Chapter 23

Alan T. Arnholt

Last compiled: May 10, 2022 at 09:29:20 AM

Contents

1 Inferences for Regression

```
bodyfat <- read.csv("./DATA/Bodyfat.csv") %>%
  clean_names()
head(bodyfat)
```

```
  density pct_bf age weight height neck chest abdomen    waist   hip thigh knee
1  1.0708   12.3  23 154.25  67.75 36.2  93.1    85.2 33.54331  94.5  59.0 37.3
2  1.0853    6.1  22 173.25  72.25 38.5  93.6    83.0 32.67717  98.7  58.7 37.3
3  1.0414   25.3  22 154.00  66.25 34.0  95.8    87.9 34.60630  99.2  59.6 38.9
4  1.0751   10.4  26 184.75  72.25 37.4 101.8    86.4 34.01575 101.2  60.1 37.3
5  1.0340   28.7  24 184.25  71.25 34.4  97.3   100.0 39.37008 101.9  63.2 42.2
6  1.0502   20.9  24 210.25  74.75 39.0 104.5    94.4 37.16535 107.8  66.0 42.0
  ankle bicep forearm wrist
1  21.9  32.0    27.4  17.1
2  23.4  30.5    28.9  18.2
3  24.0  28.8    25.2  16.6
4  22.8  32.4    29.4  18.2
5  24.0  32.2    27.7  17.7
6  25.6  35.7    30.6  18.8
```

```
ggplot(data = bodyfat, aes(x = waist, y = pct_bf)) +
  geom_point(color = "blue") +
  theme_bw() +
  labs(x = "Waist (in.)", y = "% Body Fat")
```

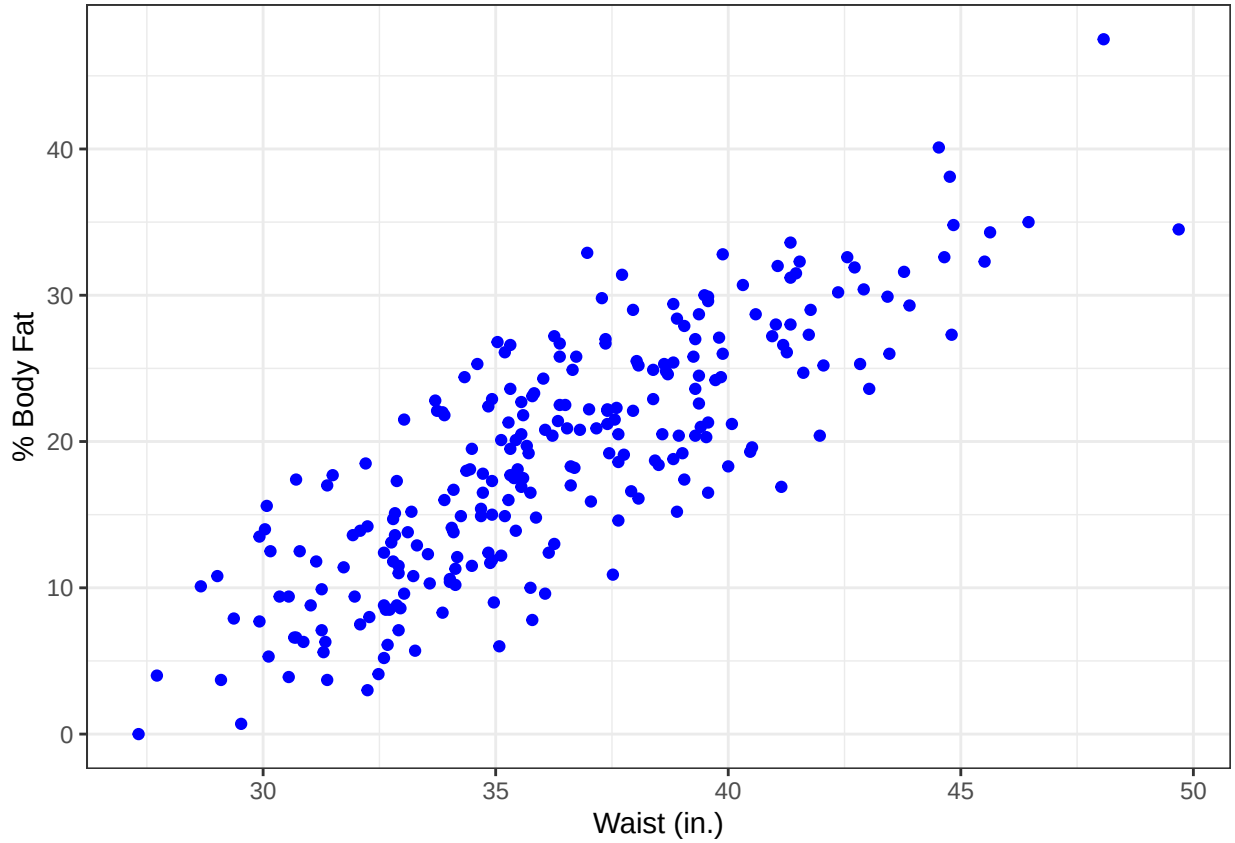


Figure 1: Percent body fat versus waist size for 250 men of various ages. The scatterplot shows a strong, positive, linear relationship.

1.1 Fitting a least squares model to Figure 1

```
mod_lm <- lm(pct_bf ~ waist, data = bodyfat)
mod_lm
```

```
Call:
lm(formula = pct_bf ~ waist, data = bodyfat)

Coefficients:
(Intercept)        waist
     -42.73         1.70
```

```
summary(mod_lm)
```

```
Call:
lm(formula = pct_bf ~ waist, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8987  -3.6453   0.1864   3.1775  12.7887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.73413    2.71651  -15.73   <2e-16 ***
waist         1.69997    0.07431   22.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.713 on 248 degrees of freedom
Multiple R-squared:  0.6785,    Adjusted R-squared:  0.6772
F-statistic: 523.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

```
library(moderndive)
get_regression_table(mod_lm)
```

```
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    -42.7     2.72      -15.7       0   -48.1    -37.4
2 waist          1.7     0.074      22.9       0     1.55     1.85
```

- Review on the board z and t scores.
- Review t statistics from regression output.
- Review confidence intervals and their derivation.

```
summary(mod_lm)$coef
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -42.734134 2.71650558 -15.73129 3.826300e-39
waist         1.699972 0.07431472  22.87530 4.846616e-63
```

```
b1 <- summary(mod_lm)$coef[2, 1]
seb1 <- summary(mod_lm)$coef[2, 2]
c(b1, seb1, b1/seb1, pt(b1/seb1, 248, lower = FALSE)*2)
```

```
[1] 1.699972e+00 7.431472e-02 2.287530e+01 4.846616e-63
```

```
#
# CI for beta_1
b1 + c(-1,1)*qt(.975, 248)*seb1
```

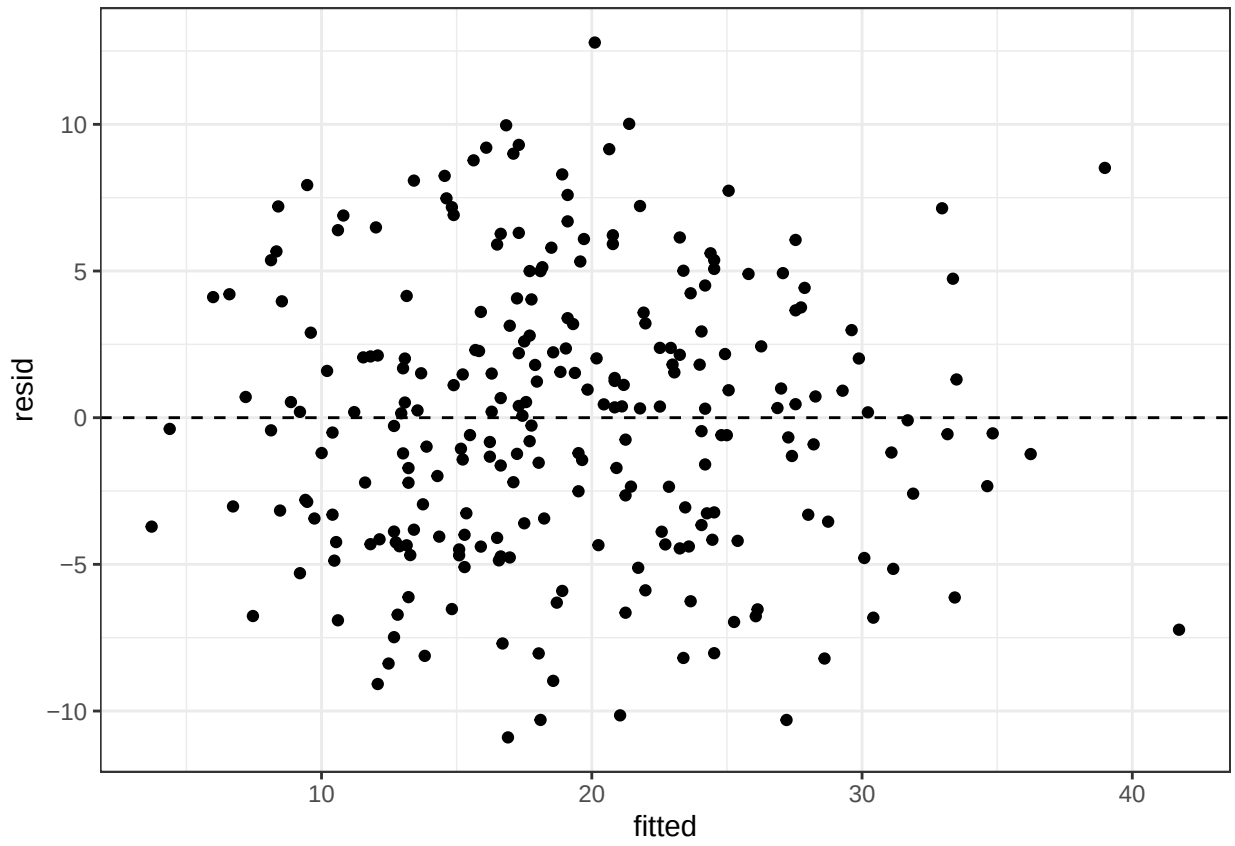
```
[1] 1.553603 1.846340
```

```
#
confint(mod_lm, level = 0.95)
```

```
                2.5 %    97.5 %
(Intercept) -48.084497 -37.38377
waist         1.553603   1.84634
```

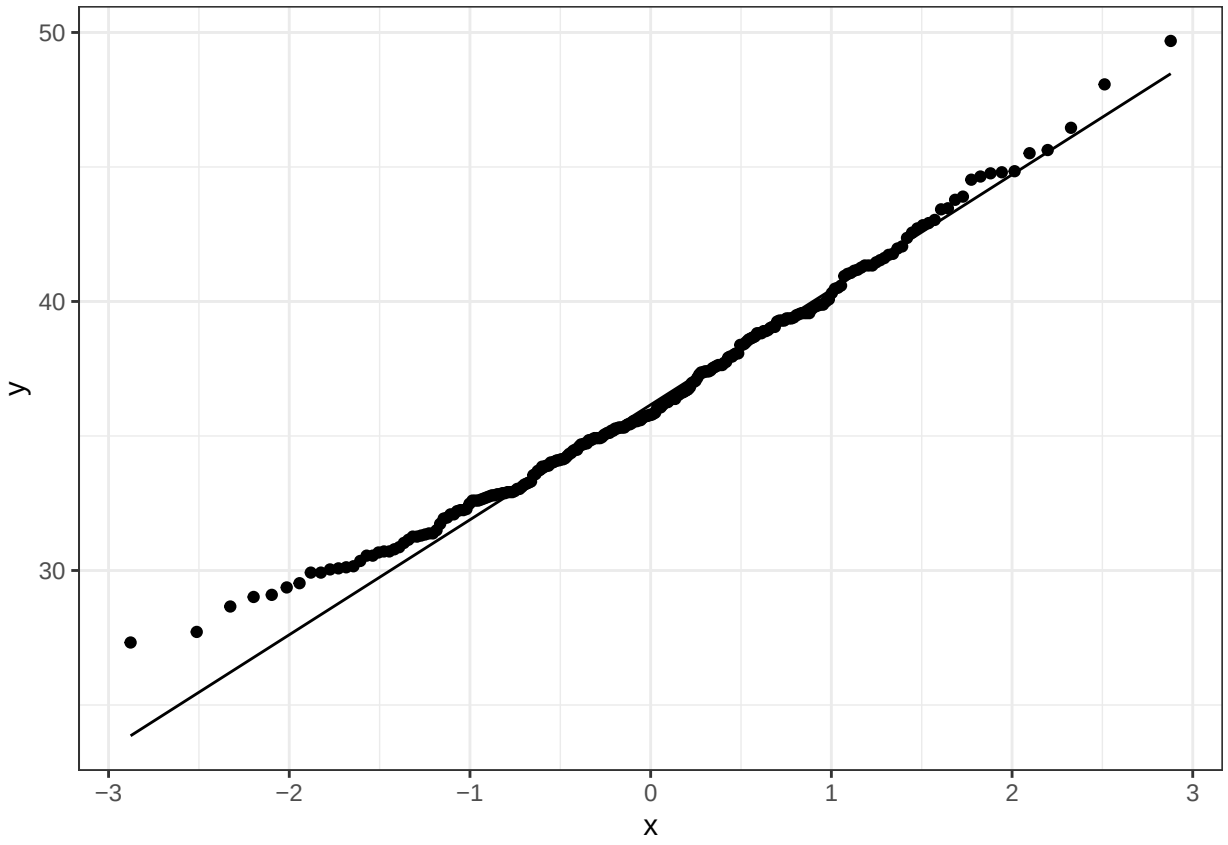
1.2 Residual and Q-Q Plots

```
# With ggplot
library(broom)
augment(mod_lm) %>%
  clean_names() -> aug_mod
ggplot(data = aug_mod, aes(x = fitted, y = resid)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, linetype = "dashed")
```

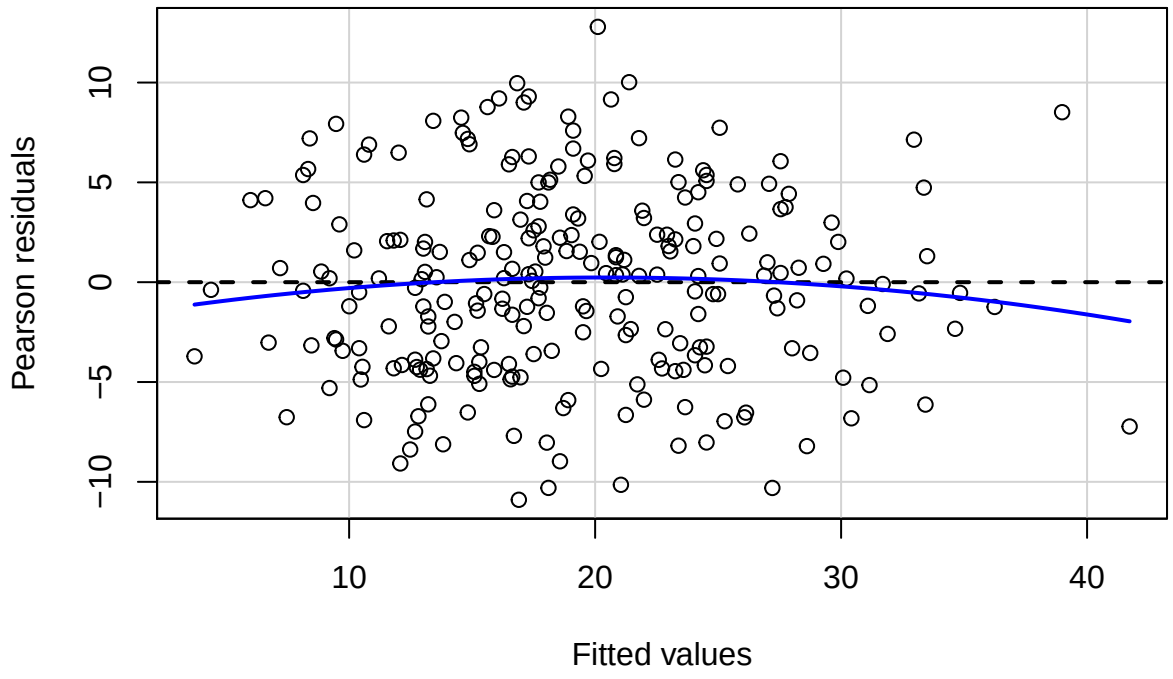



```
ggplot(data = aug_mod, aes(sample = waist)) +
  geom_qq() +
```

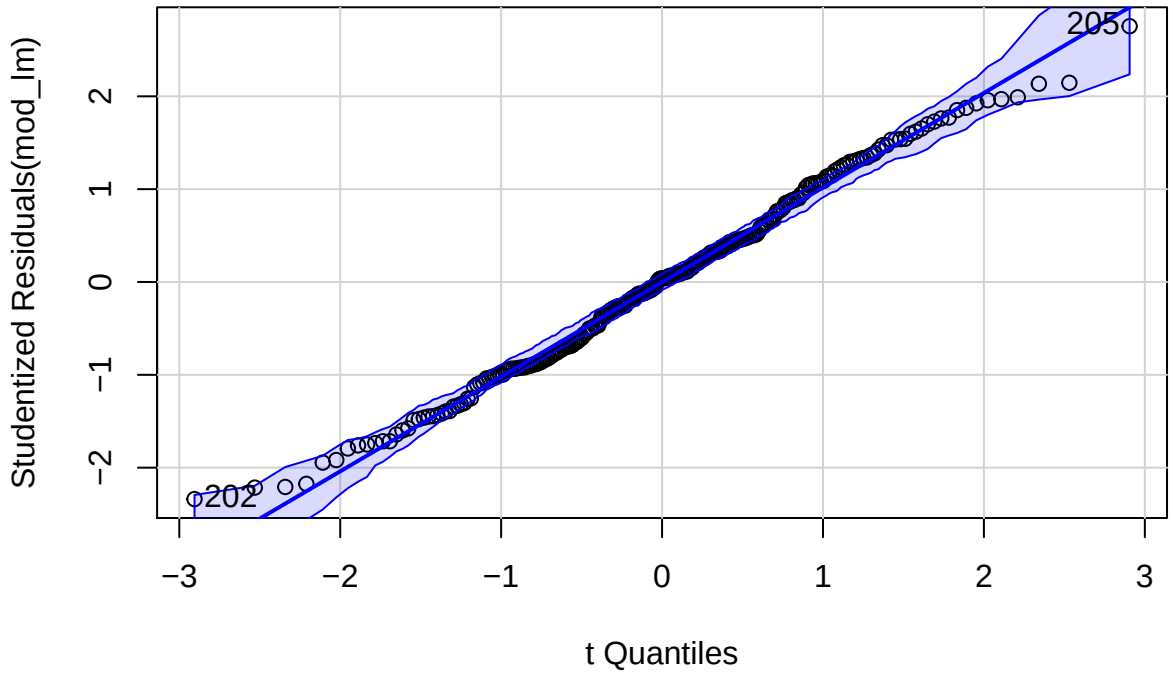
```
  geom_qq_line() +
  theme_bw()
```




```
library(car)
residualPlot(mod_lm)
```

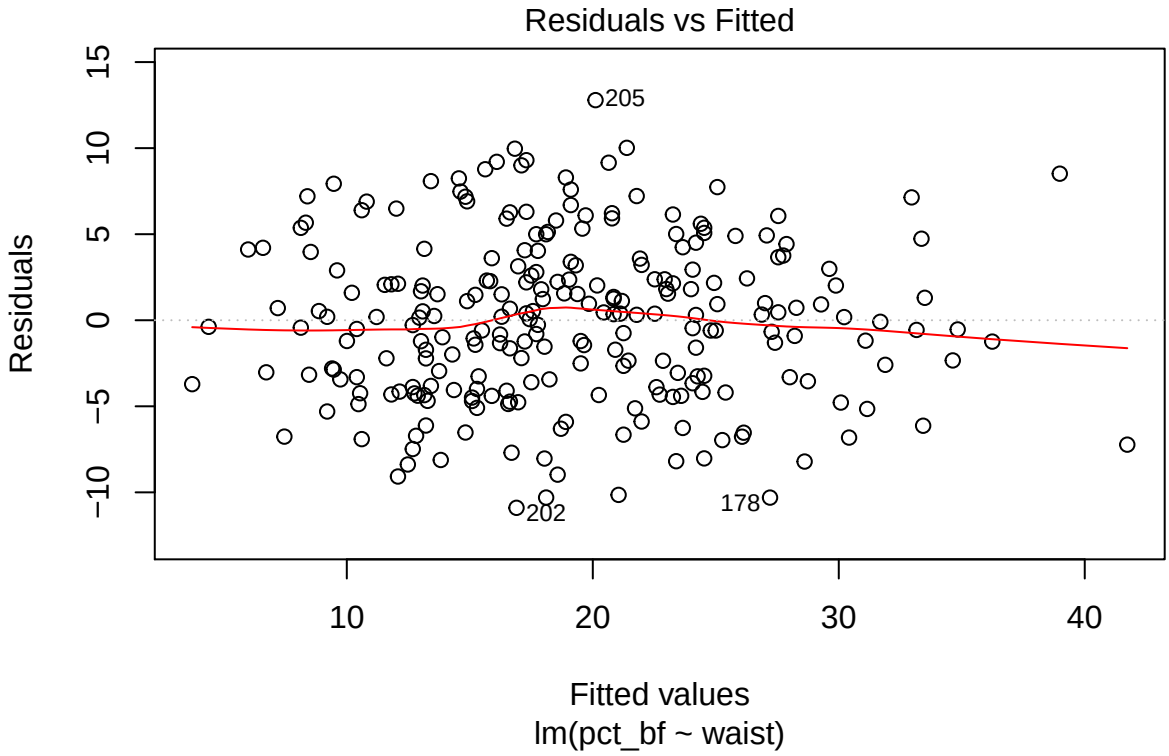


```
qqPlot(mod_lm)
```

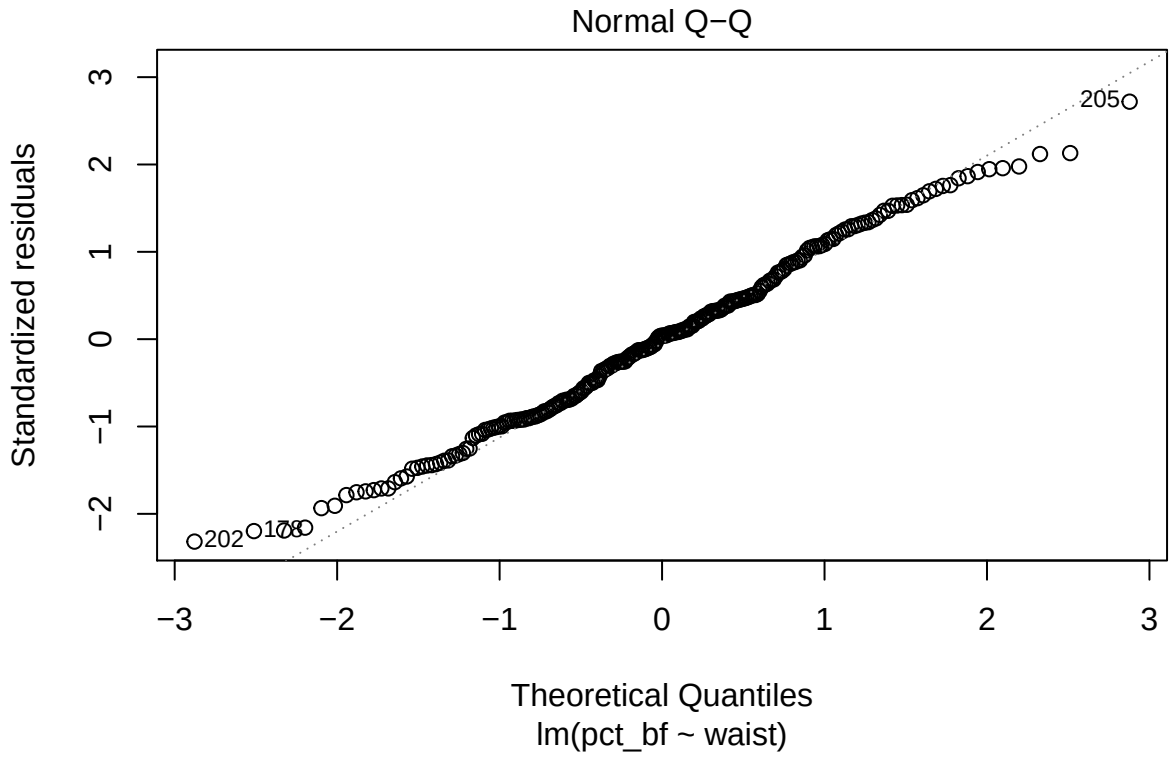



```
[1] 202 205
```

```
# Base R
plot(mod_lm, which = 1)
```




```
plot(mod_lm, which = 2)
```



1.3 Residual Standard Deviation

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

```
summary(mod_lm)
```

```
Call:
lm(formula = pct_bf ~ waist, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8987  -3.6453   0.1864   3.1775  12.7887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.73413    2.71651  -15.73   <2e-16 ***
waist         1.69997    0.07431   22.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.713 on 248 degrees of freedom
Multiple R-squared:  0.6785,    Adjusted R-squared:  0.6772
F-statistic: 523.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

```
summary(mod_lm)$sigma -> s_e
s_e
```

```
[1] 4.71257
```

```r
### By hand now
yhat <- fitted(mod_lm)
y <- bodyfat$pct_bf
se1 <- sqrt(sum((y - yhat)^2)/248)
se1
```

```
[1] 4.71257
```

1.4 Slopes Vary Revisited

```r
# Take 1000 random samples of size 250
set.seed(3)
n <- 1000
b1 <- numeric(n)
for(i in 1:n){
DF <- sample_n(bodyfat, size = 250, replace = TRUE)
mod <- lm(pct_bf ~ waist, data = DF)
b1[i] <- mod$coefficients[2]
}
ep <- quantile(b1, probs = c(0.025, 0.975))
ep
```

```
    2.5%    97.5%
1.550538 1.841983
```

1.5 Multiple Regression Inference

```r
mod_mr <- lm(pct_bf ~ waist + height, data = bodyfat)
summary(mod_mr)
```

```
Call:
lm(formula = pct_bf ~ waist + height, data = bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1692  -3.4133  -0.0977   3.0995   9.9082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.10088    7.68611  -0.403    0.687
waist        1.77309    0.07158  24.770  < 2e-16 ***
height      -0.60154    0.10994  -5.472 1.09e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.46 on 247 degrees of freedom
Multiple R-squared:  0.7132,    Adjusted R-squared:  0.7109
F-statistic: 307.1 on 2 and 247 DF,  p-value: < 2.2e-16
```

1.6 Collinearity

```
coasters <- read.csv("./DATA/Coasters_2015.csv") %>%
  clean_names() %>%
filter(name != "Xcelerator", name != "Tower of Terror")
mod_1 <- lm(duration ~ drop, data = coasters)
mod_2 <- lm(duration ~ drop + speed, data = coasters)
summary(mod_1)
```

```

Call:
lm(formula = duration ~ drop, data = coasters)

Residuals:
    Min      1Q  Median      3Q     Max
-64.869 -18.868  -0.189  17.084  82.062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.48688    9.52406   9.291 1.14e-14 ***
drop         0.38634    0.06279   6.153 2.26e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.27 on 87 degrees of freedom
  (150 observations deleted due to missingness)
Multiple R-squared:  0.3032,    Adjusted R-squared:  0.2952
F-statistic: 37.86 on 1 and 87 DF,  p-value: 2.264e-08
```

```
summary(mod_2)
```

```

Call:
lm(formula = duration ~ drop + speed, data = coasters)

Residuals:
    Min      1Q  Median      3Q     Max
-67.751 -16.483  -3.216  15.370  90.226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3932    34.0567  -0.188  0.85154
drop         -0.1399     0.1917  -0.730  0.46754
speed         2.7030     0.9346   2.892  0.00484 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.94 on 86 degrees of freedom
  (150 observations deleted due to missingness)
Multiple R-squared:  0.365, Adjusted R-squared:  0.3502
F-statistic: 24.71 on 2 and 86 DF,  p-value: 3.314e-09
```

```
# from car ---- Variance Inflation Factor vif
vif(mod_2)
```

```
   drop   speed
10.1066 10.1066
```

1.7 Confidence and Prediction Intervals

```
# Mean Body Fat for male with 38 inch waist - CI
predict(mod_lm, newdata = data.frame(waist = 38), interval = "confidence", level = 0.95)
```

```
      fit     lwr      upr
1 21.8648 21.2291 22.50049
```

```
# Prediction individual with a 38 inch waist
predict(mod_lm, newdata = data.frame(waist = 38), interval = "predict", level = 0.95)
```

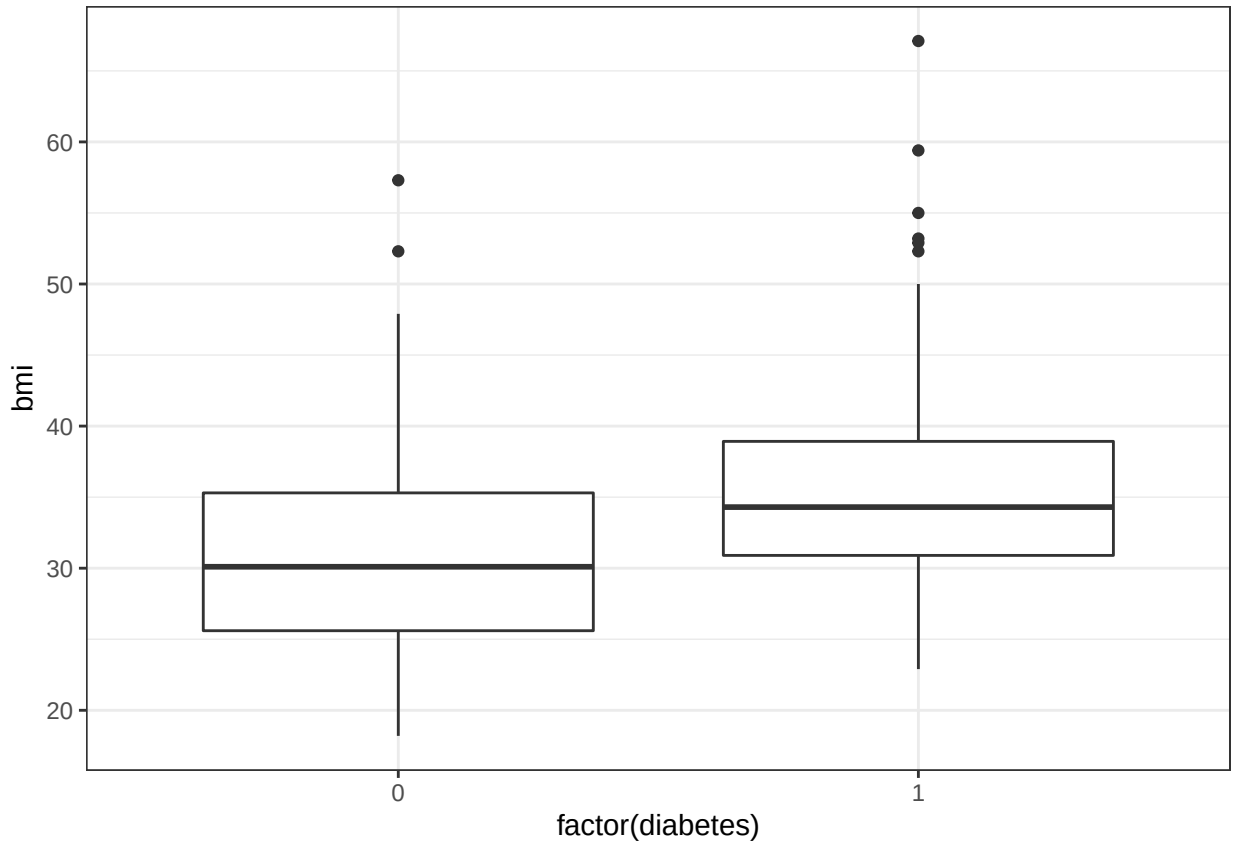
```
      fit      lwr     upr
1 21.8648 12.56129 31.1683
```

1.8 Logistic Regression

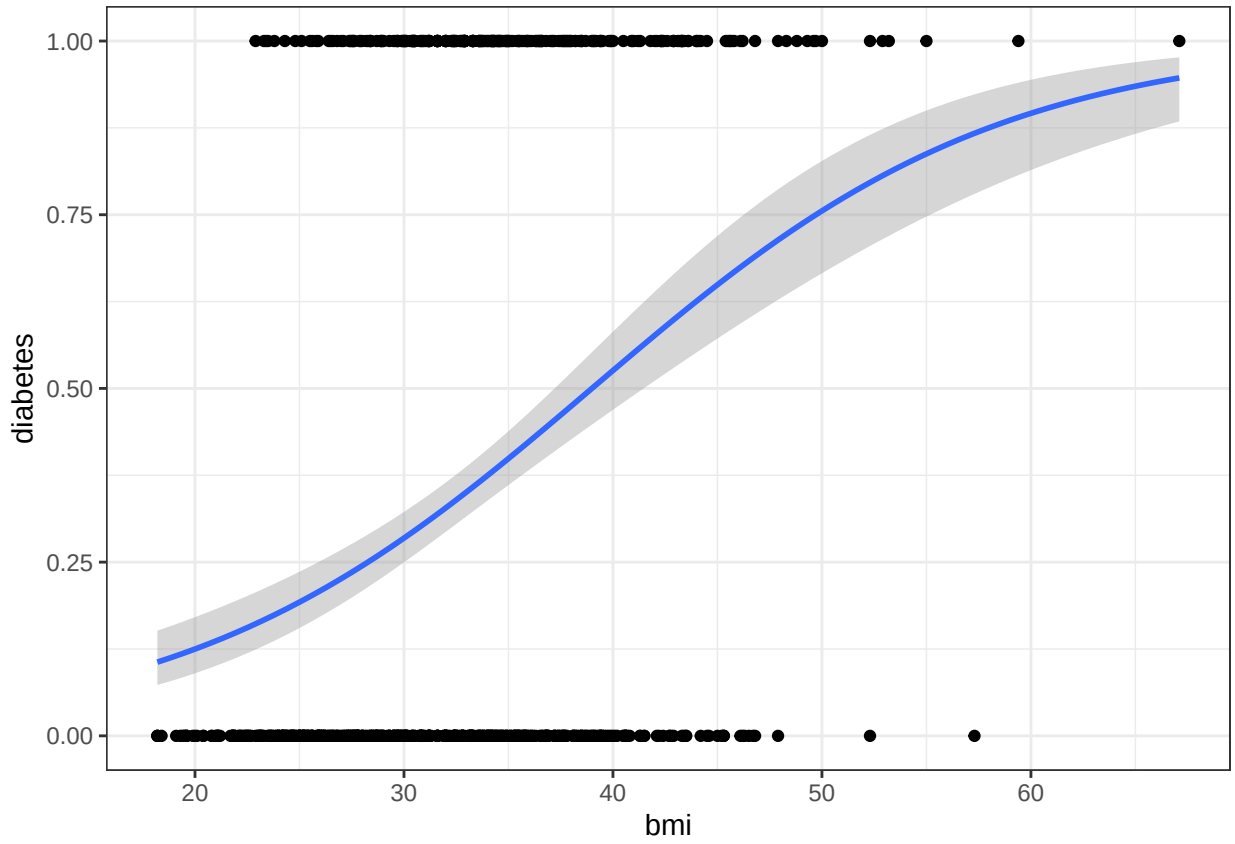
```
pima <- read.csv("./DATA/Pima_indians.csv") %>%
  clean_names() %>%
  filter(bmi != 0)
head(pima)
```

```
  diabetes  bmi age
1        1 33.6  50
2        0 26.6  31
3        1 23.3  32
4        0 28.1  21
5        1 43.1  33
6        0 25.6  30
```

```
###
ggplot(data = pima, aes(x = factor(diabetes), y = bmi)) +
  geom_boxplot() +
  theme_bw()
```




```
###
ggplot(data = pima, aes(x = bmi, y = diabetes)) +
  geom_point() +
  theme_bw() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"))
```




```
mod_lr <- glm(diabetes ~ bmi, data = pima, family = "binomial")
summary(mod_lr)
```

```
Call:
glm(formula = diabetes ~ bmi, family = "binomial", data = pima)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0094  -0.9184  -0.6598   1.2254   1.9107

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.99682    0.42885   -9.32  < 2e-16 ***
bmi          0.10250    0.01261    8.13 4.31e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 981.53  on 756  degrees of freedom
Residual deviance: 904.89  on 755  degrees of freedom
AIC: 908.89

Number of Fisher Scoring iterations: 4
```

```
predict(mod_lr, newdata = data.frame(bmi = 60), type = "response")
```

```
        1
```

```
0.8959635
```

1.9 Problems

```
earnings <- read.csv("./DATA/Graduate_Earnings.csv") %>%
  clean_names()
head(earnings)
```

```
                           school public      location  earn  sat act price
1              Princeton University      0 Princeton, NJ 62800 1510  33 61300
2  University of Michigan-Ann Arbor      1 Ann Arbor, MI 59000 1380  30 28100
3                Harvard University      0 Cambridge, MA 62900 1510  34 64800
4                   Rice University      0   Houston, TX 63700 1460  33 58600
5 University of California-Berkeley      1  Berkeley, CA 60300 1360  30 35700
6    Brigham Young University-Provo      1     Provo, UT 51800 1260  29 18500
  price_with_aid need_fraction merit_aided
1          20600          0.59          NA
2          17300          0.30        0.16
3          16500          0.58          NA
4          22400          0.39        0.11
5          18200          0.51        0.06
6          13400          0.39        0.24
```

```
mod <- lm(earn ~ sat, data = earnings)
summary(mod)
```

```
Call:
lm(formula = earn ~ sat, data = earnings)

Residuals:
     Min       1Q   Median       3Q      Max
-16385.1  -3521.6   -246.4   3191.6  24881.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14468.088   1776.682   8.143 1.75e-15 ***
sat            27.264      1.545  17.646  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5603 on 704 degrees of freedom
Multiple R-squared:  0.3067,    Adjusted R-squared:  0.3057
F-statistic: 311.4 on 1 and 704 DF,  p-value: < 2.2e-16
```

```
confint(mod)
```

```
                  2.5 %      97.5 %
(Intercept) 10979.85867 17956.31734
sat            24.23067    30.29765
```

```
mod2 <- lm(earn ~ sat + need_fraction, data = earnings)
summary(mod2)
```

```
Call:
lm(formula = earn ~ sat + need_fraction, data = earnings)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-16409  -3819   -423   2832  25658

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   23974.208   2327.479  10.301  < 2e-16 ***
sat              23.188      1.658  13.989  < 2e-16 ***
need_fraction -8500.746   1328.938  -6.397 2.94e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5409 on 684 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.3555,    Adjusted R-squared:  0.3536
F-statistic: 188.6 on 2 and 684 DF,  p-value: < 2.2e-16
```

```
mod3 <- lm(earn ~ sat + need_fraction + act, data = earnings)
summary(mod3)
```

```
Call:
lm(formula = earn ~ sat + need_fraction + act, data = earnings)

Residuals:
     Min       1Q   Median       3Q      Max
-15653.3  -3633.6   -443.2   2822.6  25111.7

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   25162.762   2340.317  10.752  < 2e-16 ***
sat              10.112      4.355   2.322  0.02053 *
need_fraction -8564.027   1319.930  -6.488 1.67e-10 ***
act             551.243    169.957   3.243  0.00124 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5372 on 683 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.3653,    Adjusted R-squared:  0.3625
F-statistic:   131 on 3 and 683 DF,  p-value: < 2.2e-16
```

```
predict(mod3, newdata = data.frame(sat = 1200, need_fraction = 0.5, act = 26), interval = "confidence")
```

```
       fit      lwr      upr
1 47347.06 46881.35 47812.78
```

```
predict(mod3, newdata = data.frame(sat = 1200, need_fraction = 0.5, act = 26), interval = "predict")
```

```
       fit     lwr      upr
1 47347.06 36789.2 57904.93
```