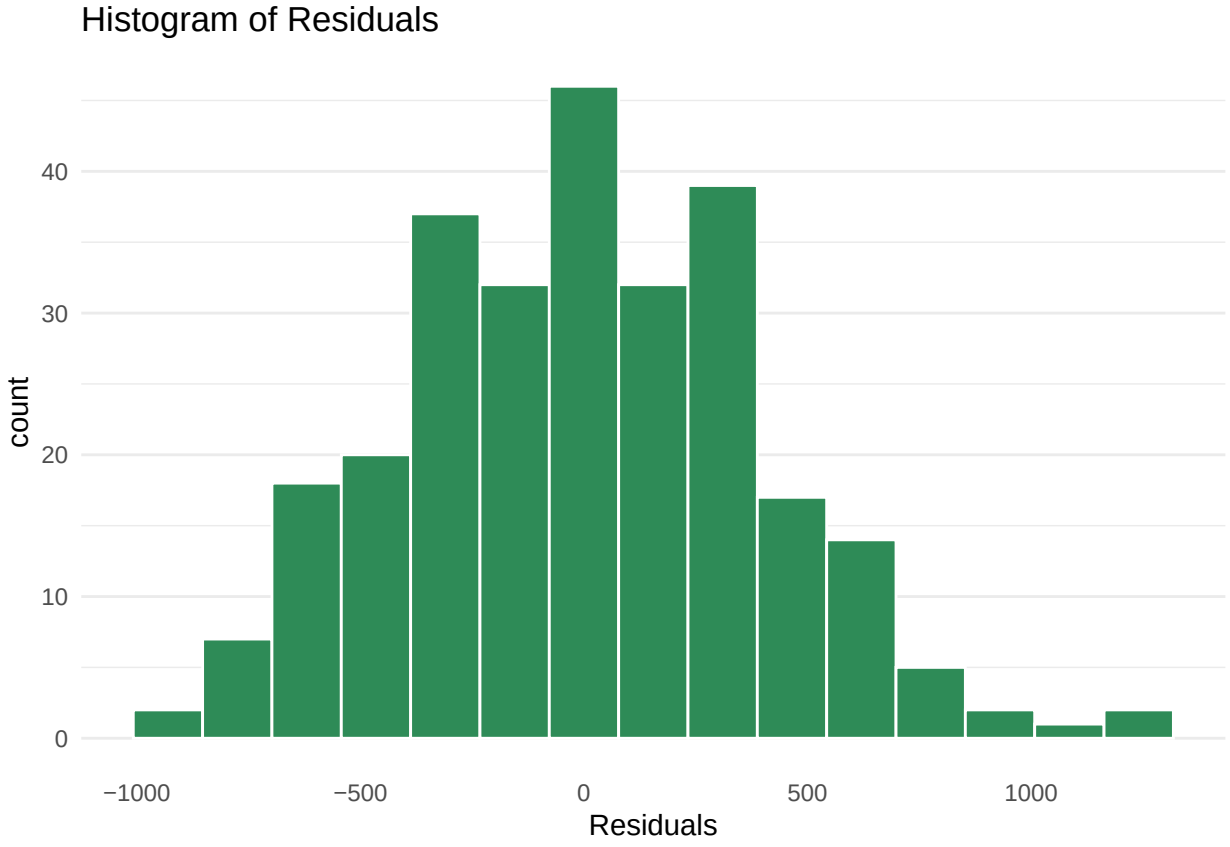
```
par(mfrow = c(1, 2))

plot(model, which = 2) # QQ

e <- residuals(model) # Residuals

qqnorm(e) # QQ
qqline(e)
```



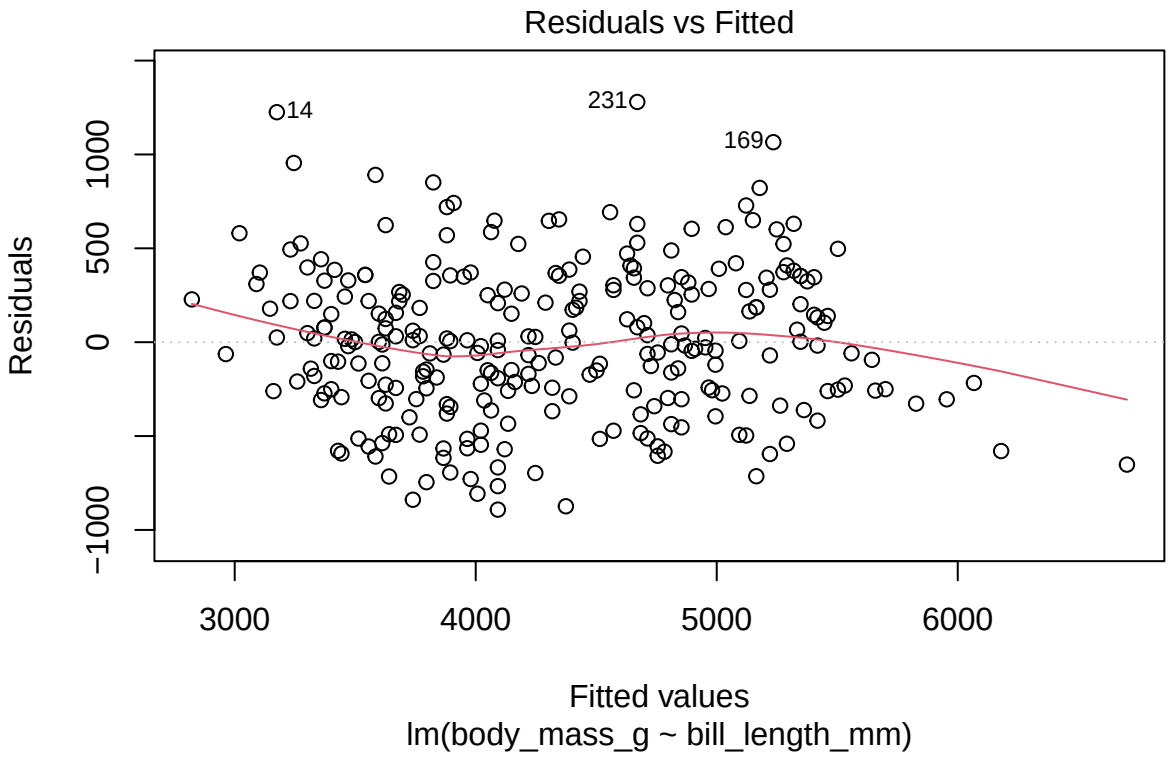


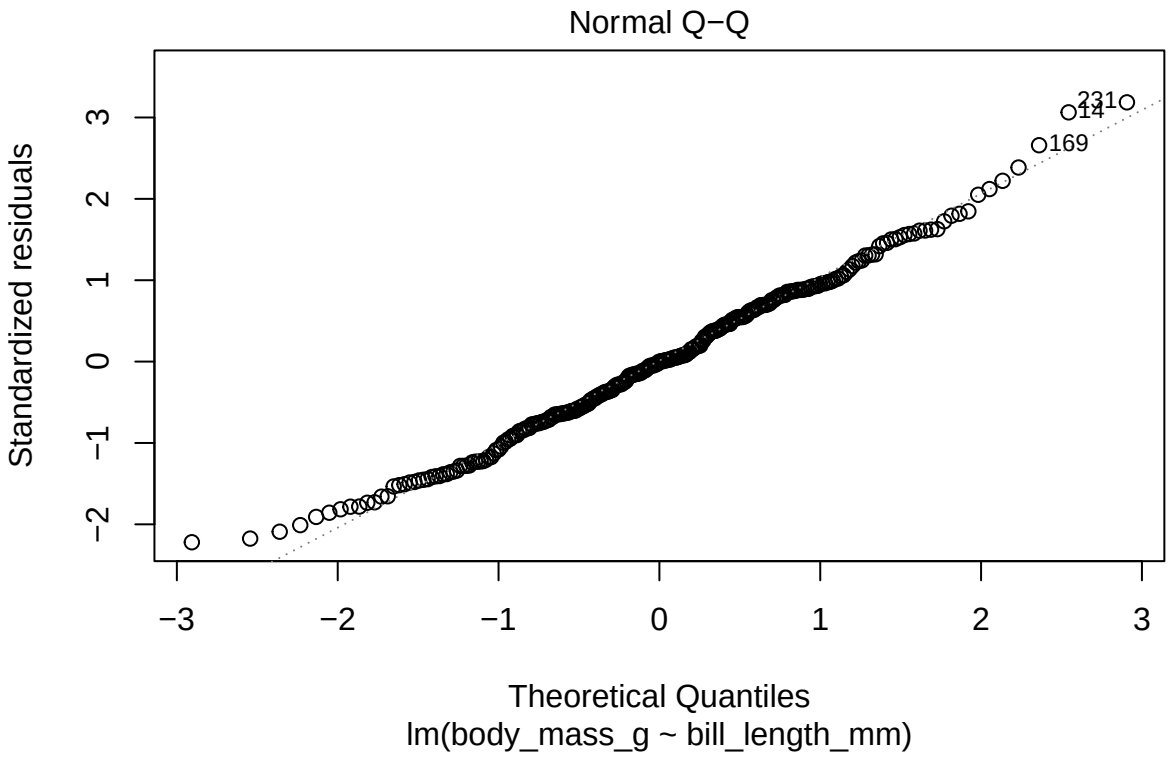
Histogram of residuals

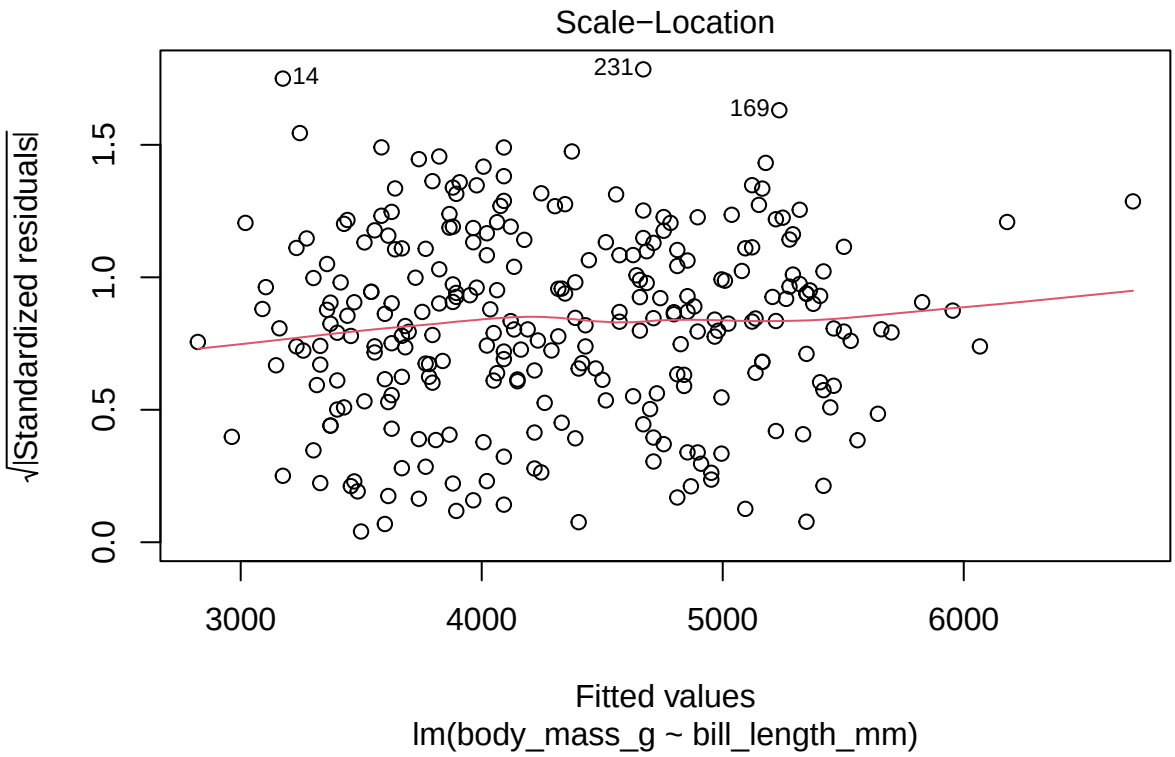
```
par(mfrow = c(1, 1))
resid_model <- tibble(residuals = residuals(model))

ggplot(data = resid_model,
       aes(x = residuals)) +
  geom_histogram(color = "white",
                 fill = "seagreen",
                 bins = 15) +
  labs(x = "Residuals",
       title = "Histogram of Residuals") +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())
```

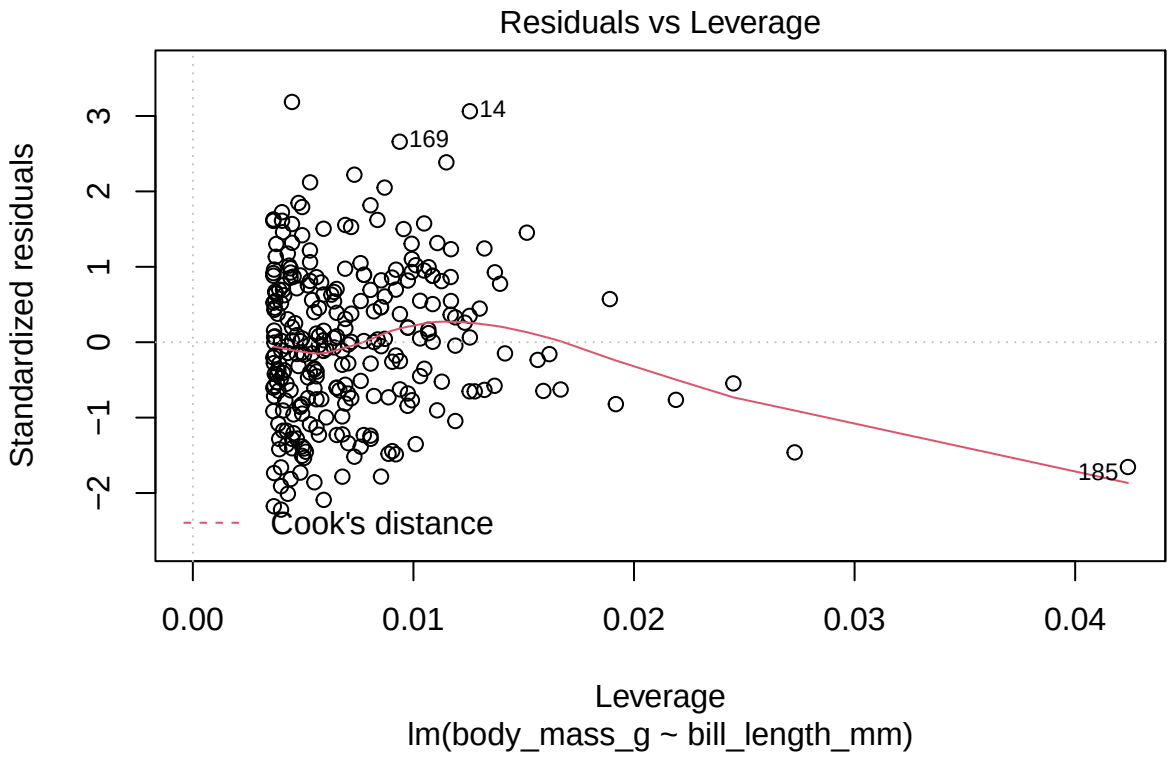



```
plot(model) #residual graph
```


Residuals vs Fitted
14
231
169
Residuals
1000
500
0
−1000
3000
4000
5000
6000
Fitted values
lm(body_mass_g ~ bill_length_mm)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(body_mass_g ~ bill_length_mm)
231
14
169
−3
−2
−1
0
1
2
3


Scale−Location
14
231
169
√|Standardized residuals|
0.0
0.5
1.0
1.5
3000
4000
5000
6000
Fitted values
lm(body_mass_g ~ bill_length_mm)




```
##In diagnostic plot 1: Residuals vs Fitted
##  If line is close to horizontal line, then the model is close to data fit.
##In diagnostic plot 2: Normal QQ
##  If points follow a almost straight line, then the residuals normally distributed.
##In diagnostic plot 3: Scale Location
##  If line is not variant, the variance is expected to be good.
##In diagnostic plot 4: Residuals vs Leverage
##  If we don't see points outside the cook's distance lines, then expect no influential points.

summary(model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_length_mm, data = penguins_noChinstrap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -891.91 -272.91   -0.82  282.47 1279.63
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1706.821    201.712  -8.462 1.65e-15 ***
## bill_length_mm   141.088      4.689  30.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 402.5 on 272 degrees of freedom
## Multiple R-squared:  0.769,  Adjusted R-squared:  0.7681
## F-statistic: 905.3 on 1 and 272 DF,  p-value: < 2.2e-16
```