Linear regression

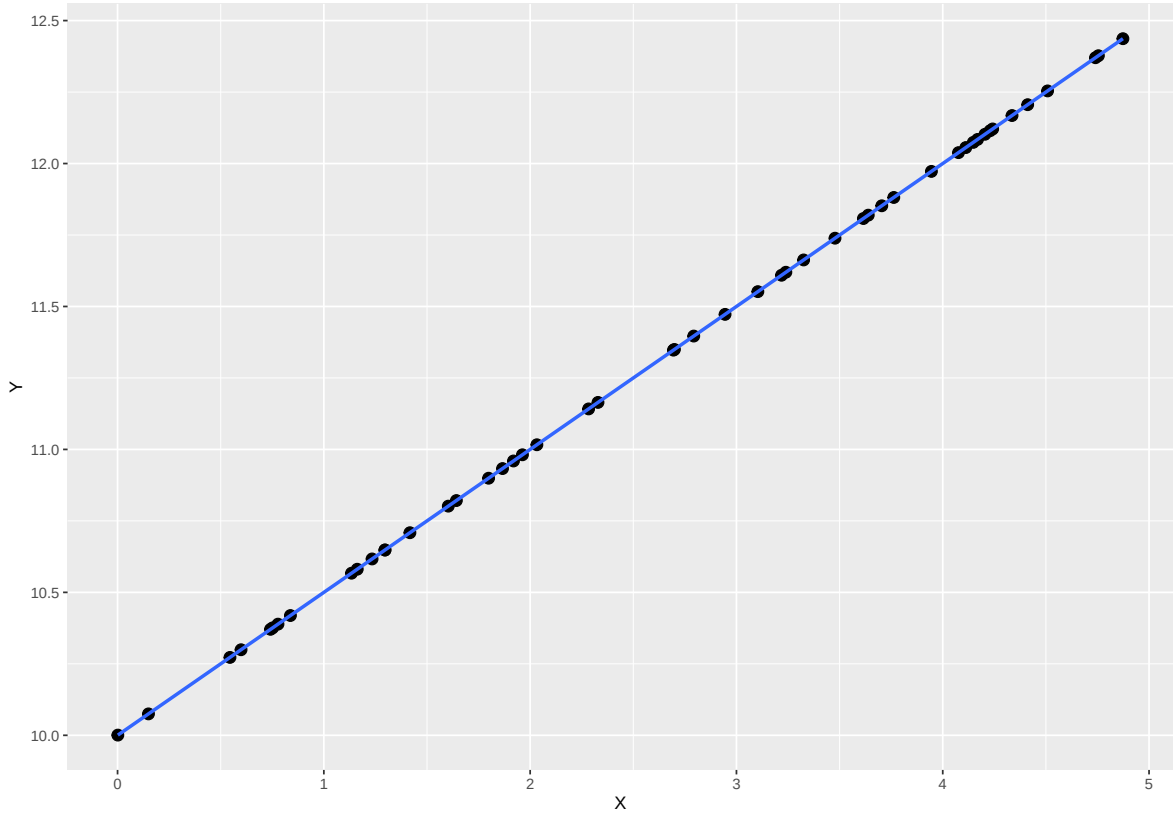
E. Pastucha

October 2024

Linear regression

A method to describe relationship within data.

$$y = \beta_0 + \beta_1 x + \varepsilon$$



Linear regression

A method to describe relationship within data.

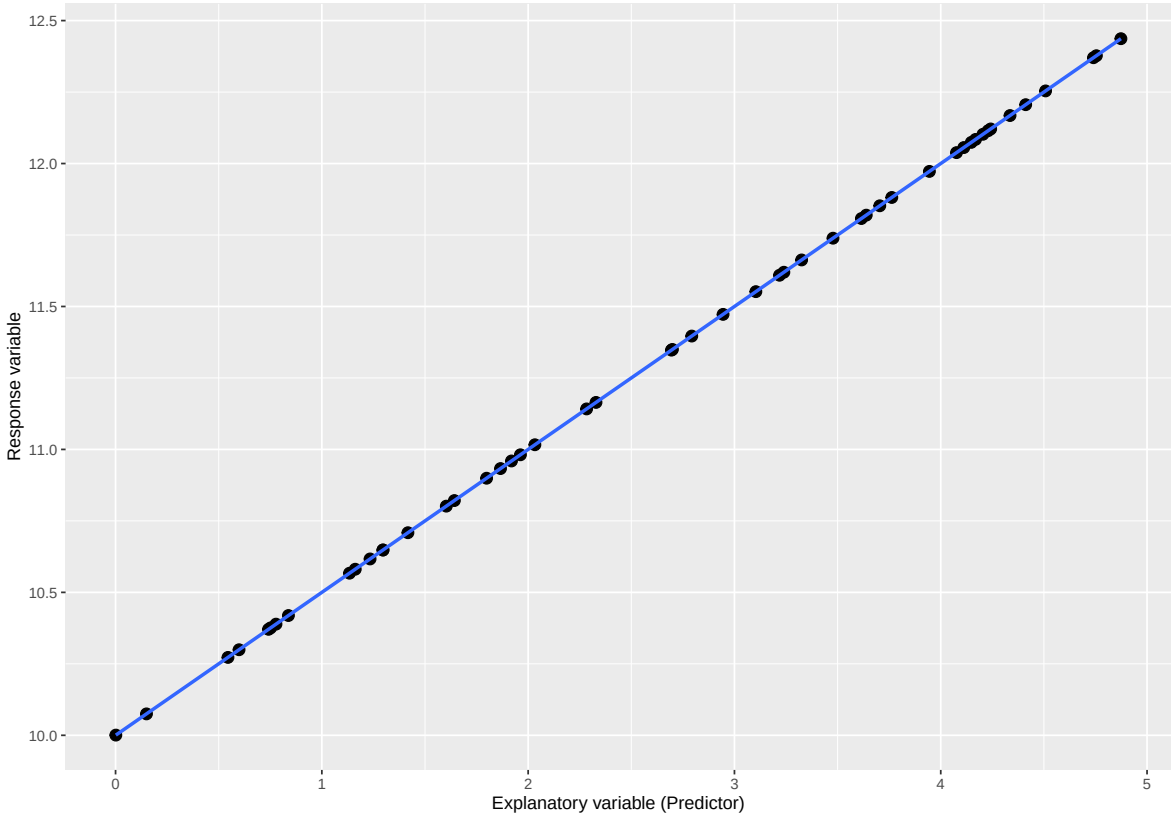
$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$y = b_0 + b_1 x + \varepsilon$$

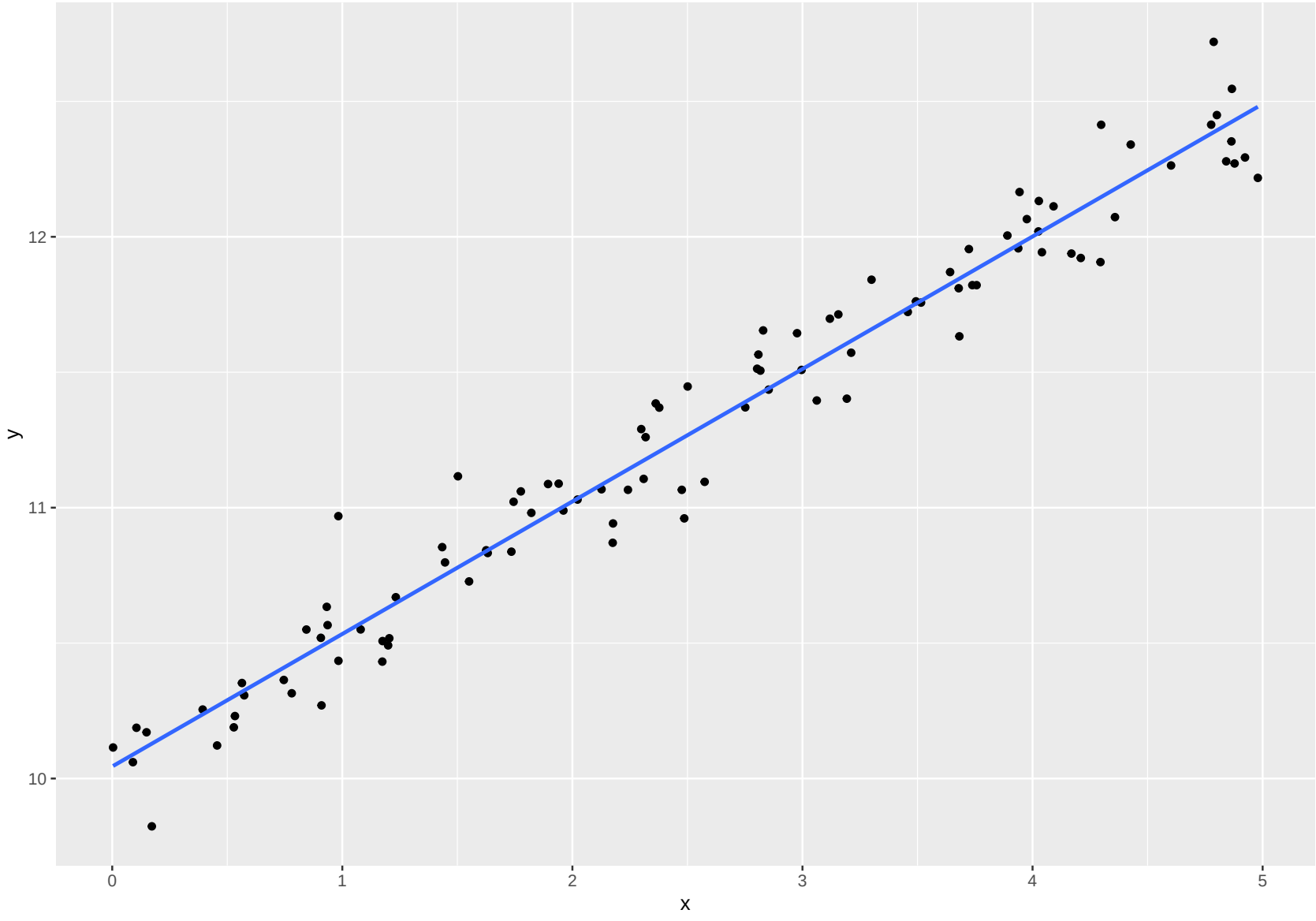
Linear regression

Explanatory (predictor) and response variables

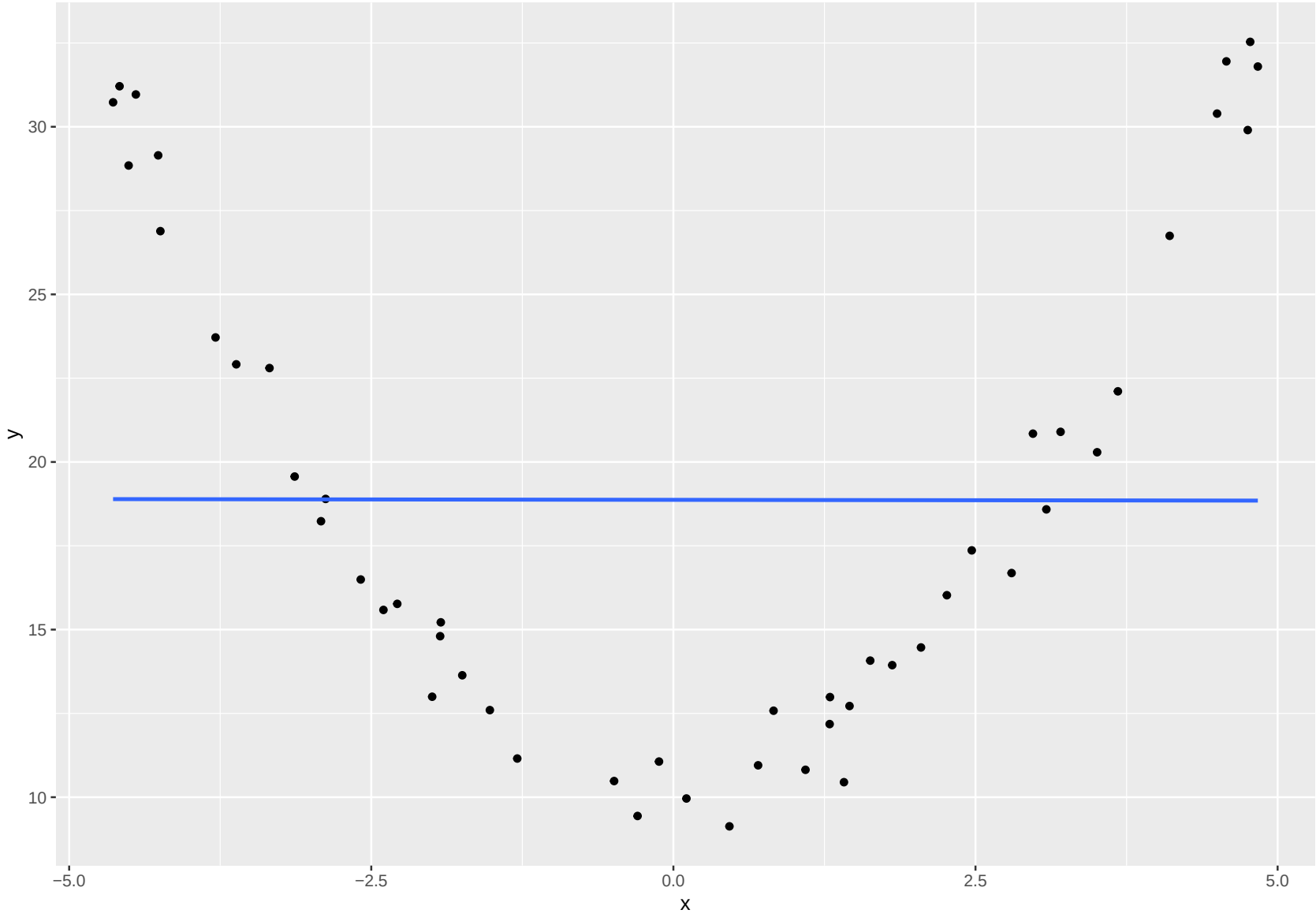
$$y = \beta_0 + \beta_1 x + \varepsilon$$



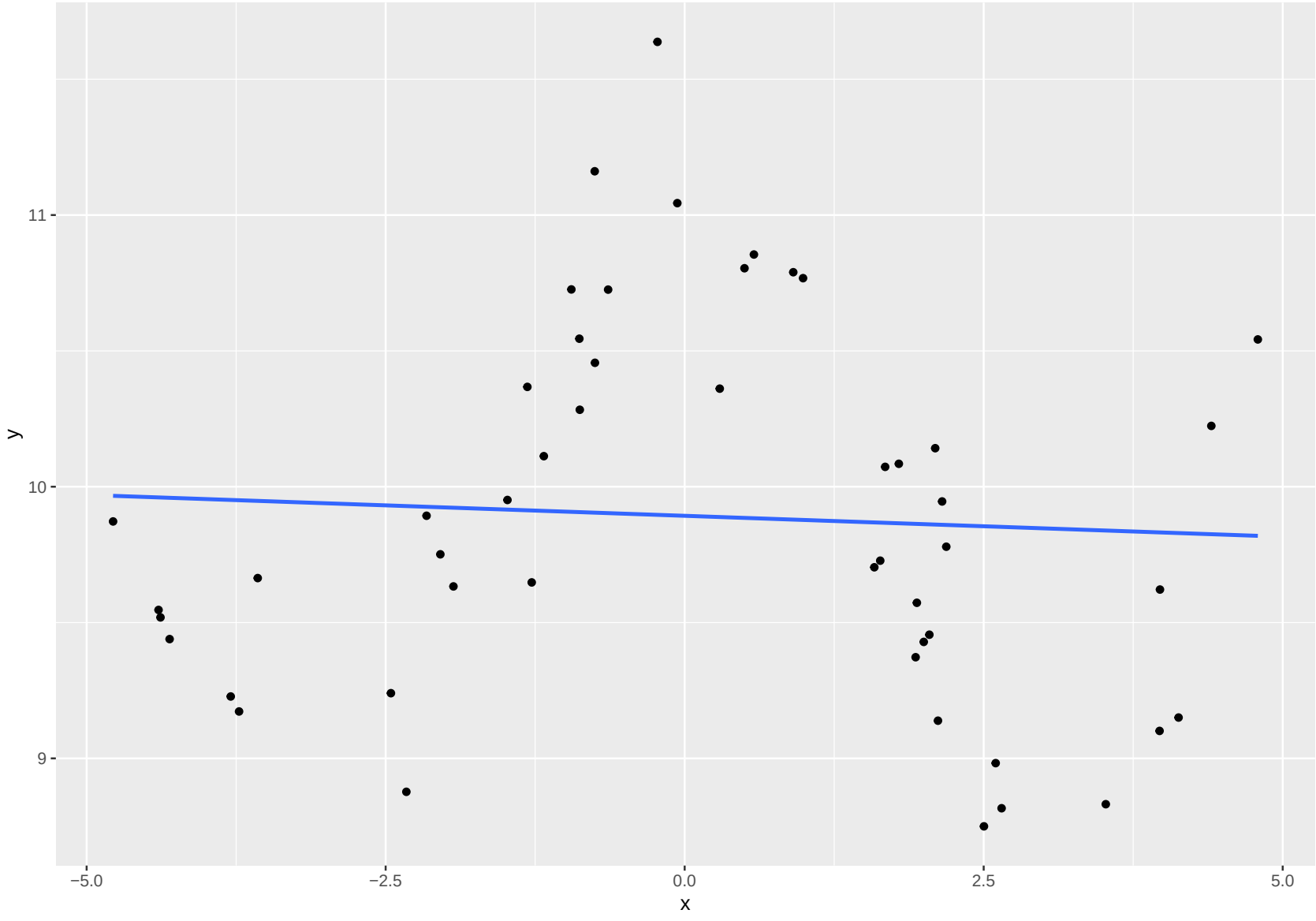
Linear regression



Linear regression



Linear regression



Linear regression

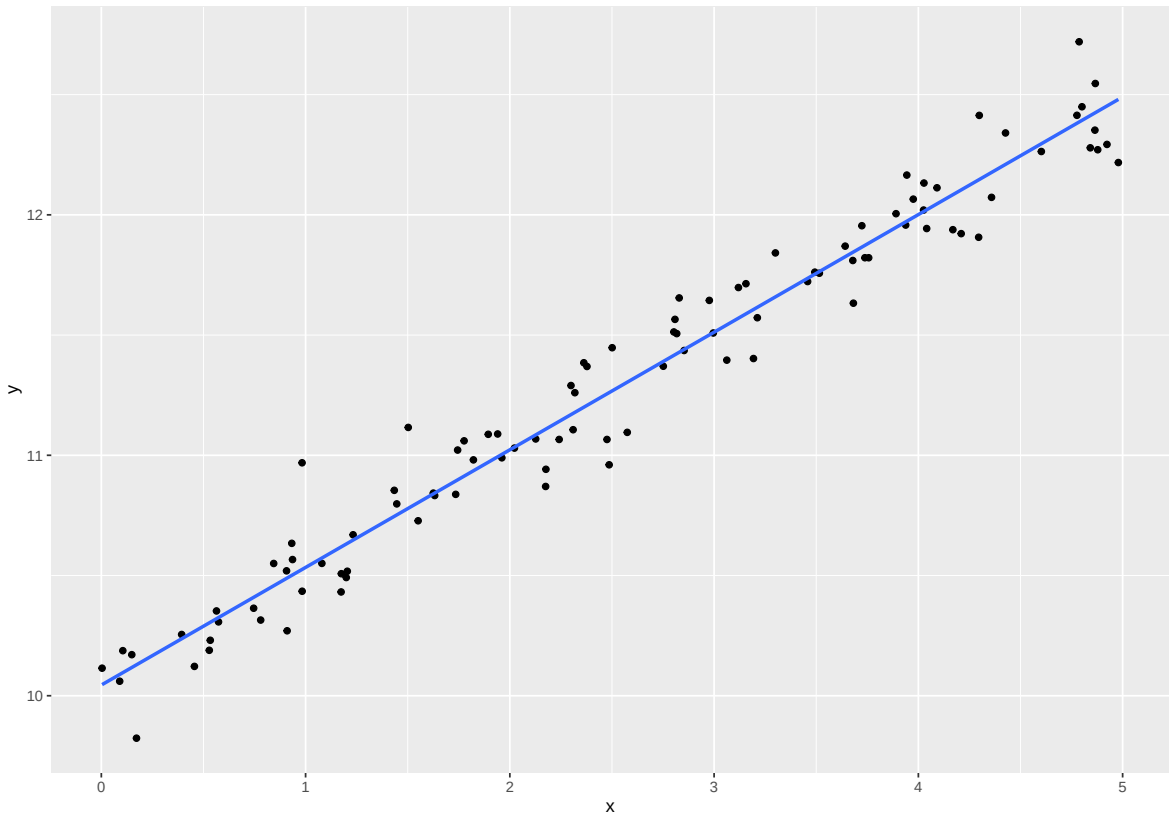
Correlation R

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

The streangth of a linear relationship.

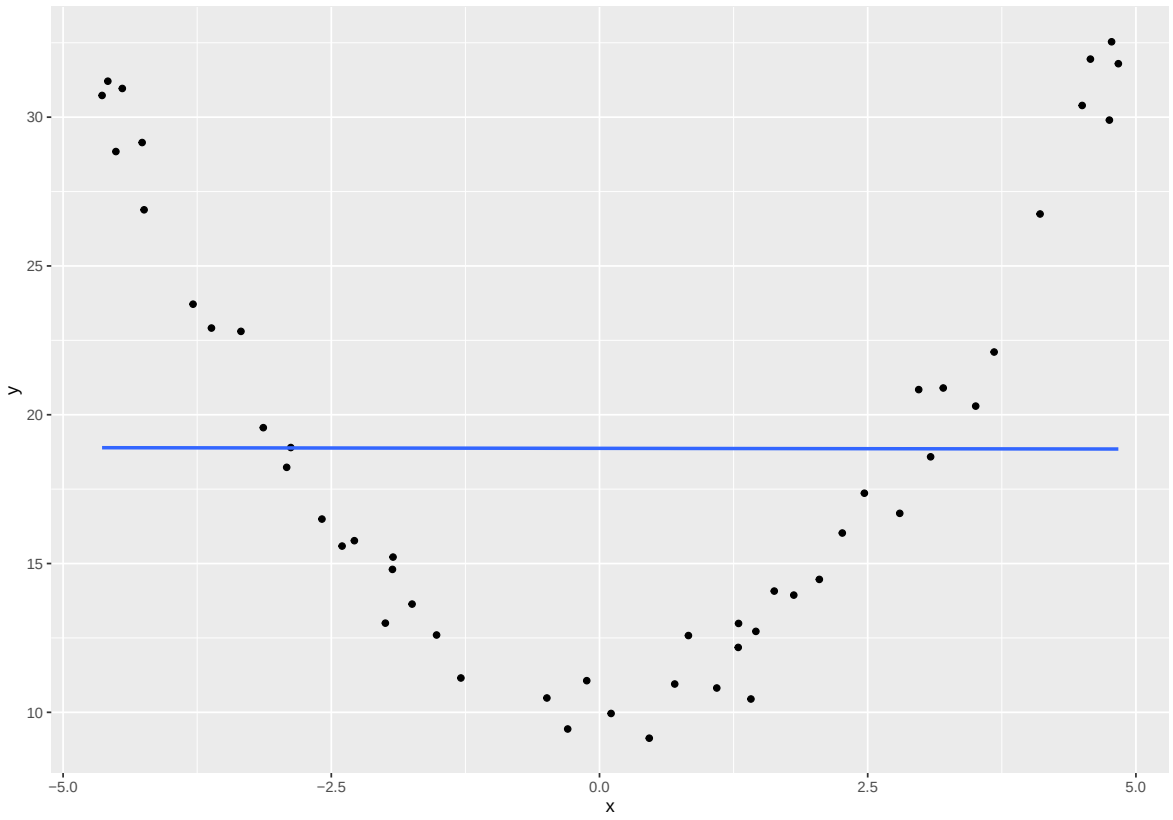
Linear regression

$R = 0.9803414$



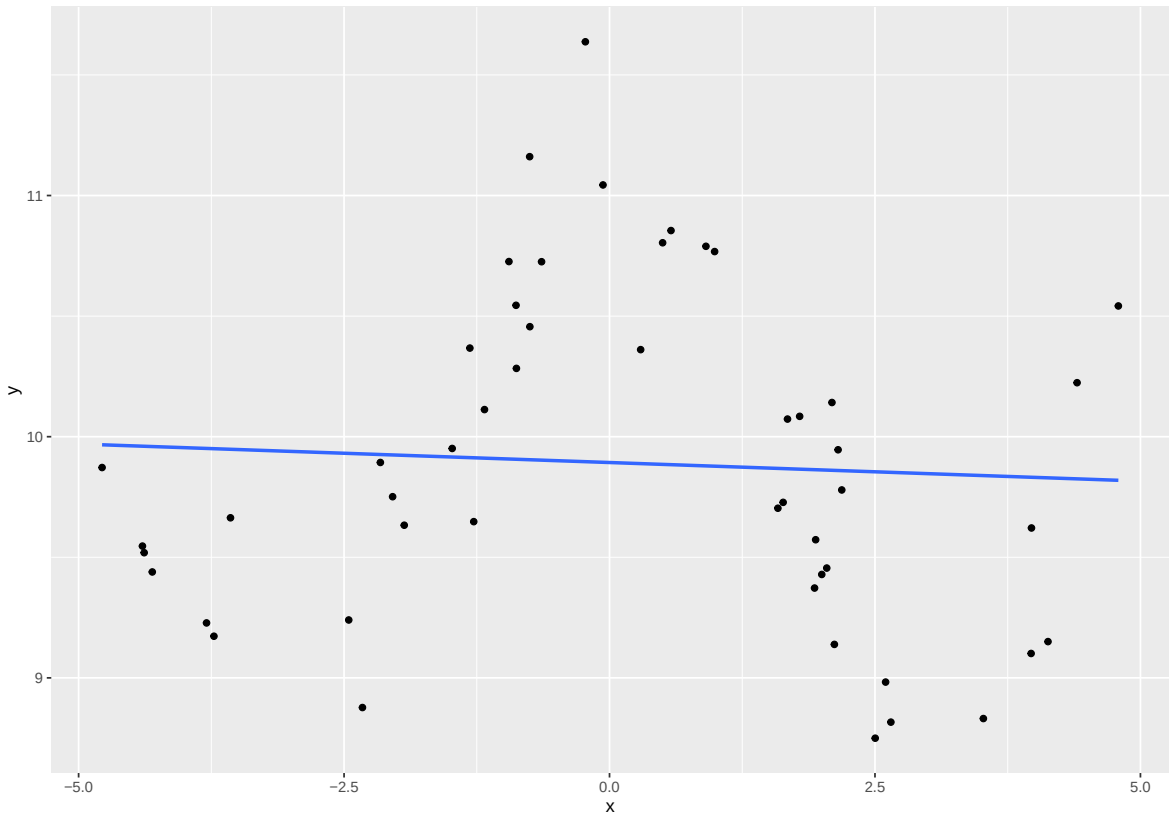
Linear regression

$R = -0.0019065$



Linear regression

$R = -0.0579671$



Linear regression

R squared - R^2

Describes the amount of variation in the response variable that is explaines by the least square fitted line.

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the response variable}} = 1 - \frac{s^2_{residuals}}{s^2_{response}}$$

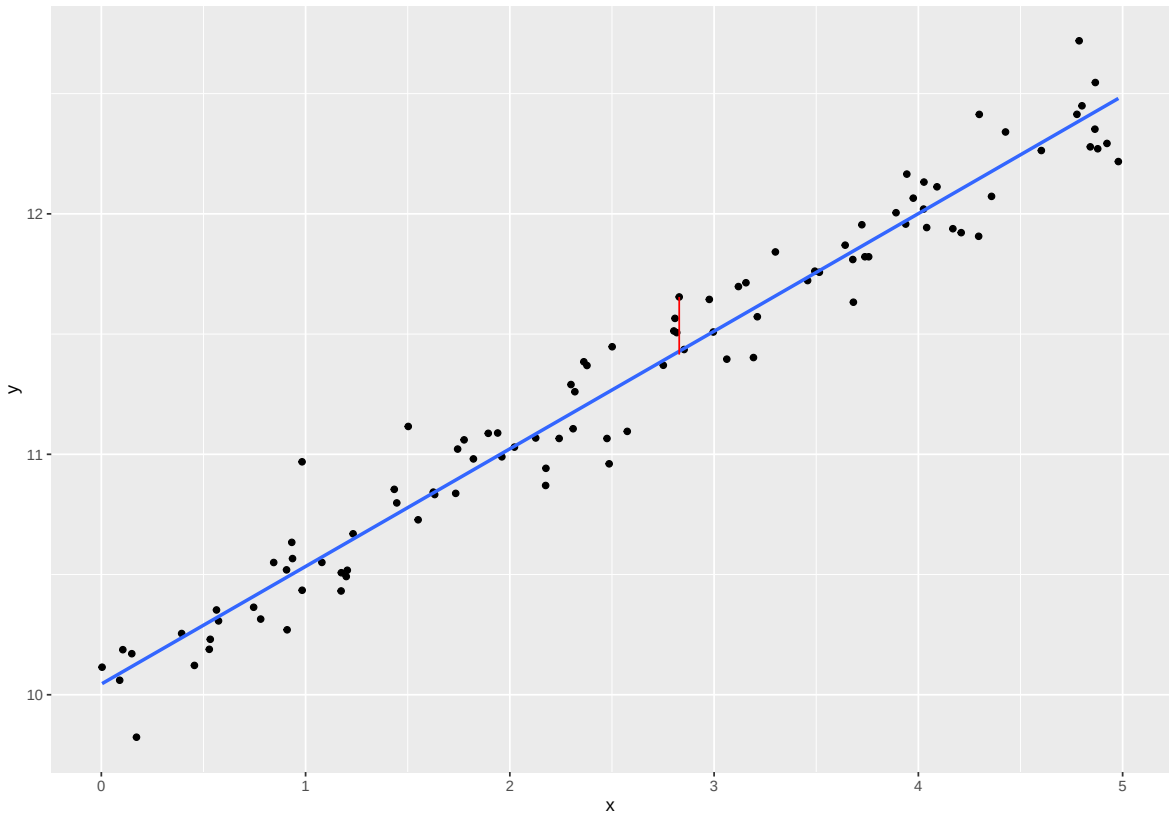
For linear, messy data $R^2 = 0.9610692$

For messy, parabolic data $R^2 = 3.6347392 \times 10^{-6}$

For messy, cosine data $R^2 = 0.0033602$

Linear regression

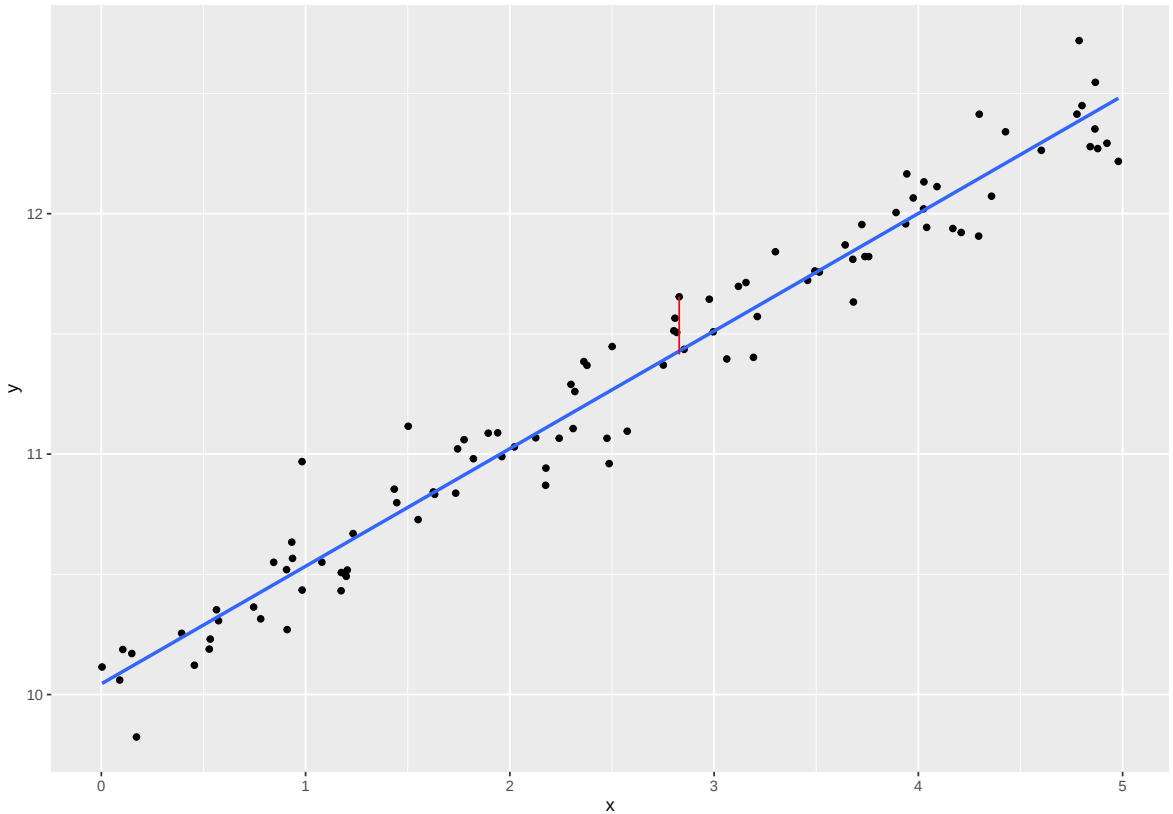
Residuals $e_i = y_i - \hat{y}_i$



Linear regression

Least squares regression

$$e_1^2 + e_2^2 + \cdots + e_n^2$$



Linear regression

How to calculate the fit? $y = b_0 + b_1 x + \varepsilon$

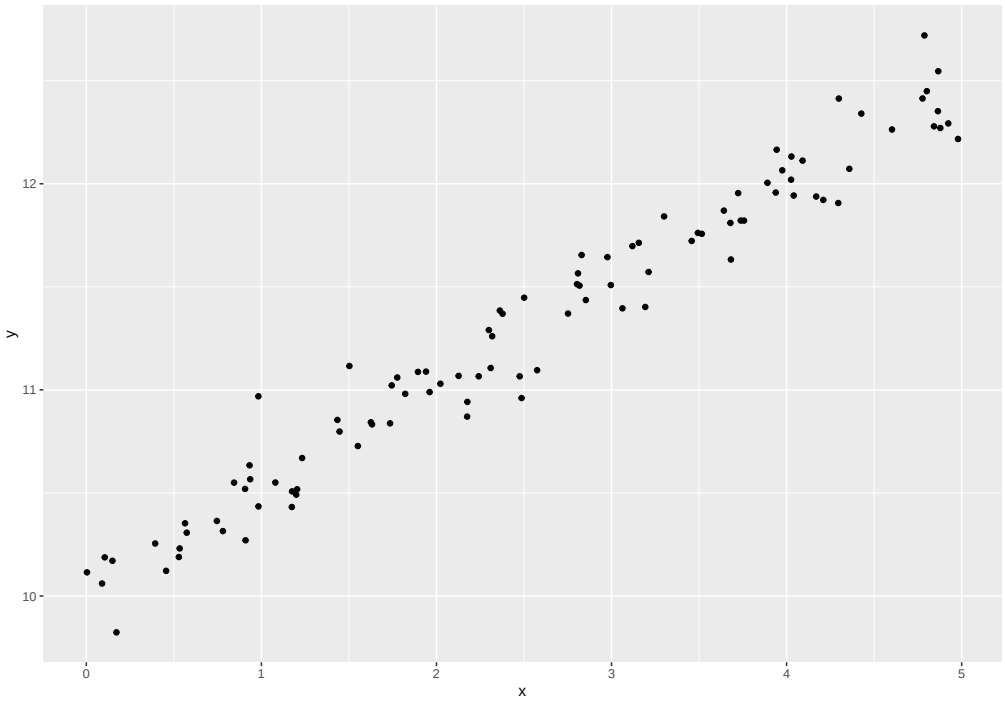
$$b_1 = \frac{s_y}{s_x} R$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Linear regression

For linear, messy data:

$10.044475, 0.4891462$



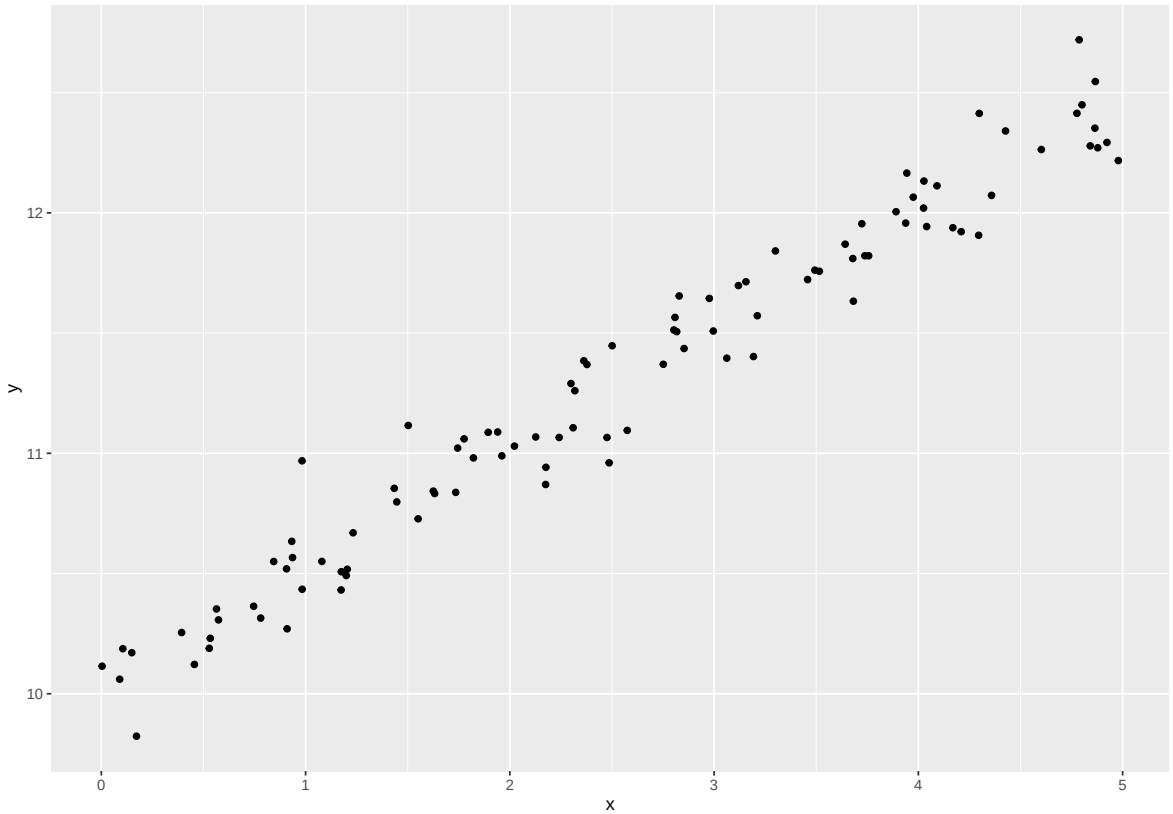
Linear regression

Conditions:

- linearity
- nearly normal residuals
- constant variability
- independent observations

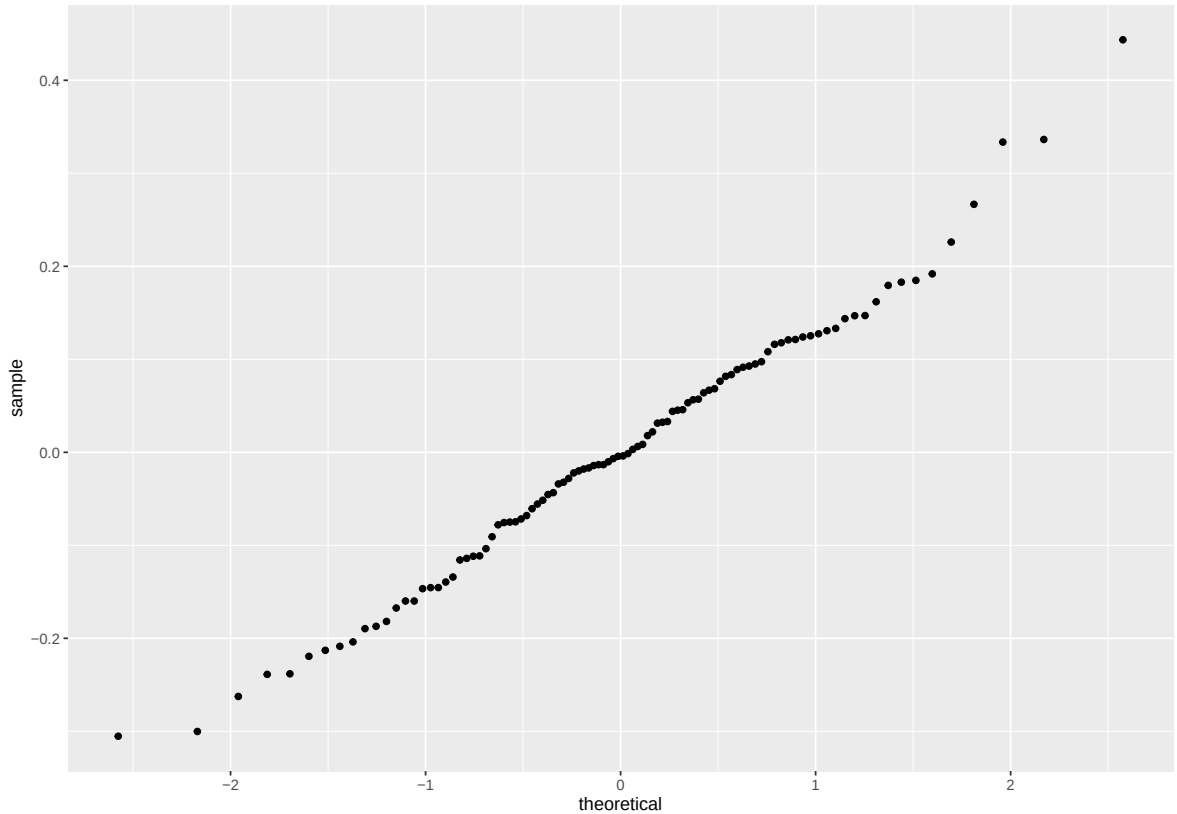
Linear regression

Conditions check - linearity:



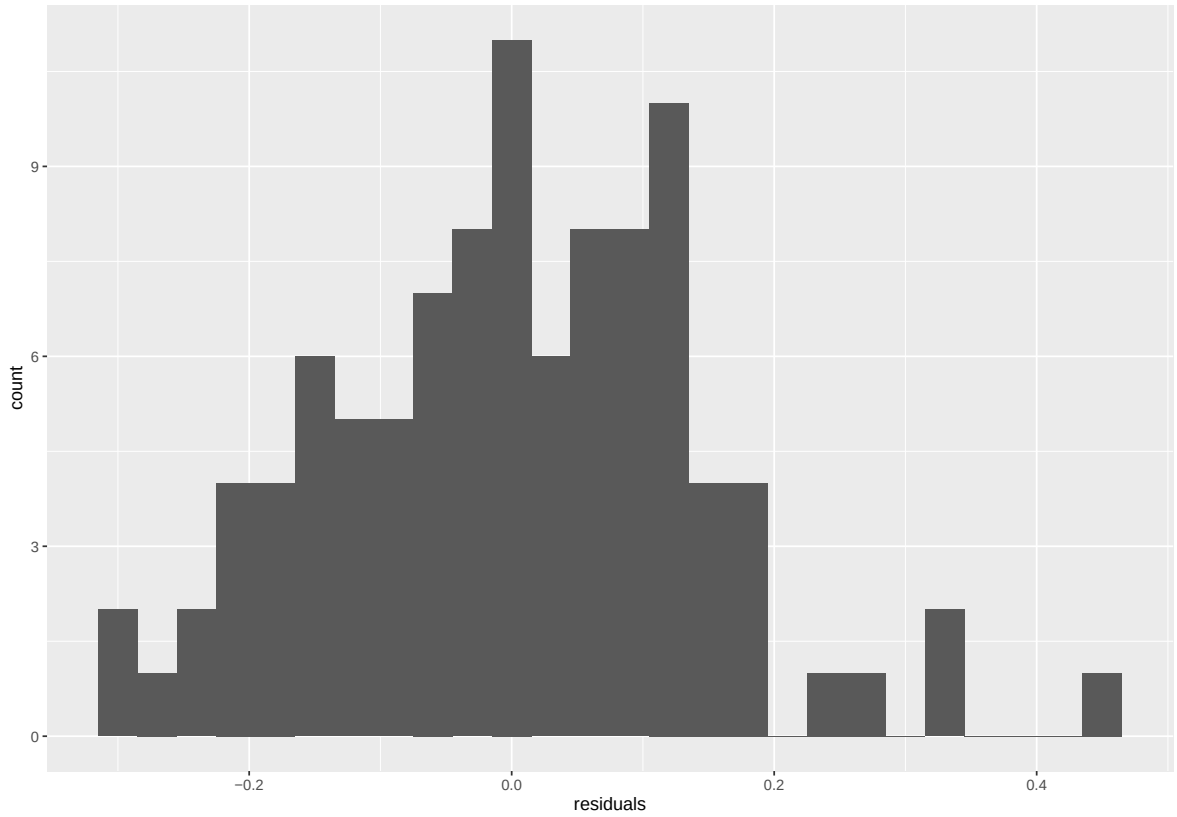
Linear regression

Conditions check - nearly normal residuals:



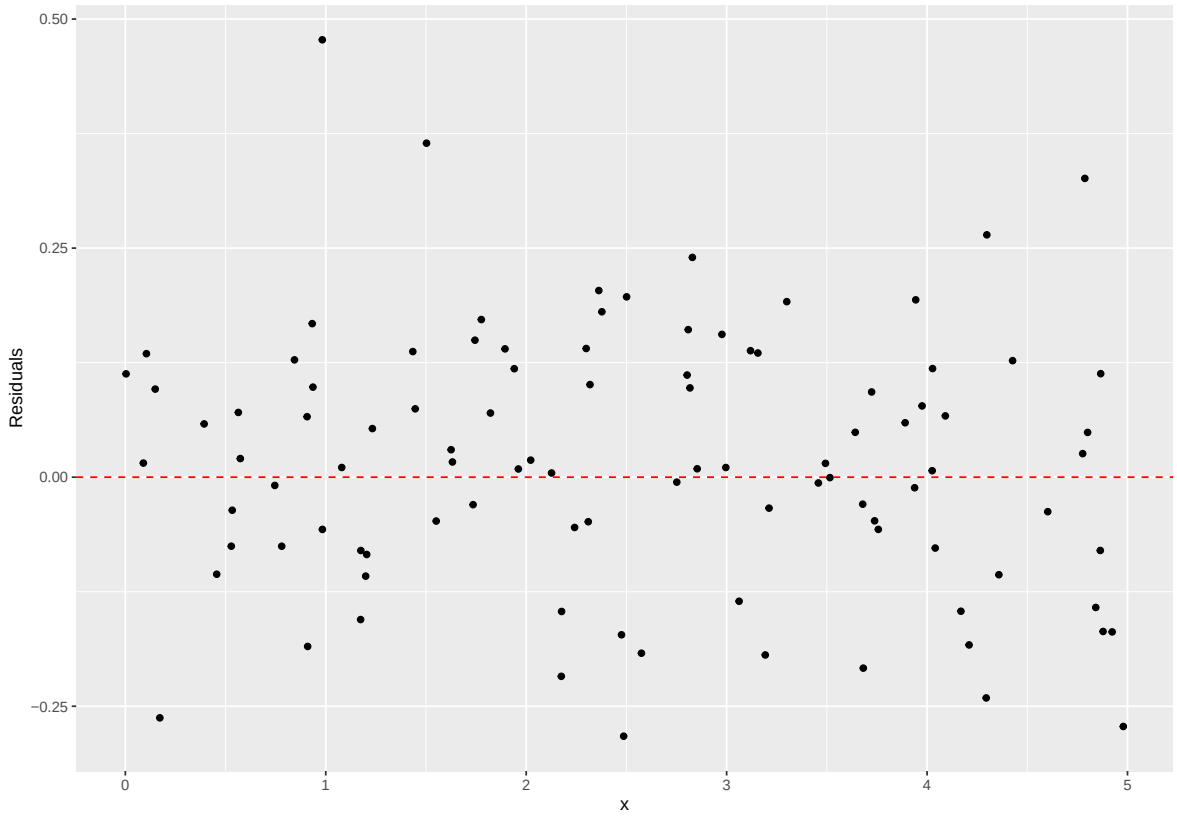
Linear regression

Conditions check - nearly normal residuals:

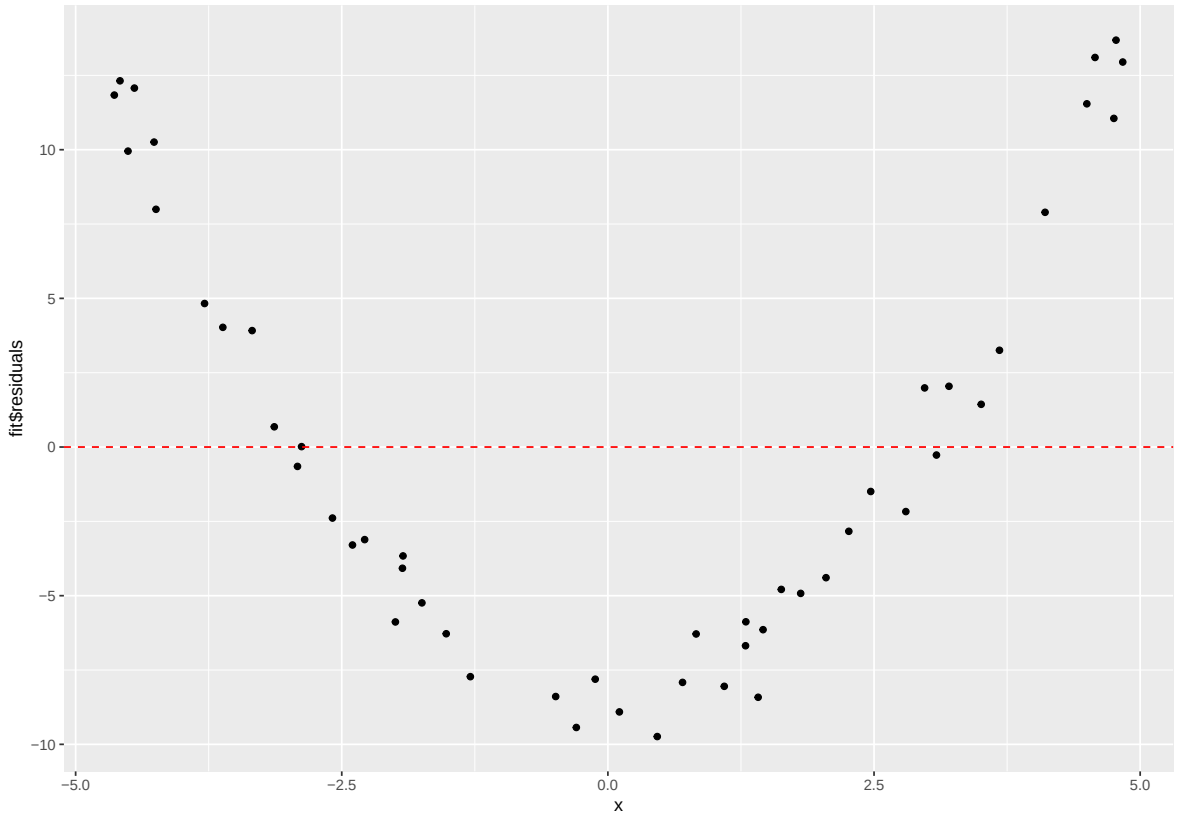


Linear regression

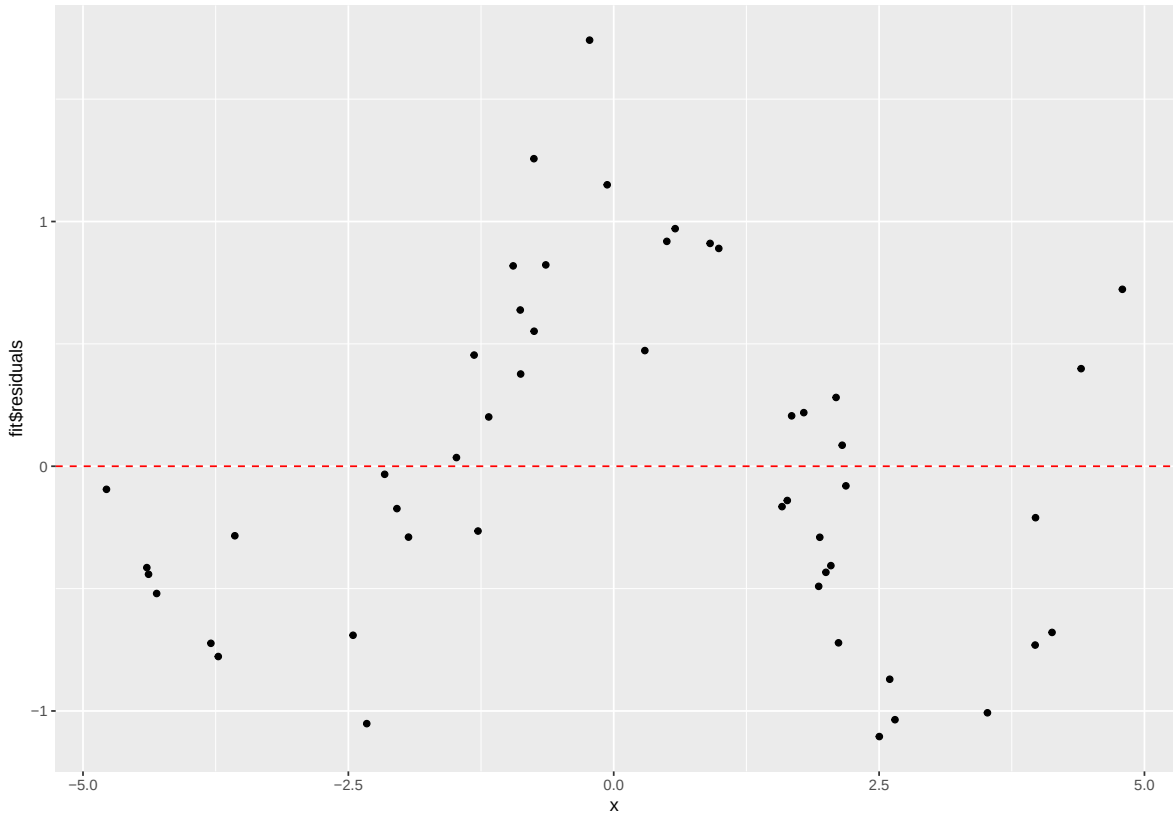
Conditions check - consant variability:



Linear regression



Linear regression



Linear regression

Conditions check - independence:

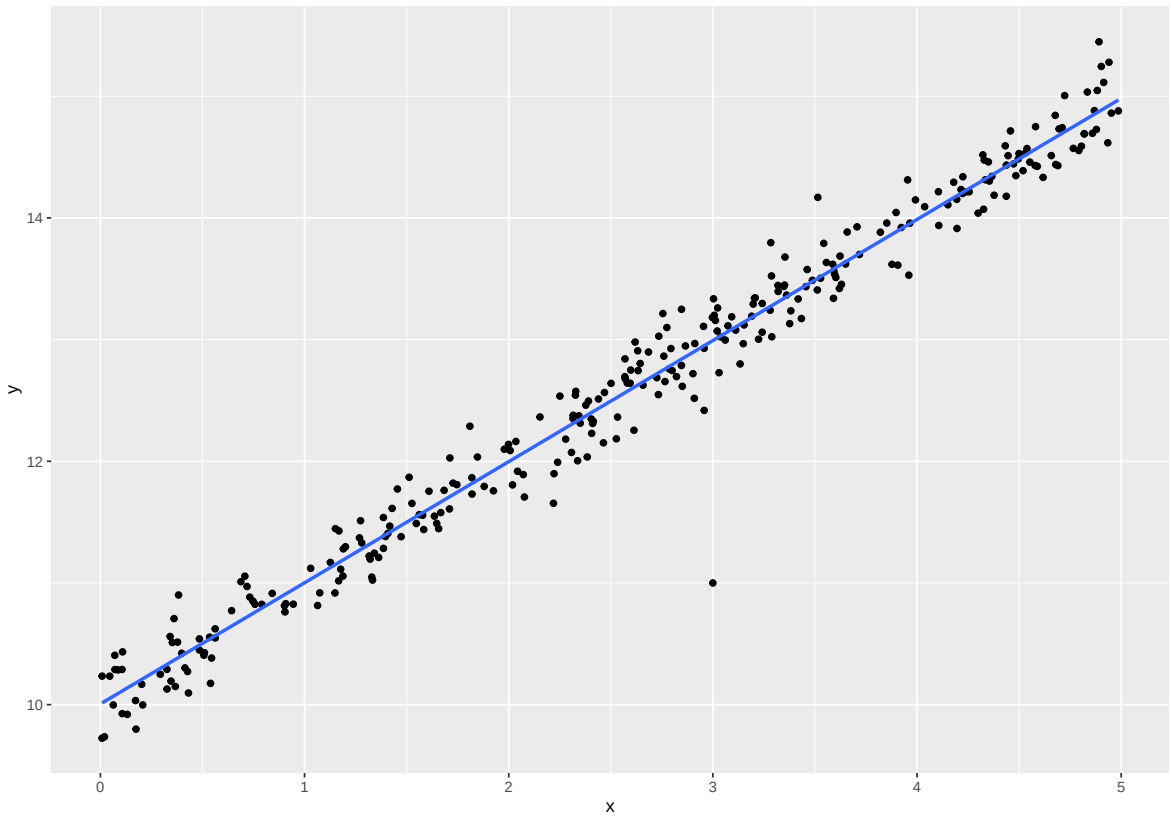
Data survey!

Linear regression

What should you pay attention to?

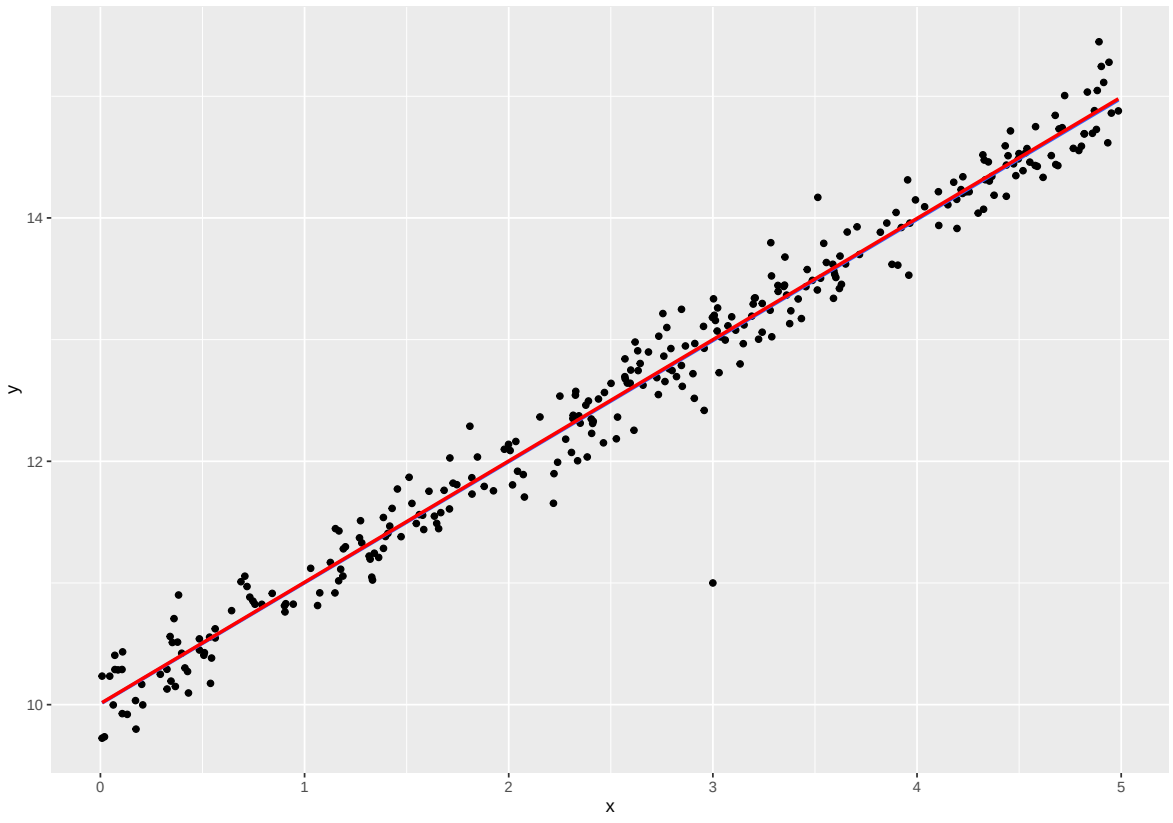
Linear regression

Outliers



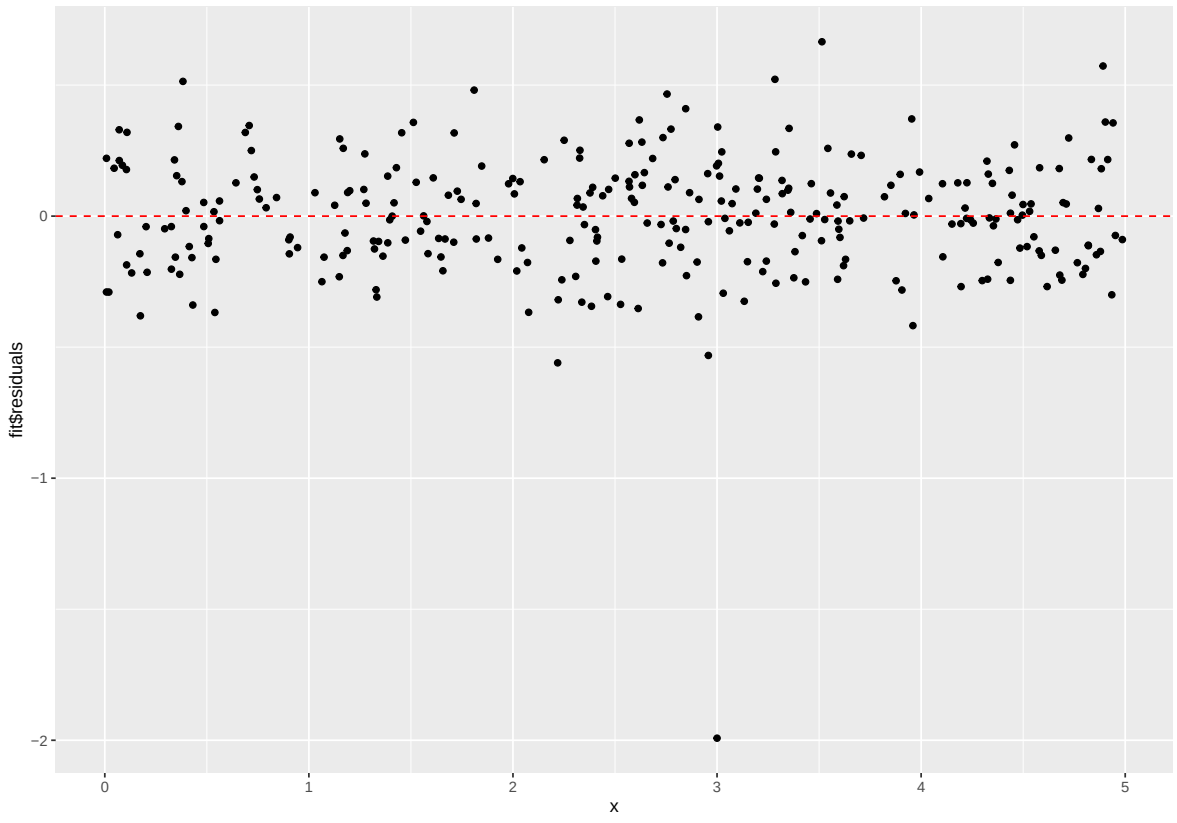
Linear regression

Outliers



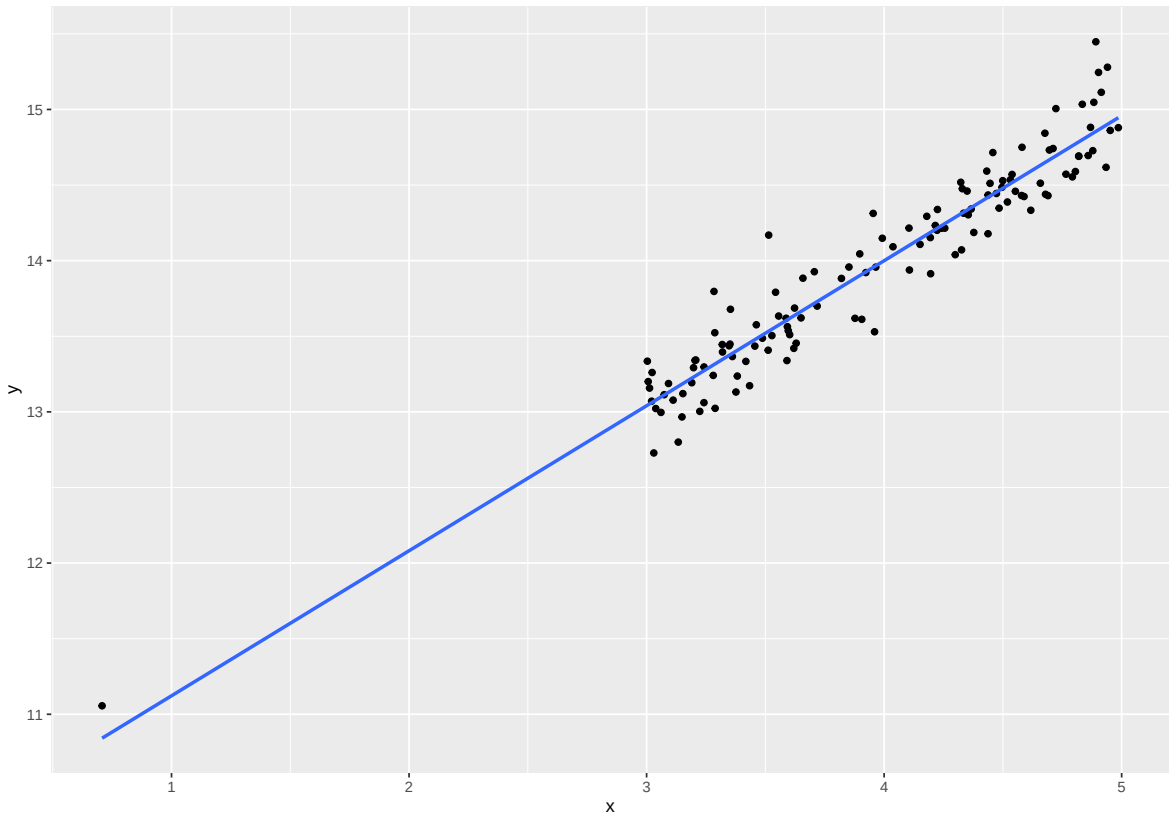
Linear regression

Outliers



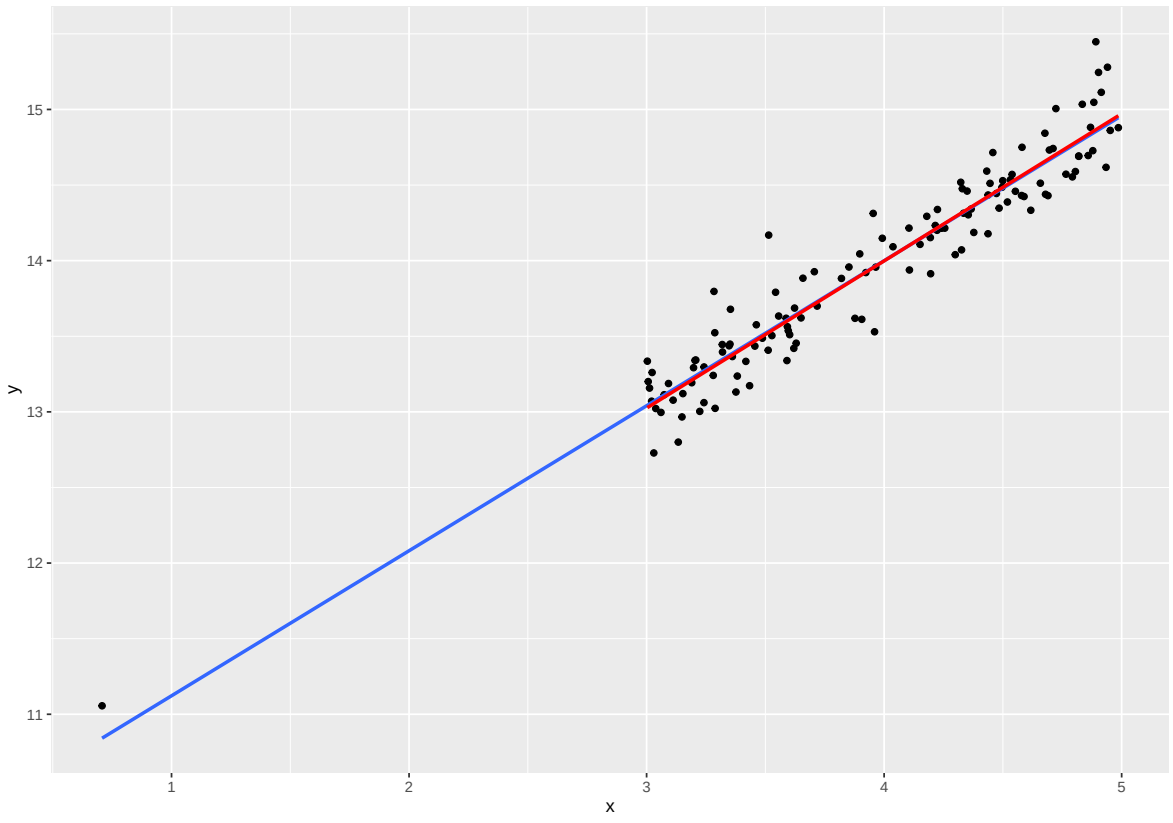
Linear regression

Outliers



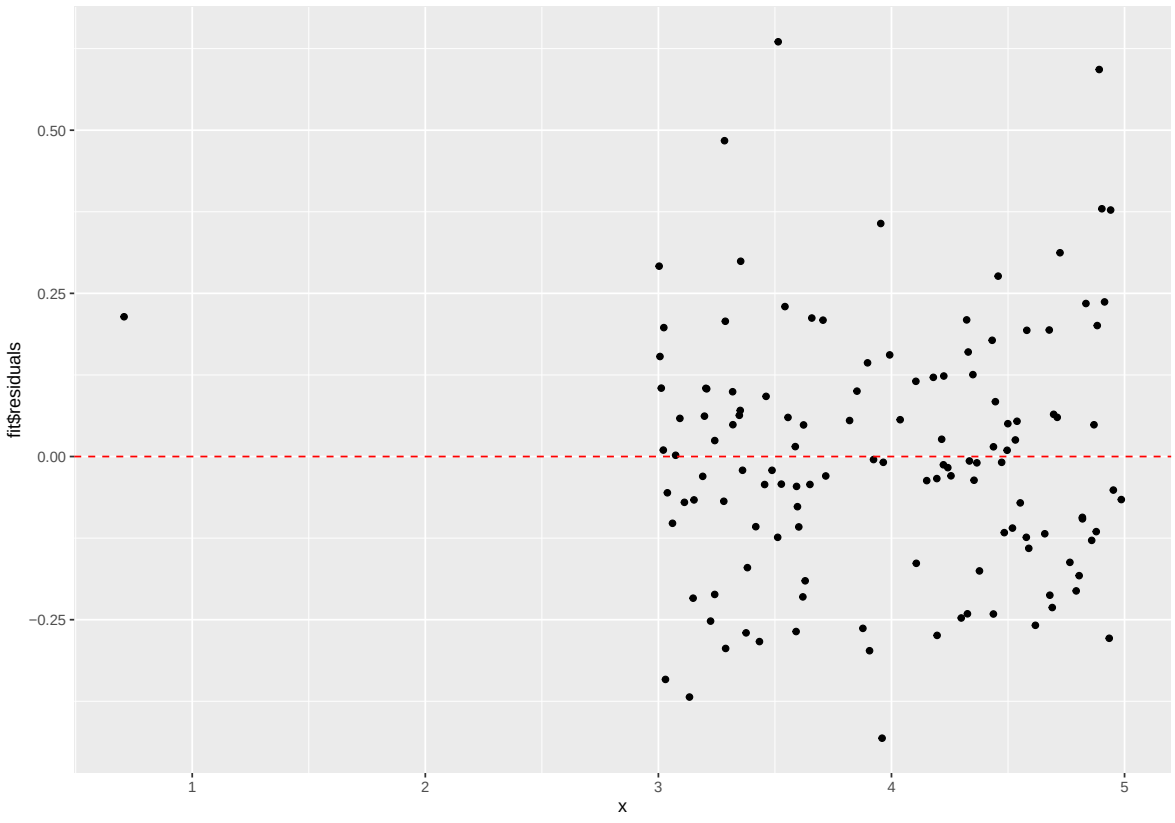
Linear regression

Outliers



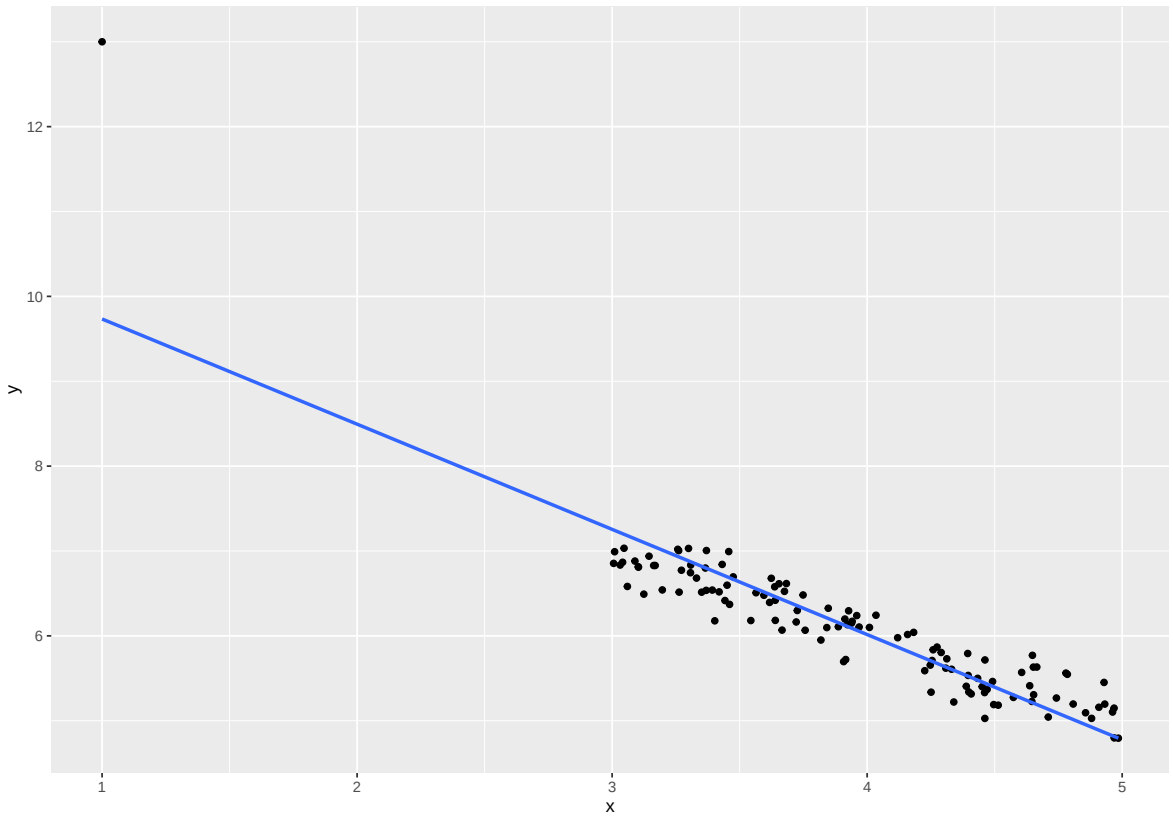
Linear regression

Outliers



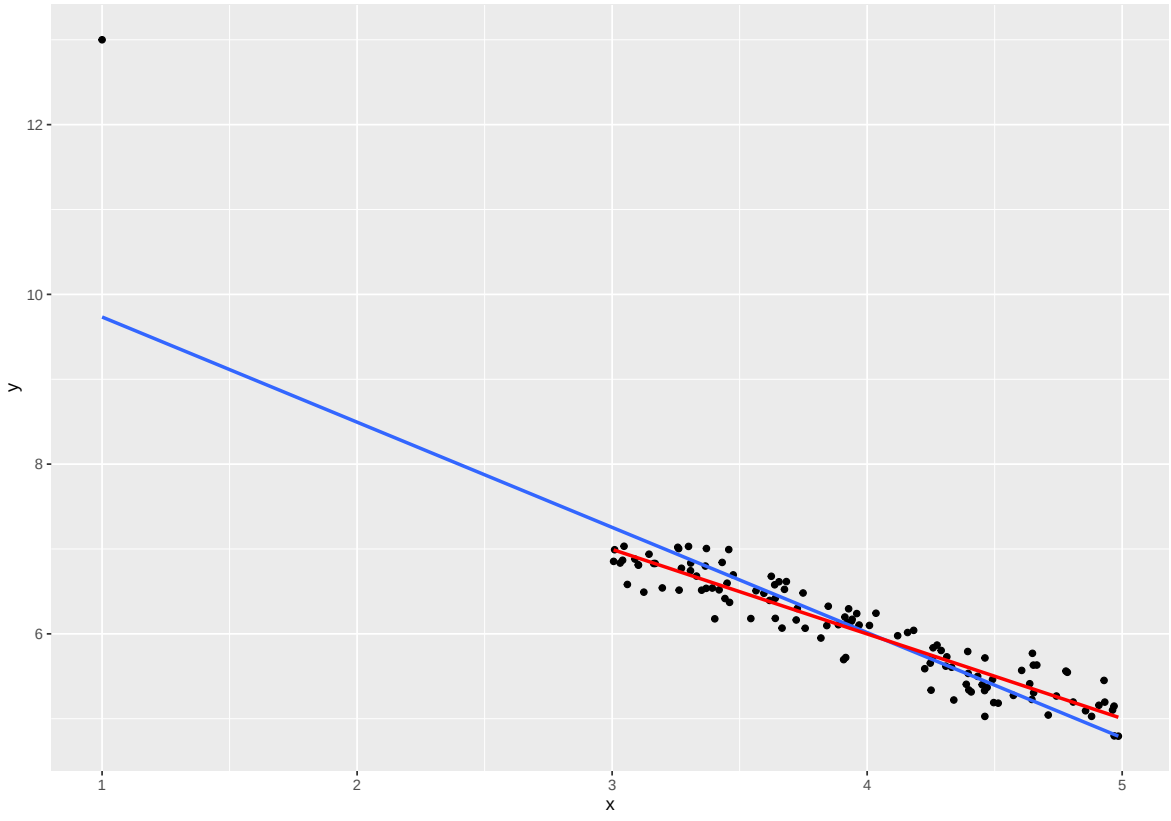
Linear regression

Outliers



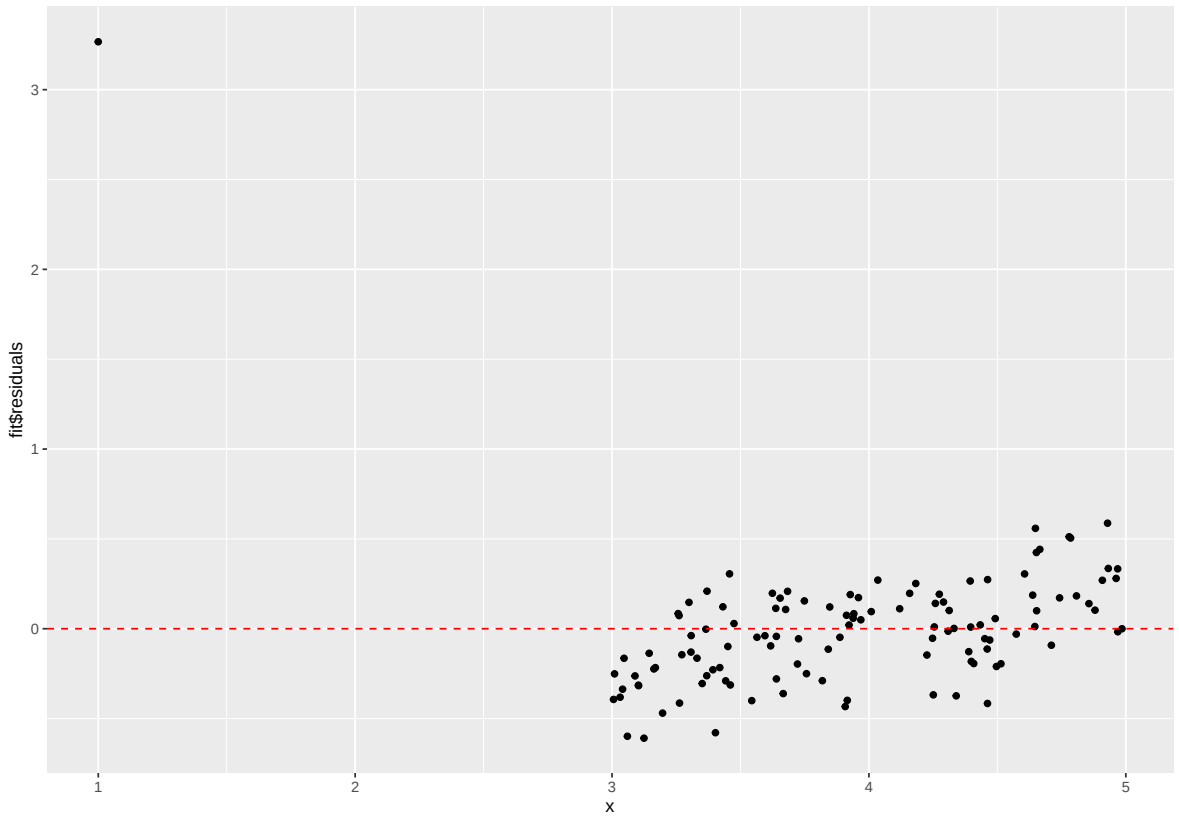
Linear regression

Outliers



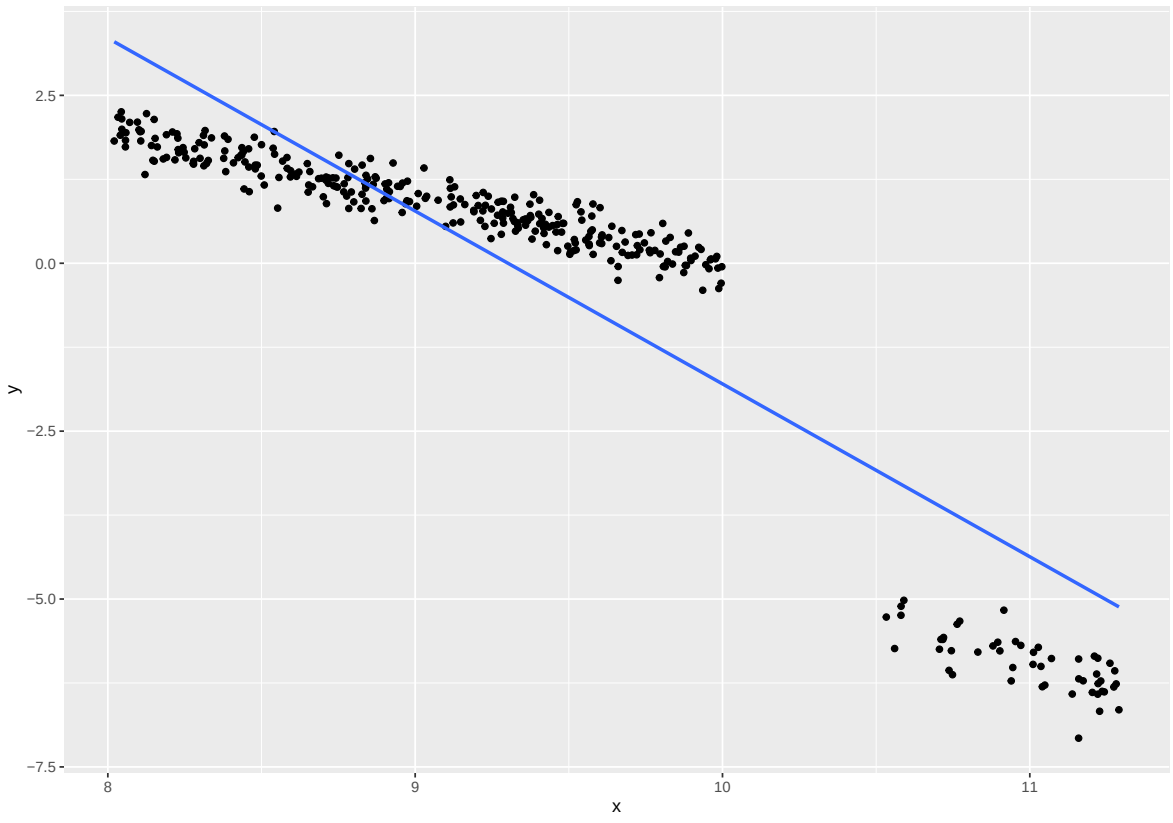
Linear regression

Outliers



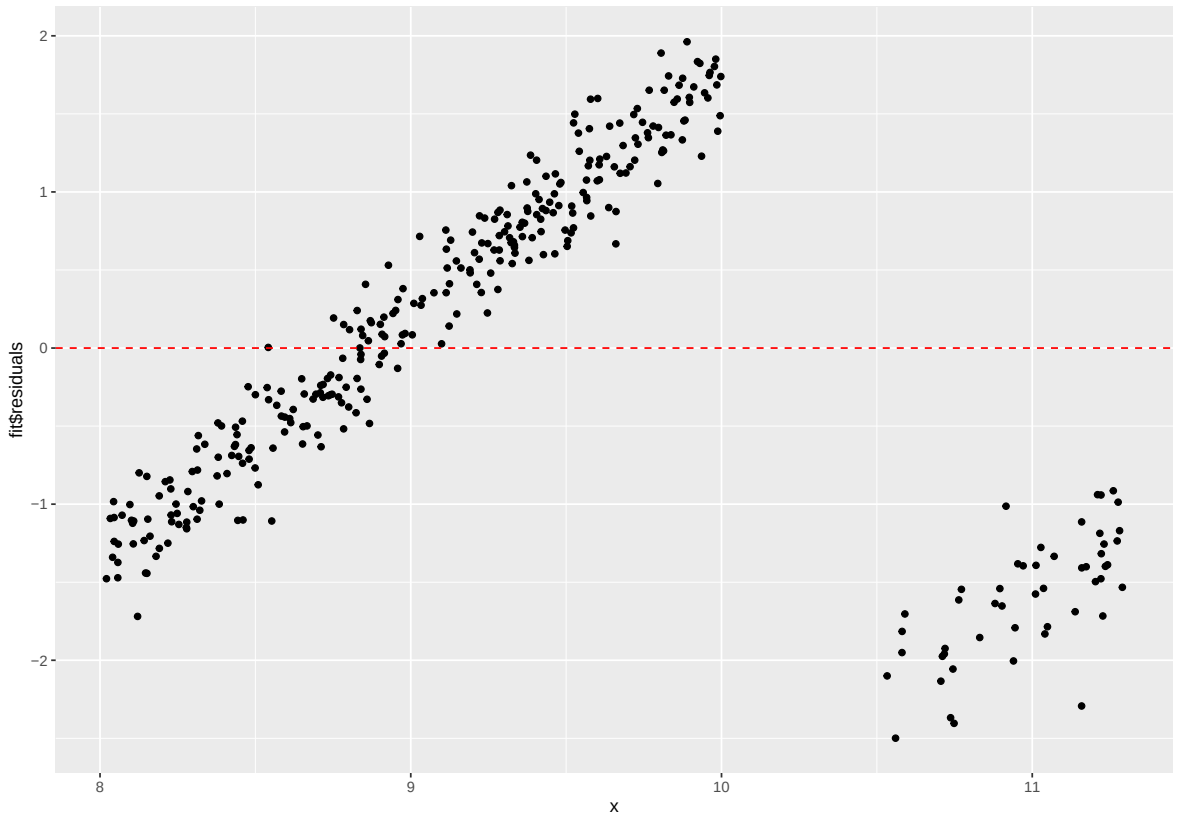
Linear regression

Outliers



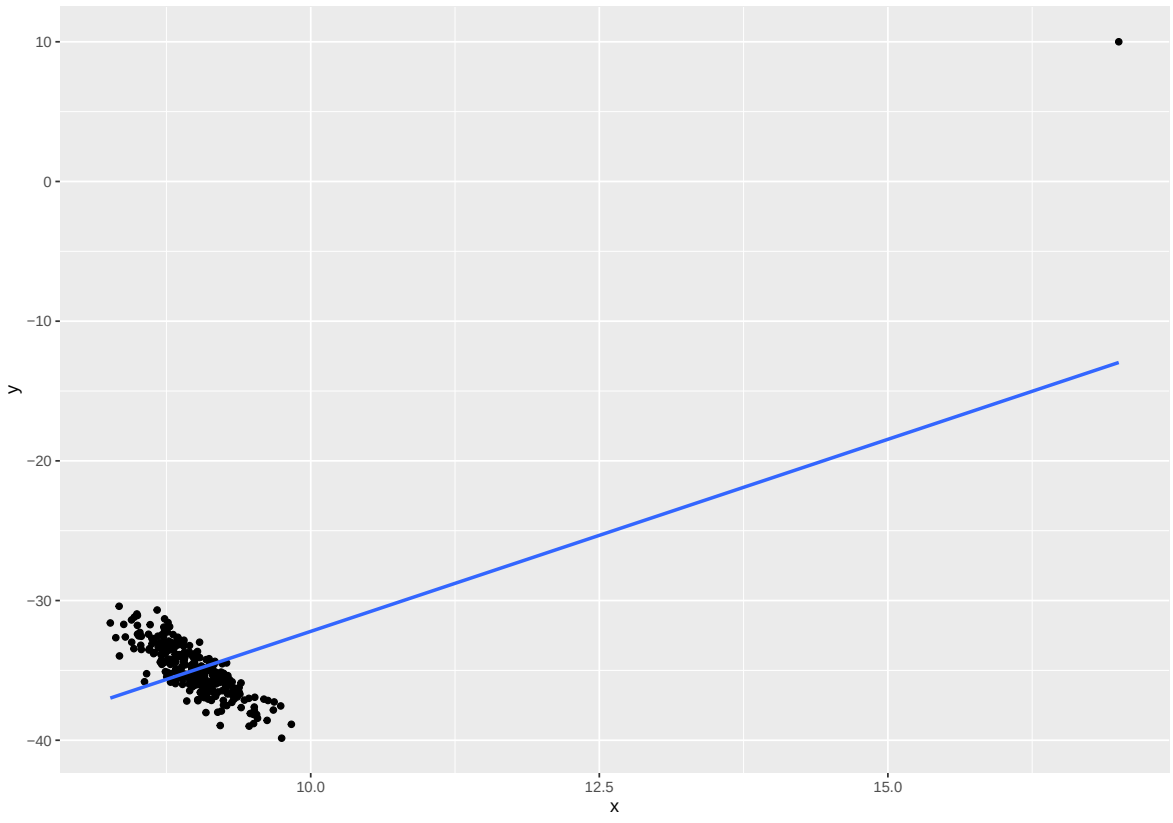
Linear regression

Outliers



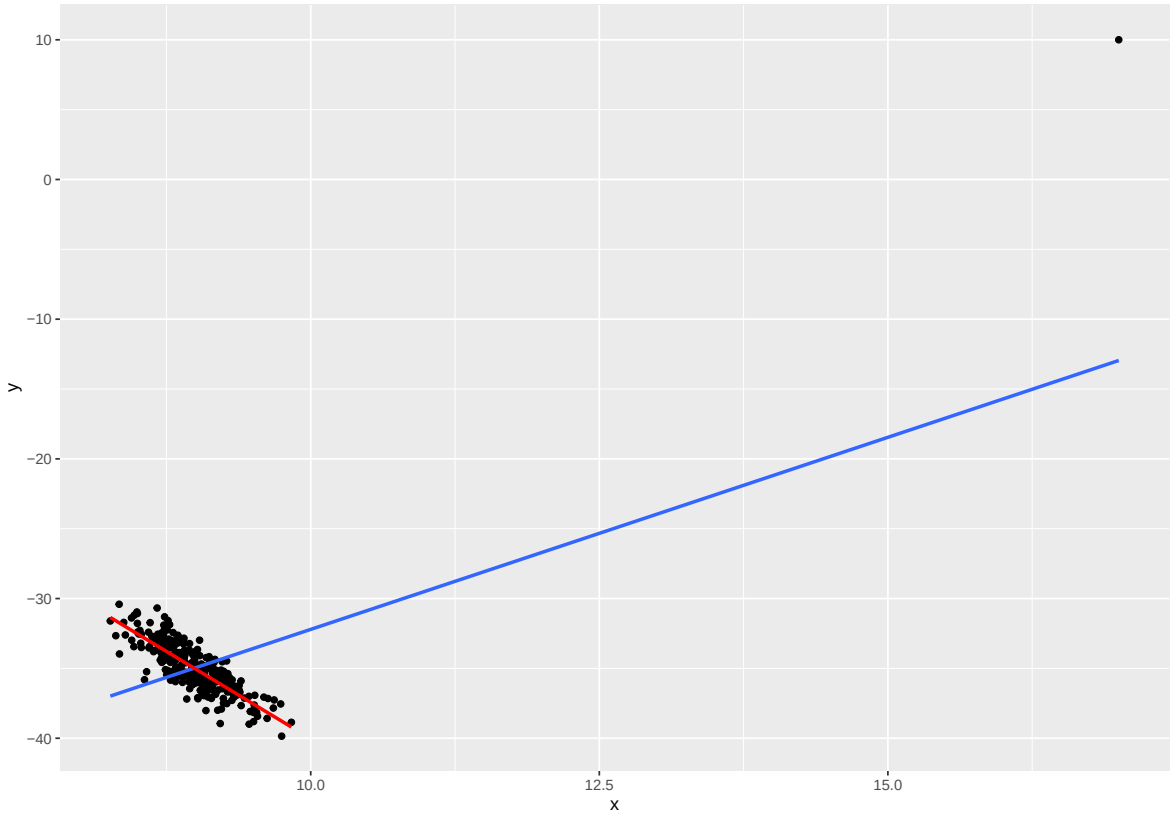
Linear regression

Outliers



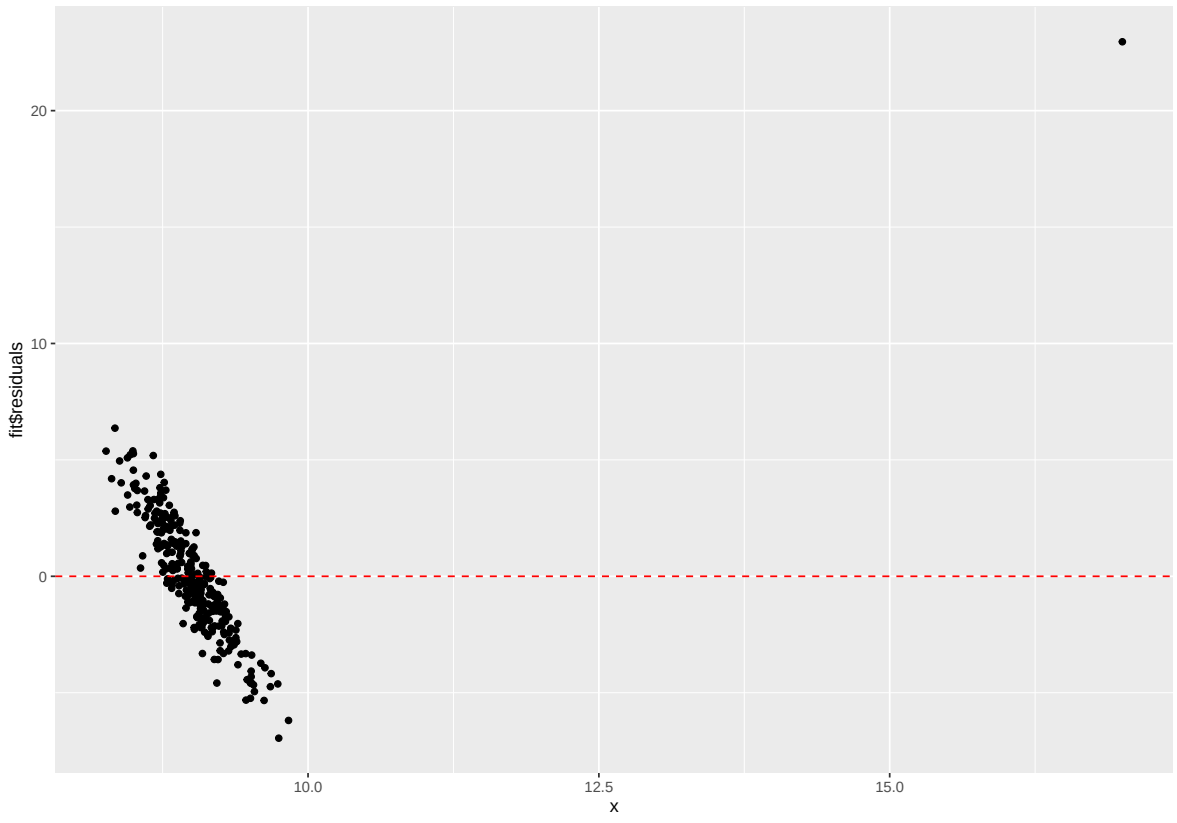
Linear regression

Outliers



Linear regression

Outliers



Linear regression

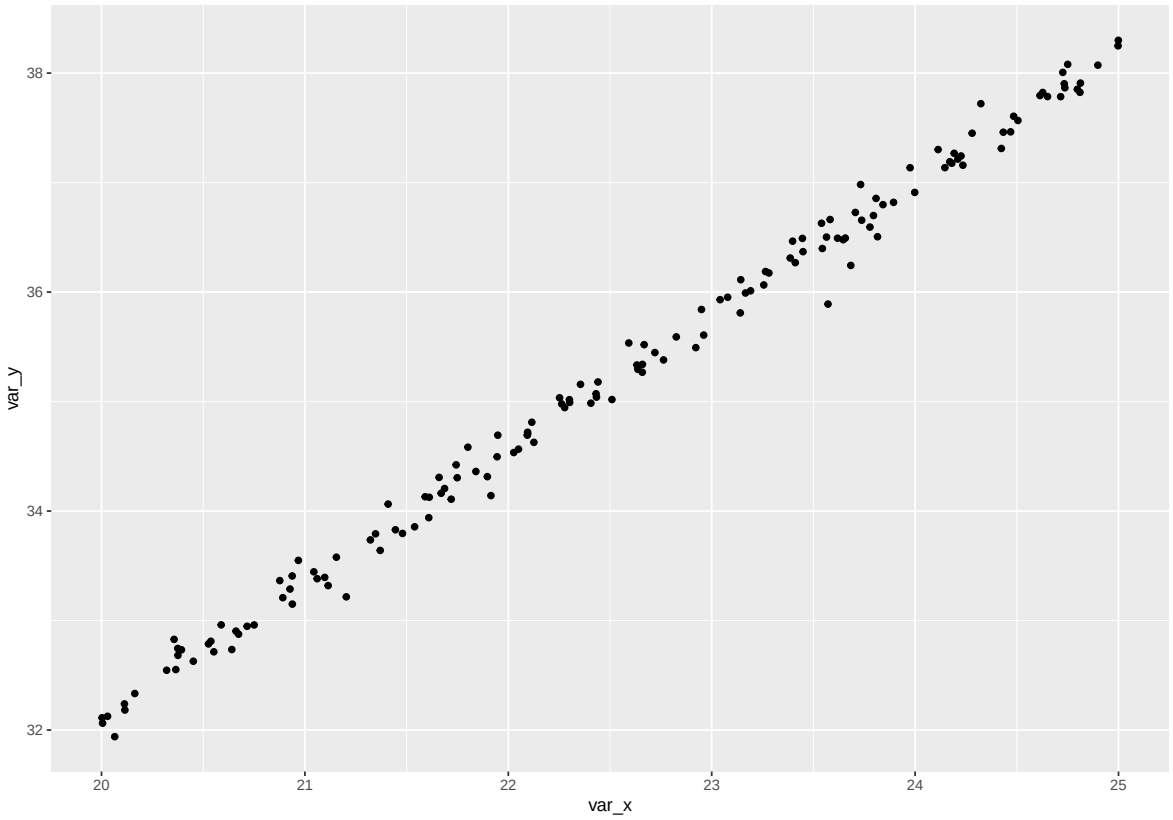
Implementation

linear_data $b_0 = 7.5$, $b_1 = 1.23$ with additional noise

row	var_x	var_y
1	21.58476	34.07142
2	22.21863	34.76214
3	20.09337	32.15816
4	23.34840	36.15238
5	23.68827	36.46497

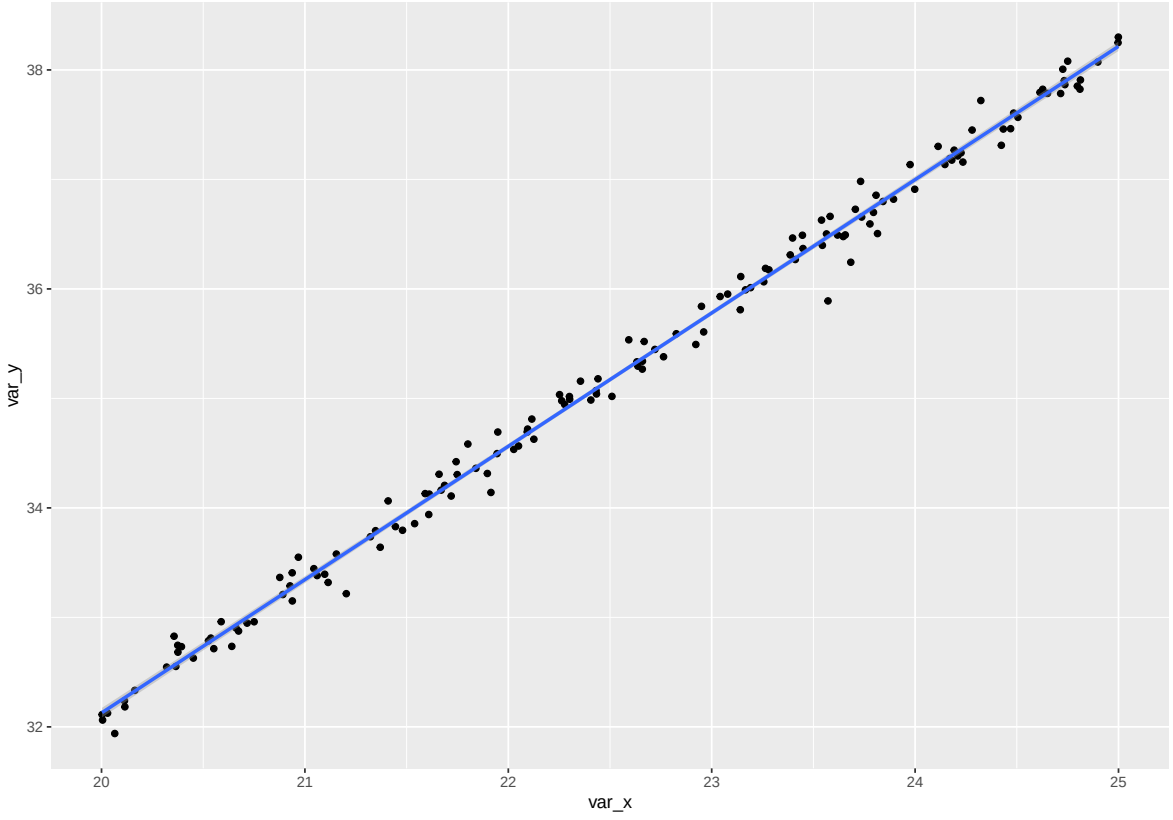
Linear regression

```
ggplot(linear_data) +
  geom_point(aes(x = var_x, y = var_y))
```



Linear regression

```
ggplot(linear_data) +
  geom_point(aes(x = var_x, y = var_y)) +
  geom_smooth(aes(x = var_x, y = var_y), method = lm)
```



Linear regression

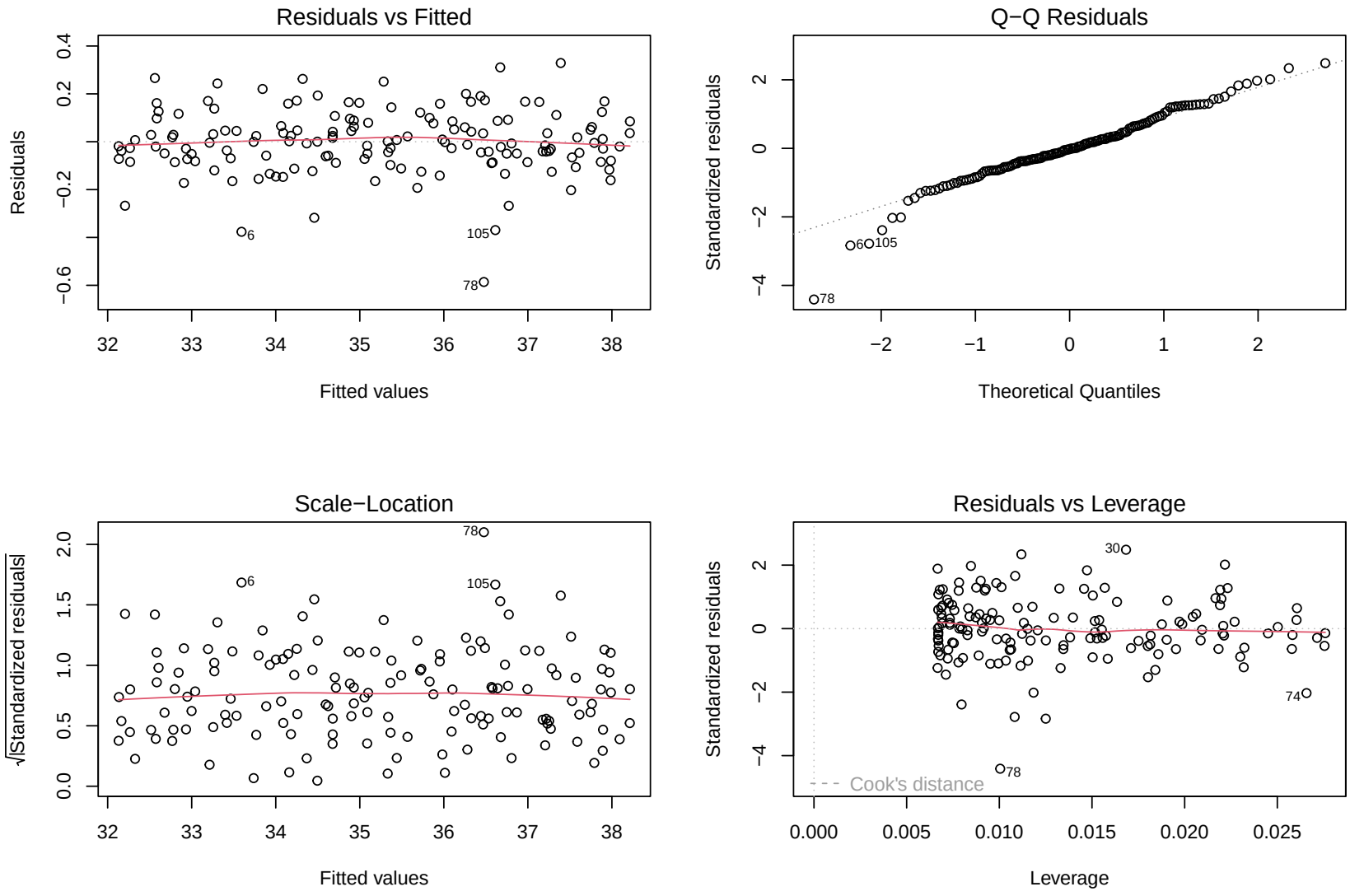
```
fit_linear_data <- lm(var_y~var_x, data = linear_data)
```

Linear regression

- linearity
- nearly normal residuals
- constant variability
- independent observations

Linear regression

```
par(mfrow = c(2, 2))
plot(fit_linear_data)
```



Linear regression

R - corelation coefficient

```
cor(linear_data$var_x, linear_data$var_y)
```

```
## [1] 0.9971393
```

Linear regression

```
summary(fit_linear_data)
```

```
##
## Call:
## lm(formula = var_y ~ var_x, data = linear_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58618 -0.07176 -0.00291  0.08356  0.32896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.775005   0.171420   45.36   <2e-16 ***
## var_x       1.217619   0.007587  160.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1335 on 148 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9942
## F-statistic: 2.576e+04 on 1 and 148 DF,  p-value: < 2.2e-16
```

Linear regression

$$var\_y = 7.7750045 + var\_x \cdot 1.2176193$$

Linear regression

Prediction:

$var\_y = 7.7750045 + var\_x \cdot 1.2176193$

New value x = 24.15

y = ?

```
summary(fit_linear_data)$coefficients[1] +
  24.15 * summary(fit_linear_data)$coefficients[2]
```

```
## [1] 37.18051
```

Linear regression

Inference

```
##
## Call:
## lm(formula = var_y ~ var_x, data = linear_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58618 -0.07176 -0.00291  0.08356  0.32896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.775005   0.171420   45.36   <2e-16 ***
## var_x       1.217619   0.007587  160.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1335 on 148 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9942
## F-statistic: 2.576e+04 on 1 and 148 DF,  p-value: < 2.2e-16
```

Linear regression

Inference - confidence intervals

We are 95% confident that *intercept* value for linear regression fit is between:

```
summary(fit_linear_data)$coefficients[1] -
  1.96 * summary(fit_linear_data)$coefficients[3]
```

```
## [1] 7.439021
```

and

```
summary(fit_linear_data)$coefficients[1] +
  1.96 * summary(fit_linear_data)$coefficients[3]
```

```
## [1] 8.110988
```

Linear regression

Inference - confidence intervals

We are 95% confident that *var_x* multiplier value for linear regression fit is between:

```
summary(fit_linear_data)$coefficients[2] -
  1.96 * summary(fit_linear_data)$coefficients[4]
```

```
## [1] 1.202749
```

and

```
summary(fit_linear_data)$coefficients[2] +
  1.96 * summary(fit_linear_data)$coefficients[4]
```

```
## [1] 1.23249
```

Linear regression

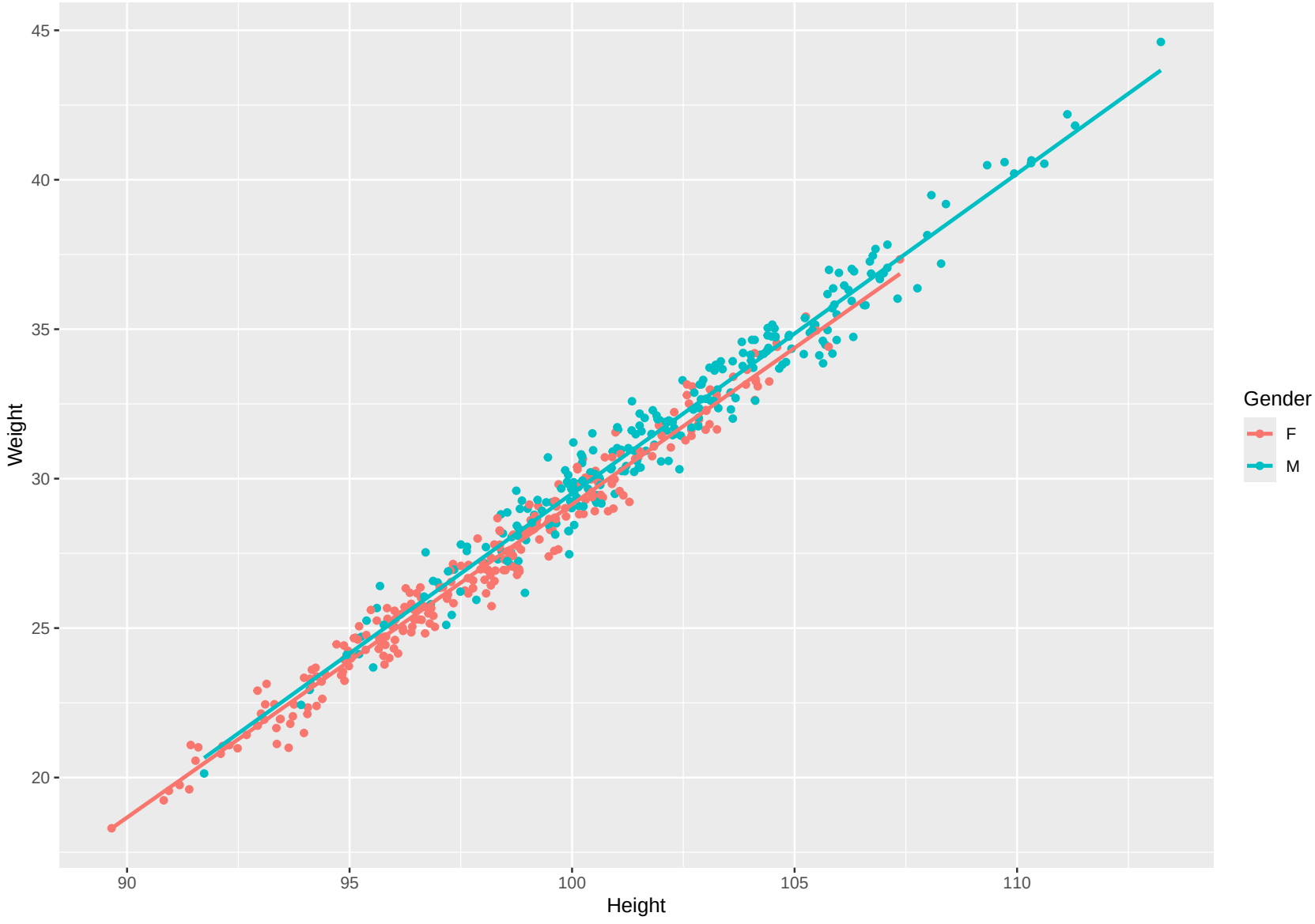
Hypothesis

```
##
## Call:
## lm(formula = var_y ~ var_x, data = linear_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58618 -0.07176 -0.00291  0.08356  0.32896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.775005   0.171420   45.36   <2e-16 ***
## var_x       1.217619   0.007587  160.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1335 on 148 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9942
## F-statistic: 2.576e+04 on 1 and 148 DF,  p-value: < 2.2e-16
```

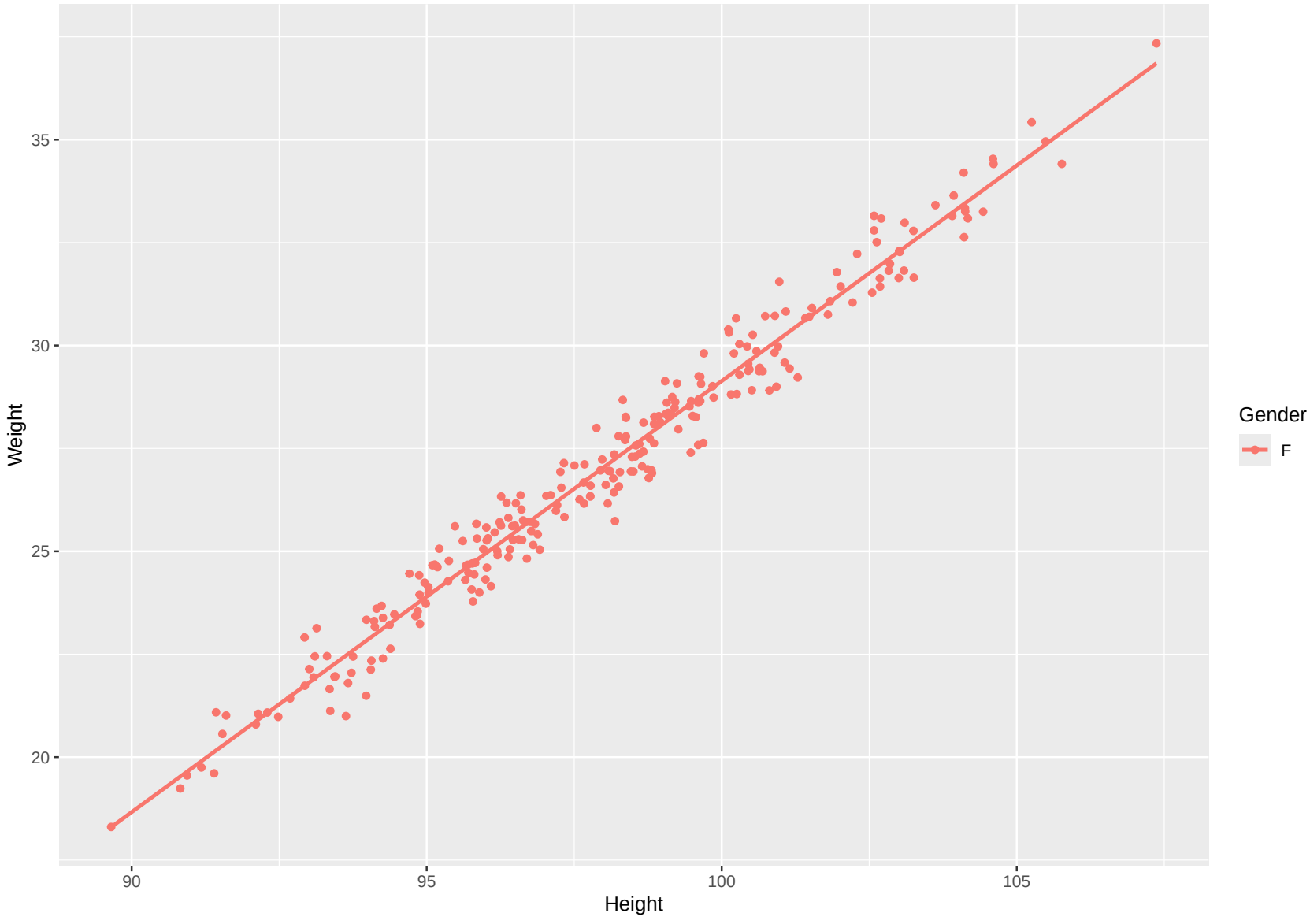
Ewoks

Id	Height	Weight	Gender
1507083	99.16915	28.98076	F
76149	103.24414	31.13418	F
1815472	103.67750	31.96588	M
2902019	101.53223	31.90404	M
1013903	97.88939	26.54659	F
877317	100.42469	30.18863	F
1351471	100.49463	29.95518	F
1430069	103.21104	32.62074	F
120495	104.74980	34.56500	M
2639518	96.41501	25.72783	F

Ewoks



Female Ewoks



Female Ewoks

Correlation

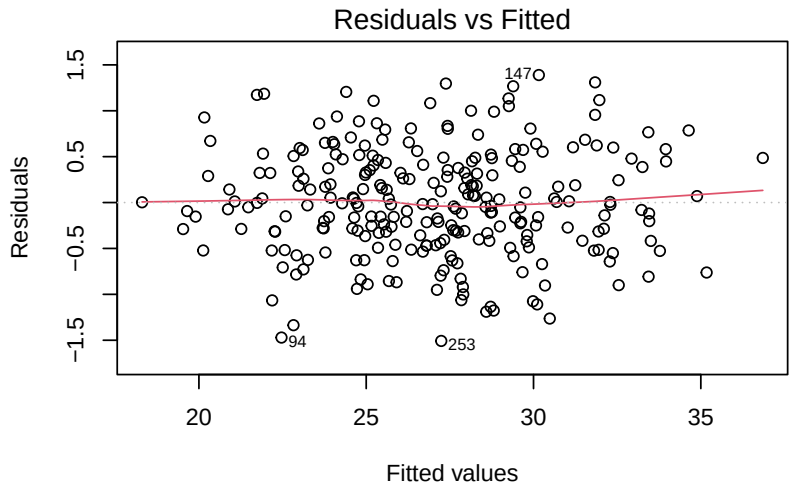
```
cor(females$Weight, females$Height)
```

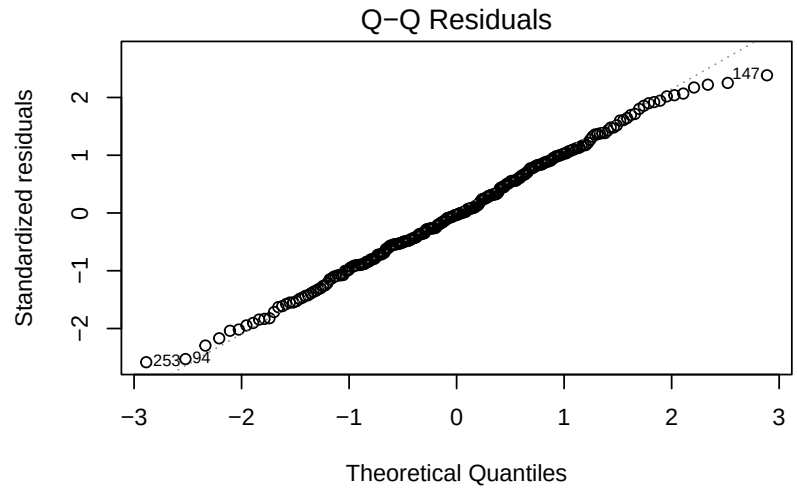
```
## [1] 0.9862896
```

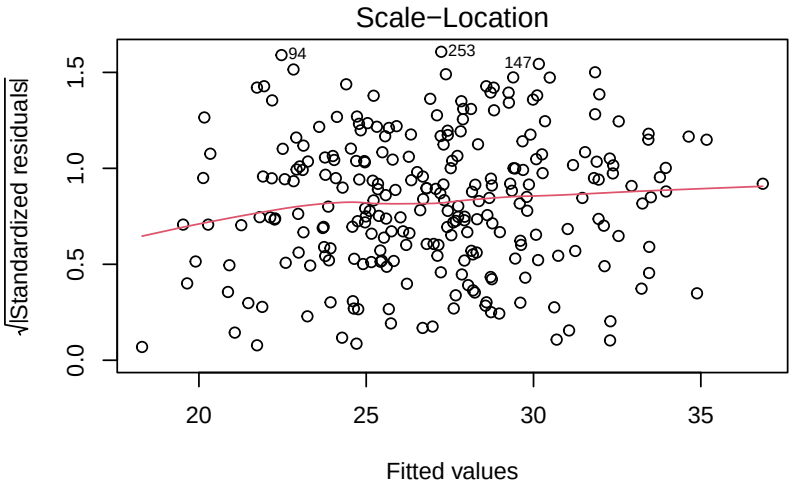
Female Ewoks

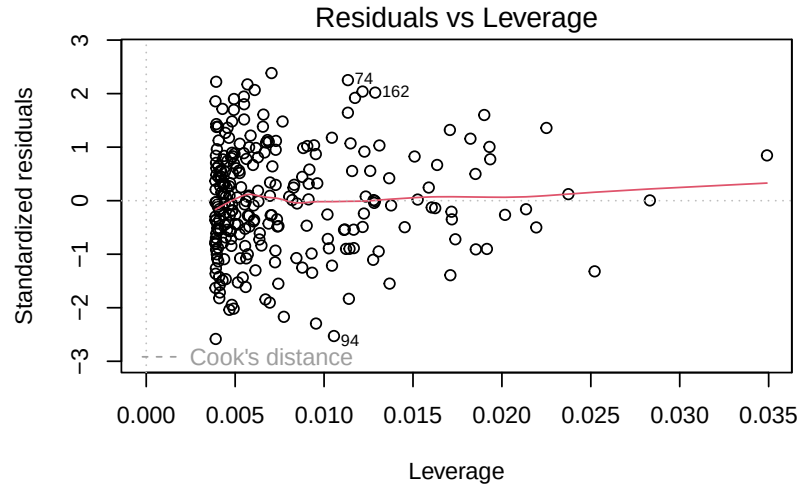
```
##
## Call:
## lm(formula = Weight ~ Height, data = females)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.50787 -0.40152 -0.01801  0.43202  1.38917
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -75.59032    1.07579  -70.27   <2e-16 ***
## Height        1.04727    0.01097   95.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5848 on 255 degrees of freedom
## Multiple R-squared:  0.9728, Adjusted R-squared:  0.9727
## F-statistic:  9109 on 1 and 255 DF,  p-value: < 2.2e-16
```

Female Ewoks





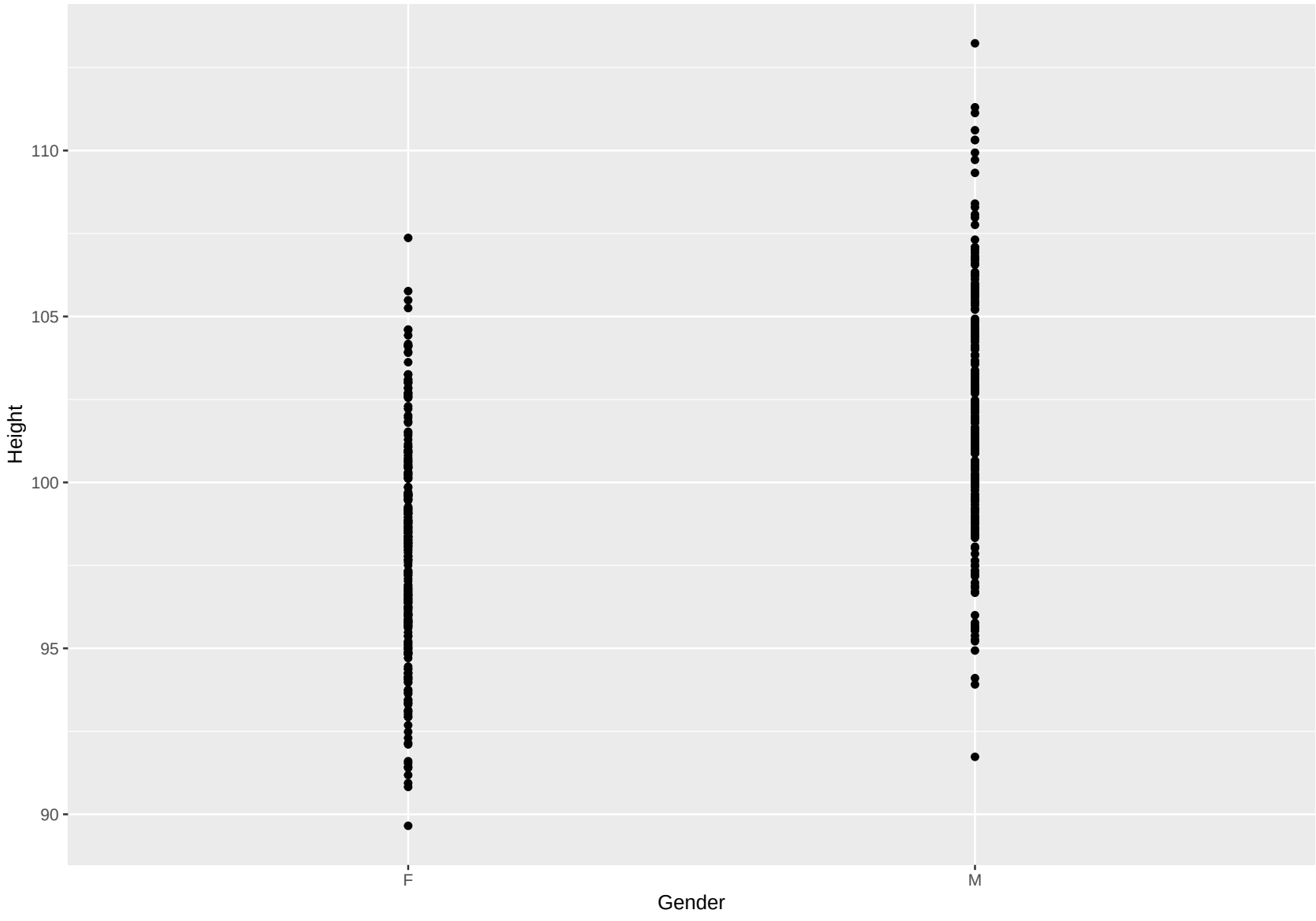




Ewoks

Linear regression for categorical data?

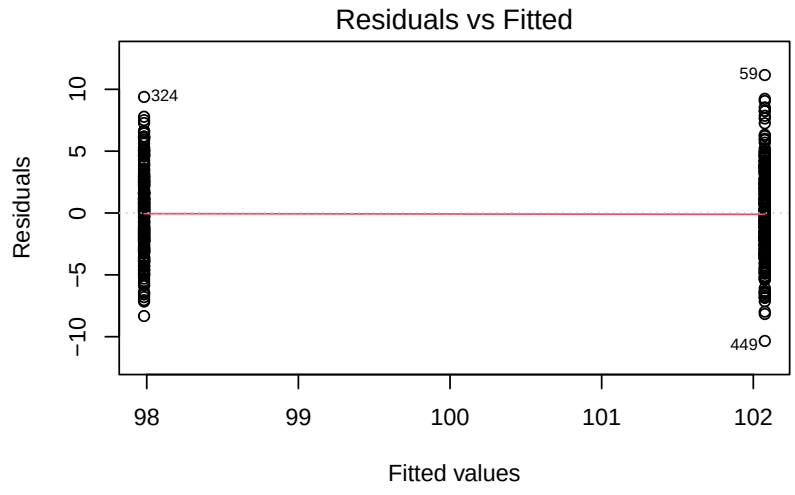
Ewoks

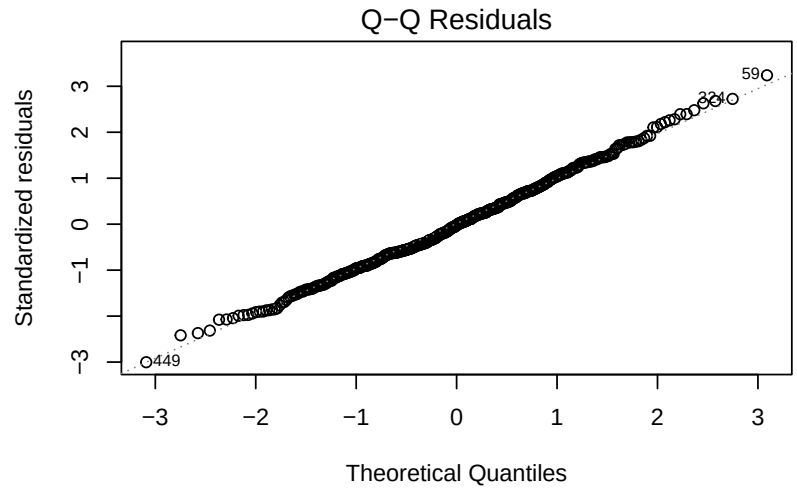


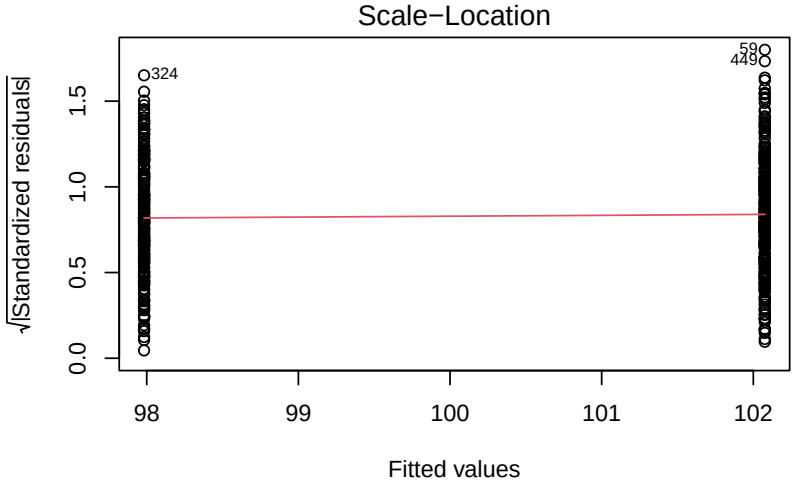
Ewoks

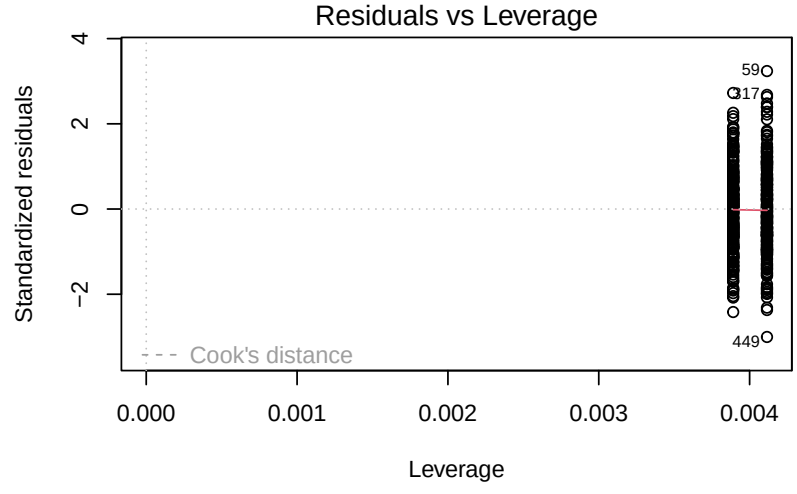
```
##
## Call:
## lm(formula = Height ~ Gender, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3442  -2.2219  -0.0878   2.3203  11.1546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.9824     0.2153  455.11   <2e-16 ***
## GenderM       4.0934     0.3088   13.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.451 on 498 degrees of freedom
## Multiple R-squared:  0.2608, Adjusted R-squared:  0.2593
## F-statistic: 175.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

Ewoks


Residuals vs Fitted
Residuals
Fitted values
324
59
449
98
99
100
101
102
10
5
0
−5
−10


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
59
324
449
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
324
59
449
98
99
100
101
102
0.0
0.5
1.0
1.5


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
59
317
449
0.000
0.001
0.002
0.003
0.004
4
2
0
−2