PSTAT 126

Lab 7 Part 2

Roupen Khanjian

Spring 2021

```r
library(faraway) # Functions and Datasets for Books by Julian Faraway
library(alr4) # Data to Accompany Applied Linear Regression 4th Edition
library(tidyverse) # Easily Install and Load the 'Tidyverse'
library(lmtest) # Testing Linear Regression Models
```

Contents

- Note on Homework. Make sure you are knitting as you go along. Don't wait until you do all your analysis and then knit.

ANOVA using the aov function

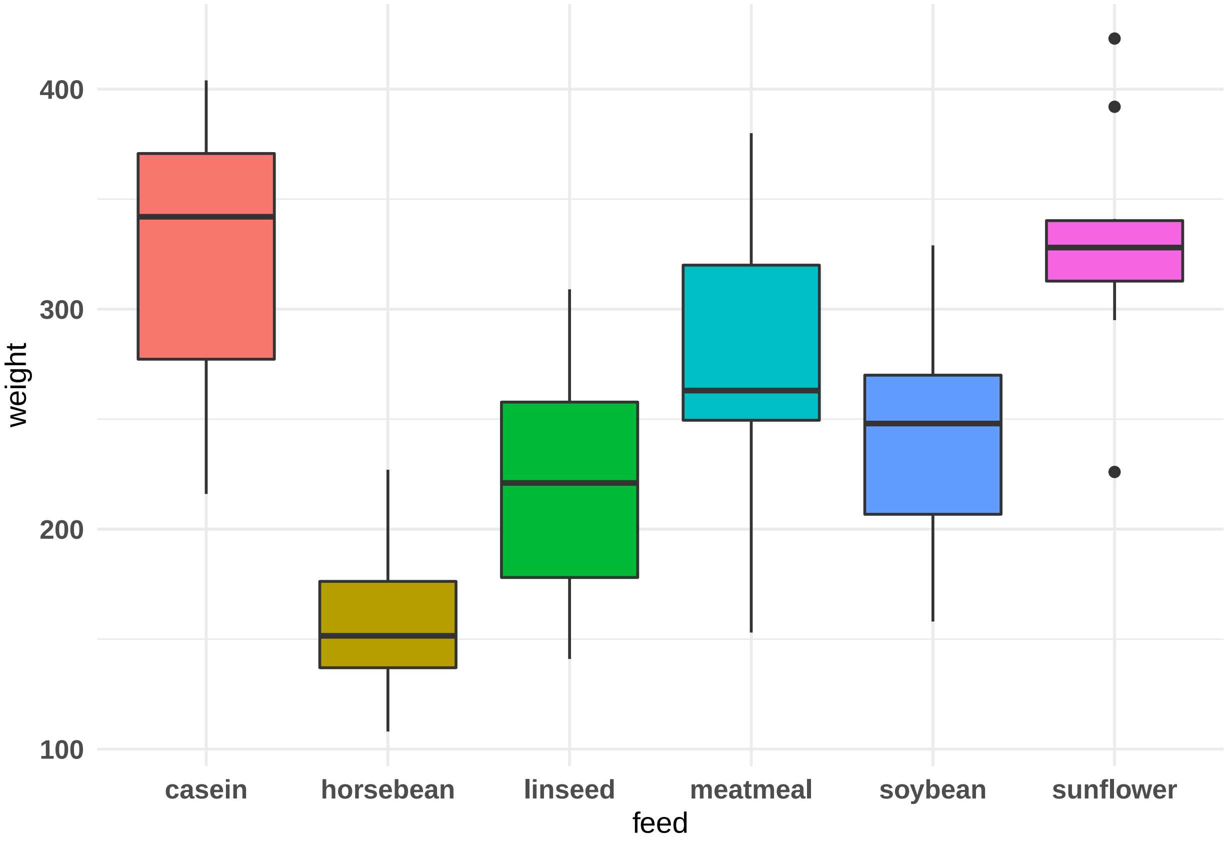
```
data("chickwts")
glimpse(chickwts)
```

```
## Rows: 71
## Columns: 2
## $ weight <dbl> 179, 160, 136, 227, 217, 168, 108, 124, 143, 140, 309, 229, 181~
## $ feed   <fct> horsebean, horsebean, horsebean, horsebean, horsebean, horsebea~
```

```
table(chickwts$feed)
```

```
## 
##    casein horsebean   linseed  meatmeal   soybean sunflower 
##        12        10        12        11        14        12
```

```
ggplot(data = chickwts,
       aes(x = feed, y = weight)) +
  geom_boxplot(aes(fill = feed),
               show.legend = FALSE) +
  theme_minimal() +
  theme(axis.text = element_text(face = "bold",
                                 size = 10))
```



```
chick_wt_aov <- aov(weight ~ feed, data = chickwts)
chick_wt_aov
```

```
## Call:
##    aov(formula = weight ~ feed, data = chickwts)
##
## Terms:
##                      feed Residuals
## Sum of Squares  231129.2  195556.0
## Deg. of Freedom        5        65
##
## Residual standard error: 54.85029
## Estimated effects may be unbalanced
```

```
summary(chick_wt_aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(weight ~ feed, data = chickwts))
```

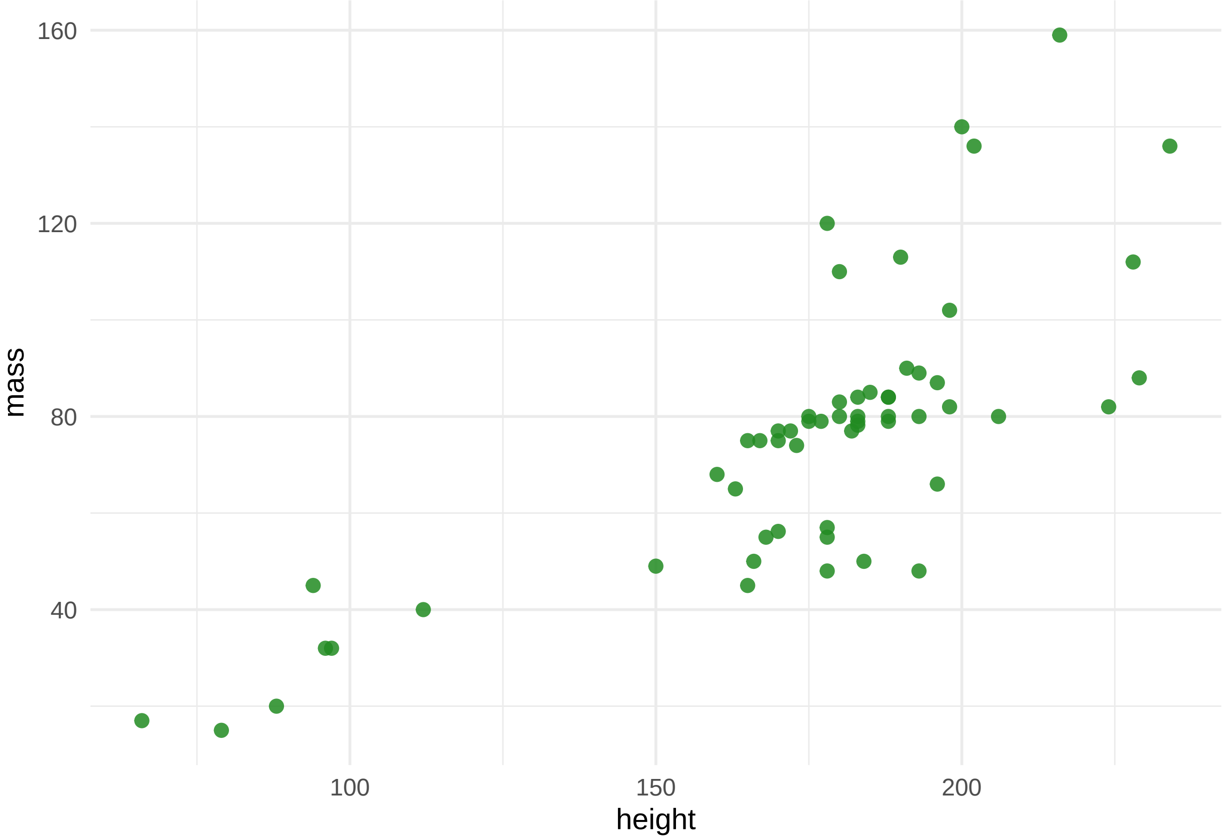
```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## feed       5 231129   46226  15.365 5.936e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Data example

```
data("starwars")
head(starwars)
```

```
## # A tibble: 6 x 14
##   name      height  mass hair_color   skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>        <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sk~     172    77 blond        fair       blue            19   male  mascu~
## 2 C-3PO        167    75 <NA>         gold       yellow         112   none  mascu~
## 3 R2-D2         96    32 <NA>         white, bl~ red             33   none  mascu~
## 4 Darth V~     202   136 none         white      yellow          41.9 male  mascu~
## 5 Leia Or~     150    49 brown        light      brown           19   fema~ femin~
## 6 Owen La~     178   120 brown, grey light       blue            52   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

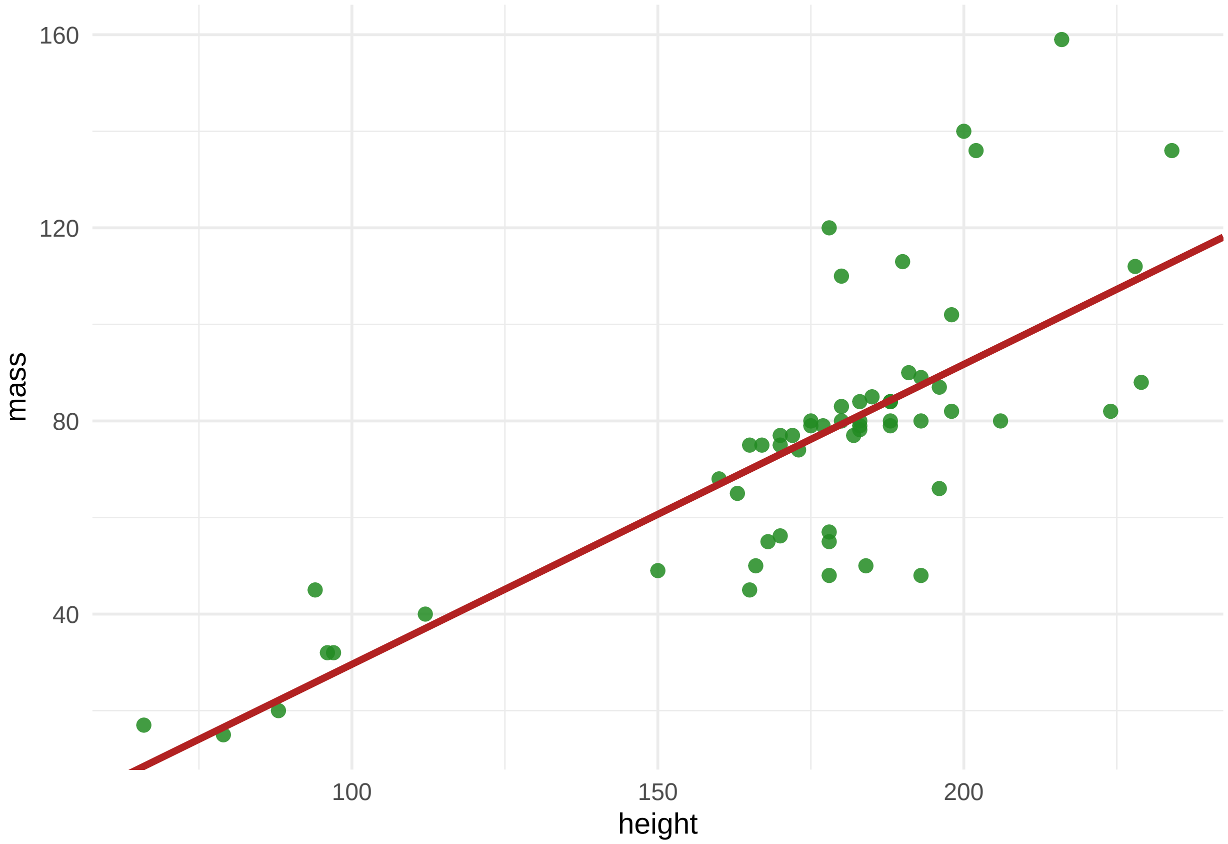
```
starwars_subset <- starwars %>% # subset data
  select(height, mass) %>% # select mass as response and height as predictor
  drop_na() %>% # remove missing values
  filter(mass < 500) # filter for mass under 500 kg

ggplot(data = starwars_subset,
       aes(x = height, y = mass)) +
  geom_point(color = "forestgreen",
             alpha = 0.85, size = 2) +
  theme_minimal()
```



```r
model_sw <- lm(mass ~ height, starwars_subset)
summary(model_sw)
```

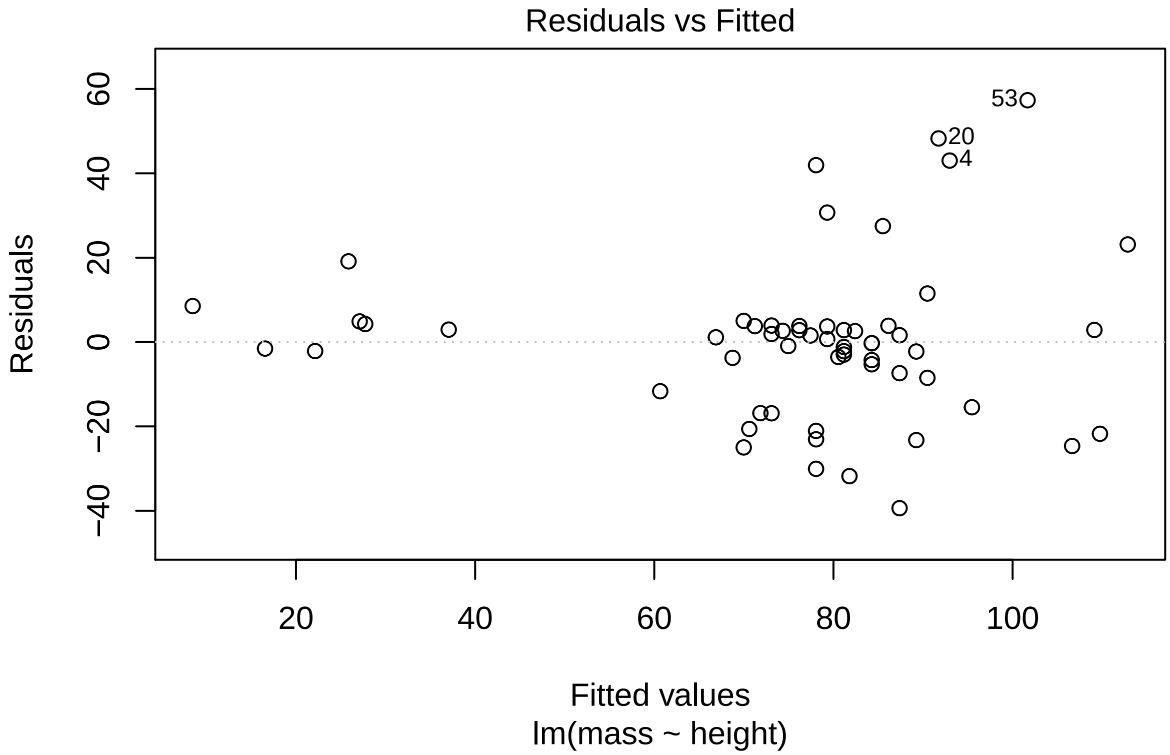
```
##
## Call:
## lm(formula = mass ~ height, data = starwars_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.382  -8.212   0.211   3.846  57.327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.54076   12.56053  -2.591   0.0122 *
## height        0.62136    0.07073   8.785 4.02e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.14 on 56 degrees of freedom
## Multiple R-squared:  0.5795, Adjusted R-squared:  0.572
## F-statistic: 77.18 on 1 and 56 DF,  p-value: 4.018e-12
```

```r
ggplot(data = starwars_subset,
       aes(x = height, y = mass)) +
  geom_point(color = "forestgreen",
             alpha = 0.85, size = 2) +
  geom_abline(aes(intercept = coef(model_sw)[1],
                  slope = coef(model_sw)[2]),
              color = "firebrick", size = 1.25) +
  theme_minimal()
```

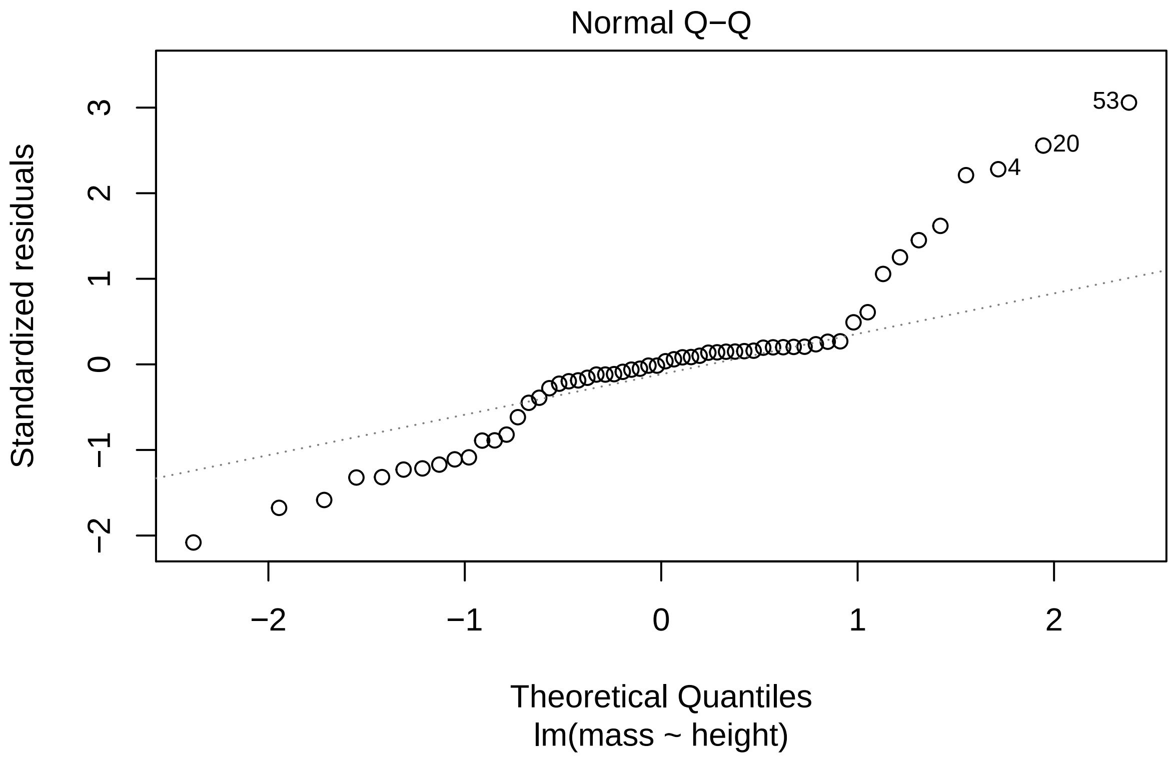


- check model assumptions graphically and using formal tests.

```
plot(model_sw, which = 1, add.smooth = F)
```



```r
plot(model_sw, which = 2)
```



```r
bptest(model_sw)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_sw
## BP = 4.827, df = 1, p-value = 0.02802
```

```r
shapiro.test(resid(model_sw))
```

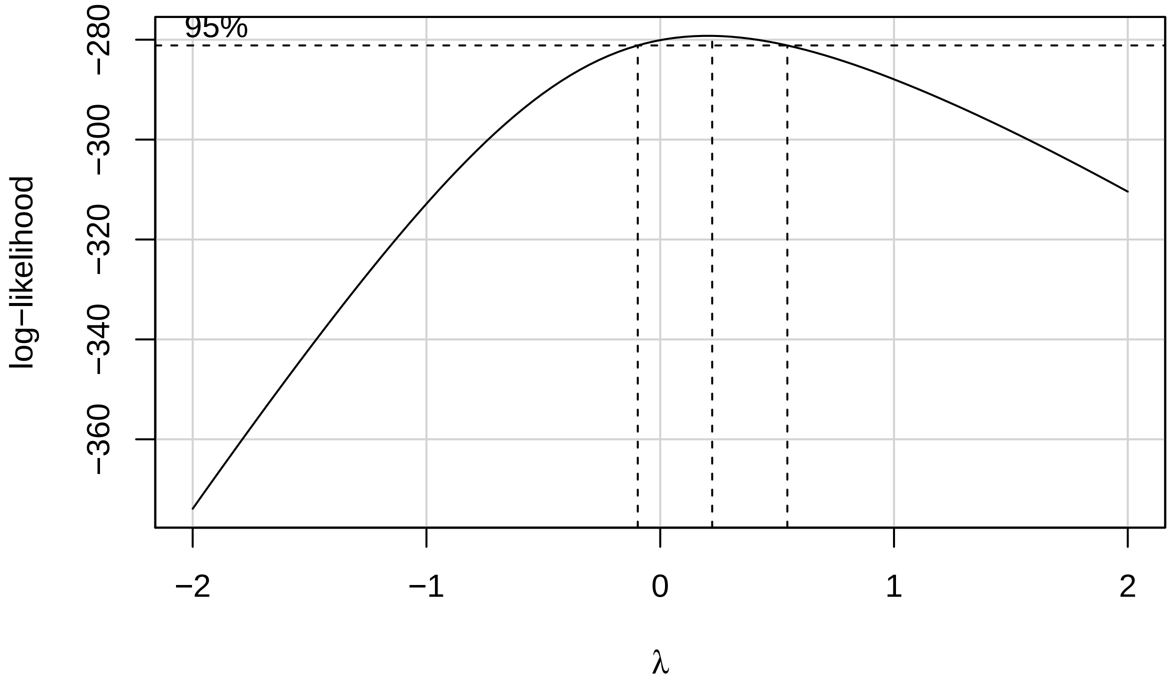
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_sw)
## W = 0.912, p-value = 0.0004688
```

```r
shapiro.test(residuals(model_sw))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_sw)
## W = 0.912, p-value = 0.0004688
```

- assumptions are violated
- try response transformation

```r
bc <- boxCox(model_sw)
```



```
bc$x[which.max(bc$y)]
```

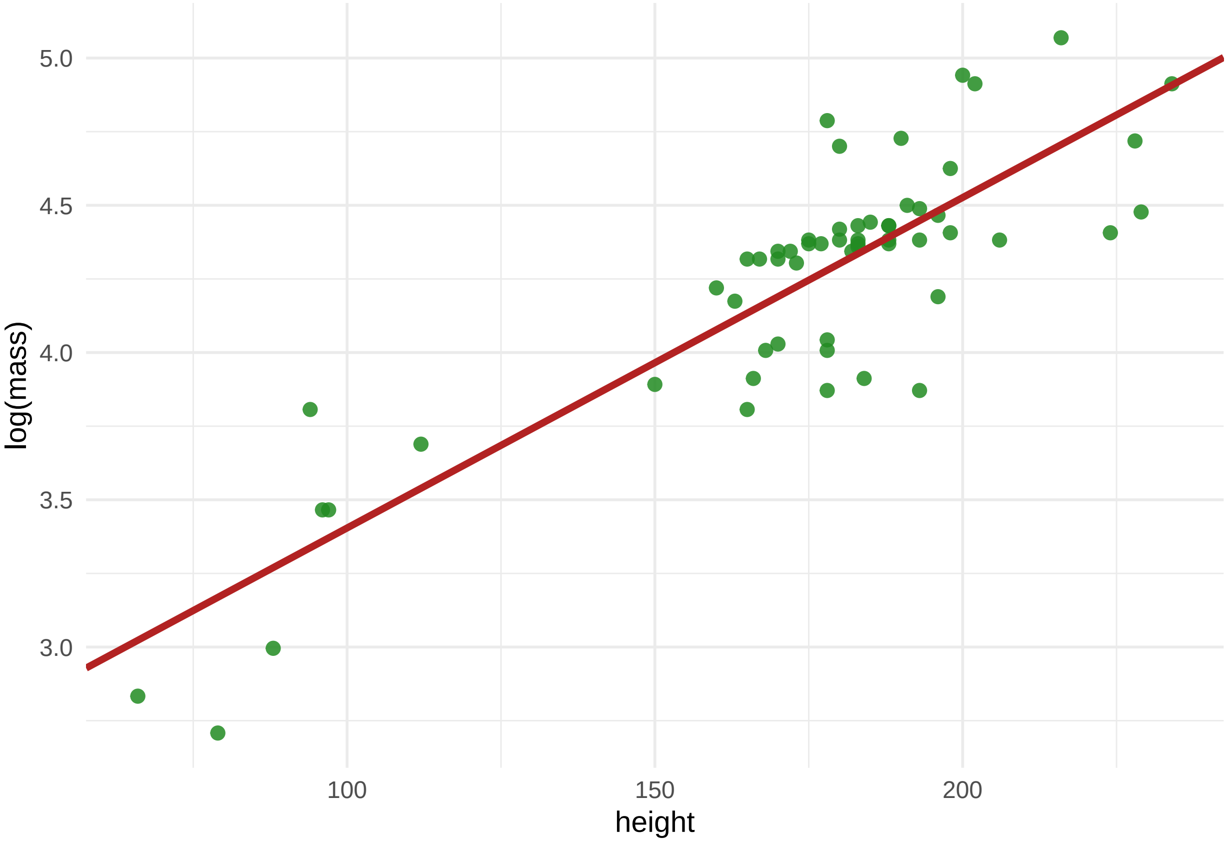
```
## [1] 0.2222222
```

lets try to log transform response

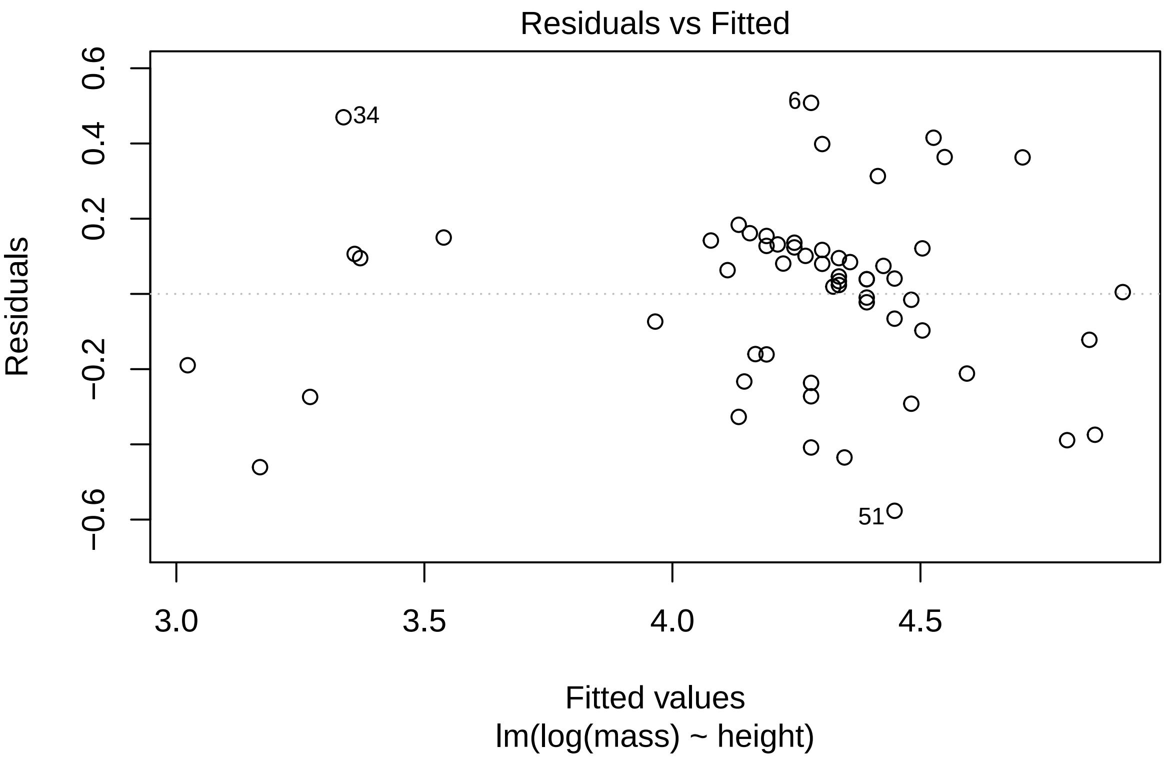
```
model_sw2 <- lm(log(mass) ~ height, data = starwars_subset)
summary(model_sw2)
```

```
##
## Call:
## lm(formula = log(mass) ~ height, data = starwars_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57664 -0.16064  0.03993  0.12668  0.50795
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2822928  0.1589110   14.36   <2e-16 ***
## height      0.0112205  0.0008948   12.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2421 on 56 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7327
## F-statistic: 157.2 on 1 and 56 DF,  p-value: < 2.2e-16
```

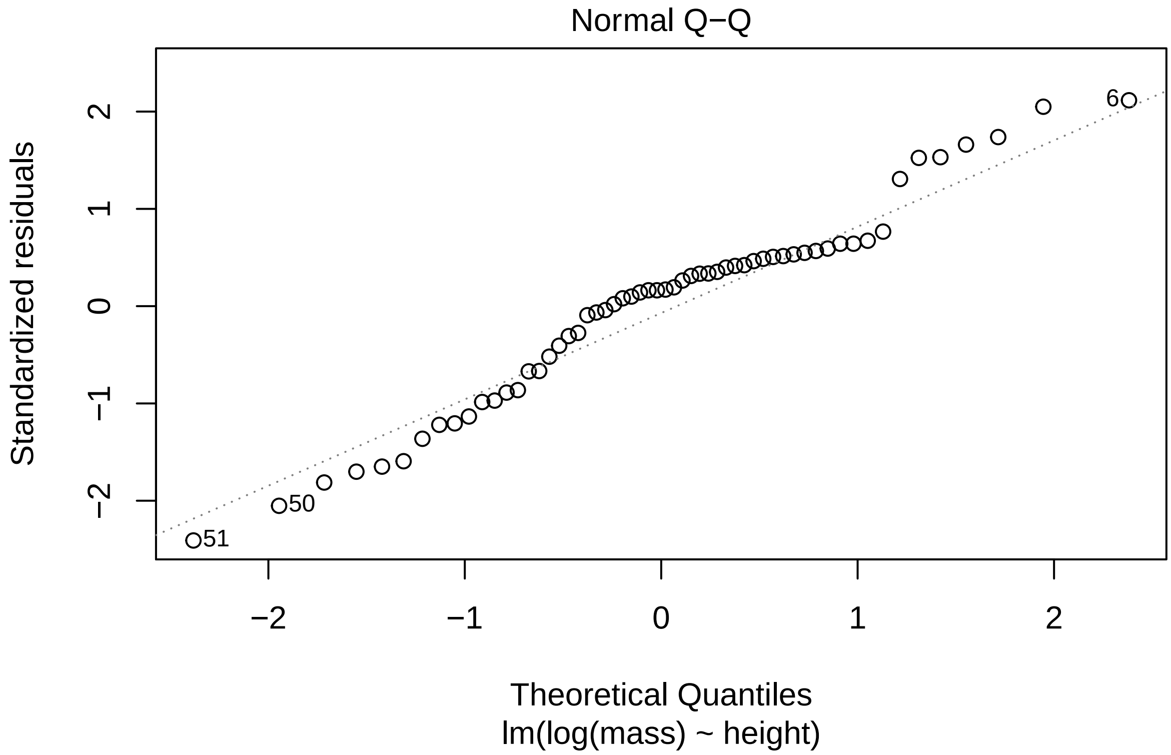
```
ggplot(data = starwars_subset,
       aes(x = height, y = log(mass))) +
  geom_point(color = "forestgreen",
             alpha = 0.85, size = 2) +
  geom_abline(aes(intercept = coef(model_sw2)[1],
                  slope = coef(model_sw2)[2]),
              color = "firebrick", size = 1.25) +
  theme_minimal()
```



```
plot(model_sw2, which = 1, add.smooth = F)
```



```r
plot(model_sw2, which = 2)
```



```r
bptest(model_sw2)
```

```
##
##  studentized Breusch-Pagan test
```

```
##
## data:  model_sw2
## BP = 0.10309, df = 1, p-value = 0.7482
```

```
shapiro.test(resid(model_sw2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_sw2)
## W = 0.96702, p-value = 0.1156
```

- Should we add a polynomial term maybe? Let's see!

```
model_sw3 <- lm(log(mass) ~ height + I(height^2), data = starwars_subset)
summary(model_sw3)
```

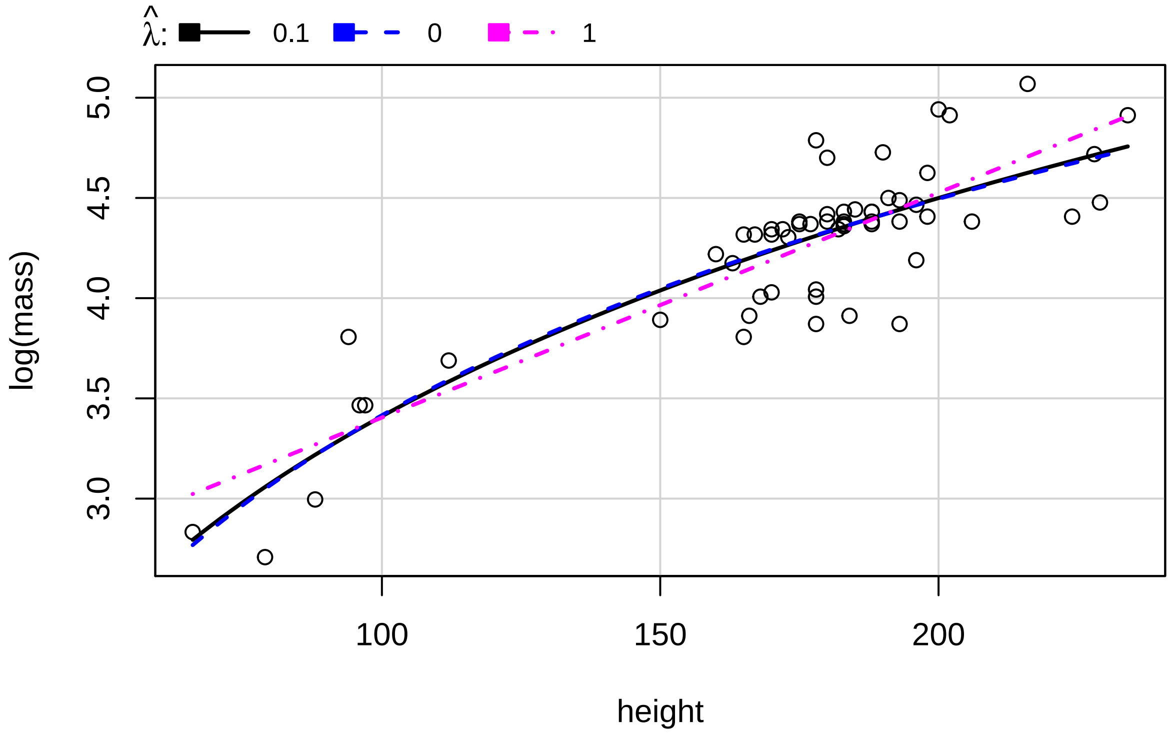
```
##
## Call:
## lm(formula = log(mass) ~ height + I(height^2), data = starwars_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5761 -0.1617  0.0276  0.1086  0.5069
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.674e+00  3.976e-01   4.209 9.55e-05 ***
## height       2.008e-02  5.392e-03   3.724 0.000464 ***
## I(height^2) -2.957e-05  1.776e-05  -1.665 0.101554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2384 on 55 degrees of freedom
## Multiple R-squared:   0.75,  Adjusted R-squared:  0.7409
## F-statistic: 82.49 on 2 and 55 DF,  p-value: < 2.2e-16
```

```
anova(model_sw2, model_sw3)
```

```
## Analysis of Variance Table
##
## Model 1: log(mass) ~ height
## Model 2: log(mass) ~ height + I(height^2)
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     56 3.2828
## 2     55 3.1252  1   0.15757 2.773 0.1016
```

- should we transform the predictor variable?

```
invTranPlot(log(mass) ~ height, data = starwars_subset,
            lambda = c(0, 1), optimal = TRUE)
```



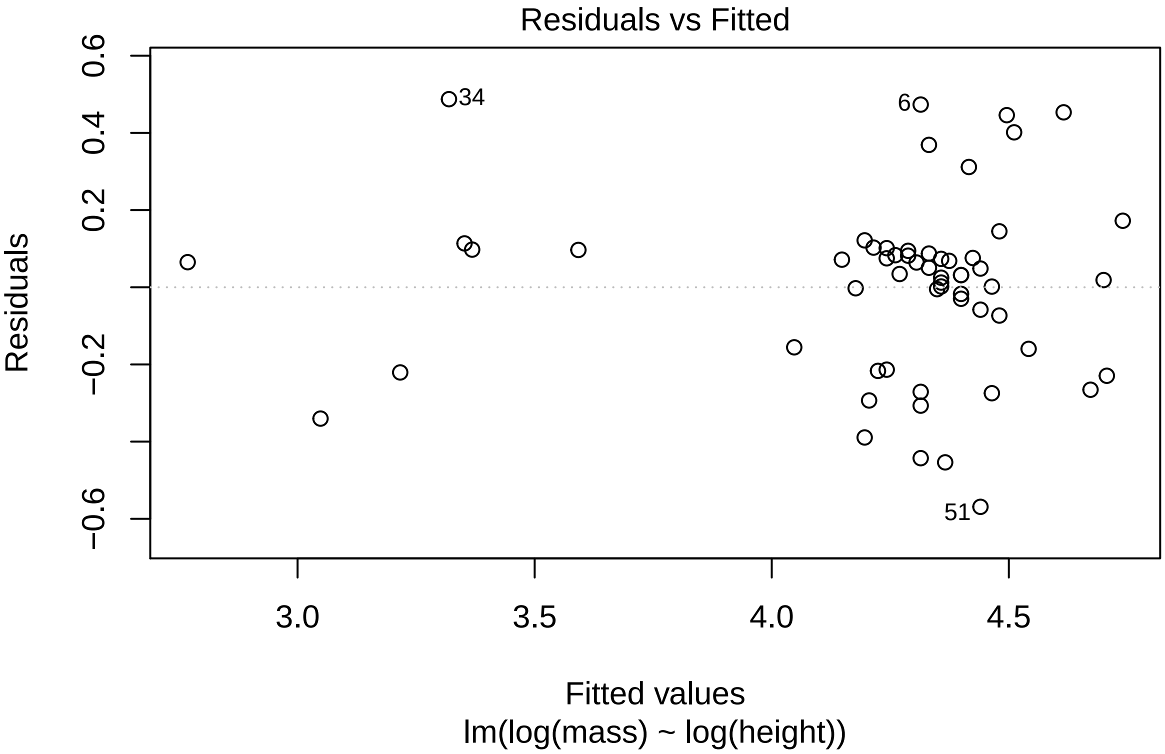
```
##      lambda      RSS
## 1 0.1019086 3.096695
## 2 0.0000000 3.099013
## 3 1.0000000 3.282768
```

- From the plot above can try a log transformation of predictor variable.

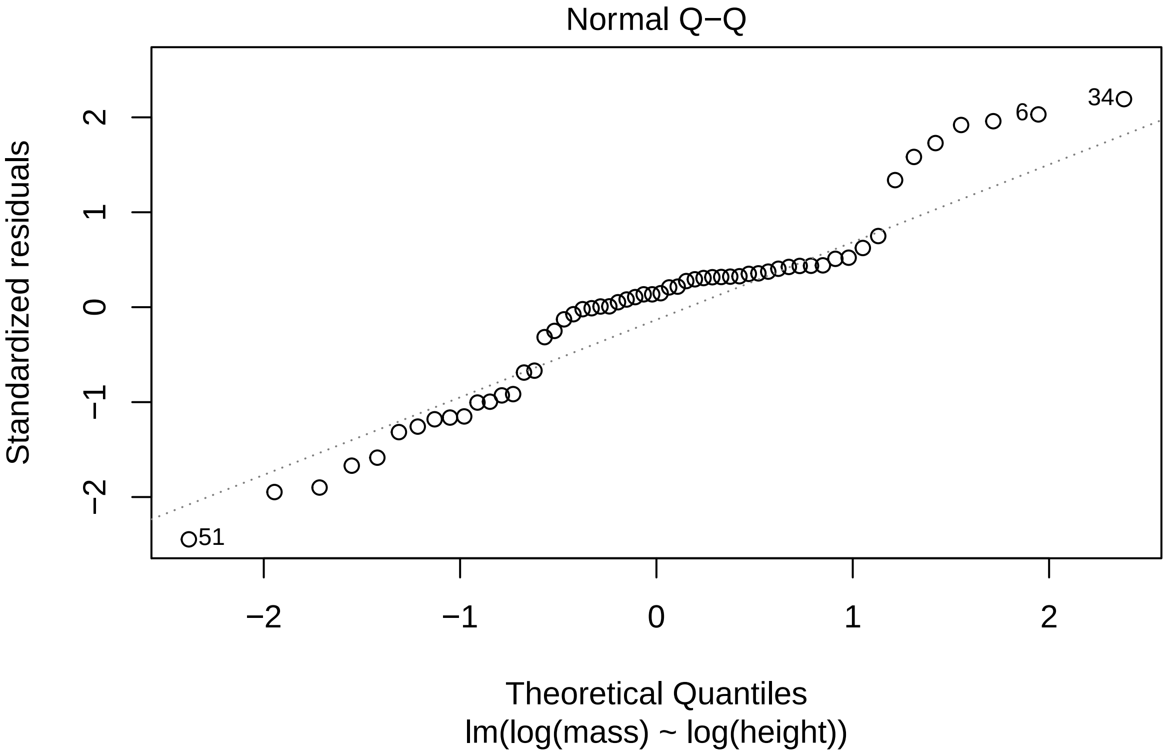
```
model_sw4 <- lm(log(mass) ~ log(height), data = starwars_subset)
summary(model_sw4)
```

```
##
## Call:
## lm(formula = log(mass) ~ log(height), data = starwars_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56899 -0.15870  0.03293  0.09607  0.48741
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.7599     0.6142  -6.122 9.61e-08 ***
## log(height)   1.5582     0.1195  13.034  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2352 on 56 degrees of freedom
## Multiple R-squared:  0.7521, Adjusted R-squared:  0.7477
## F-statistic: 169.9 on 1 and 56 DF,  p-value: < 2.2e-16
```

```r
plot(model_sw4, which = 1, add.smooth = F)
```



```r
plot(model_sw4, which = 2)
```



```r
bptest(model_sw4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_sw4
## BP = 0.00033249, df = 1, p-value = 0.9855
```

```
shapiro.test(resid(model_sw4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_sw4)
## W = 0.94641, p-value = 0.01249
```

- Normality assumption is violated.

Therefore, the most appropriate model for this data would be model 2

```
summary(model_sw2)
```

```
##
## Call:
## lm(formula = log(mass) ~ height, data = starwars_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57664 -0.16064  0.03993  0.12668  0.50795
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2822928  0.1589110   14.36   <2e-16 ***
## height      0.0112205  0.0008948   12.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2421 on 56 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7327
## F-statistic: 157.2 on 1 and 56 DF,  p-value: < 2.2e-16
```