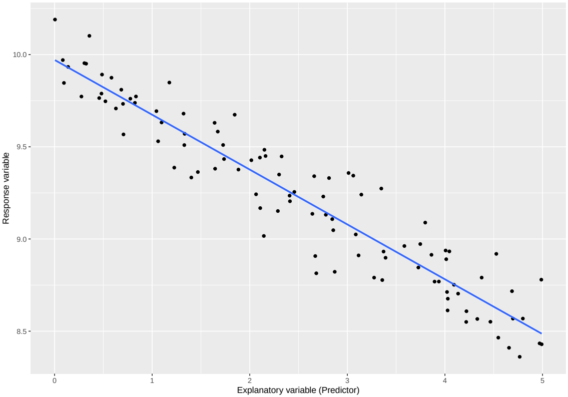
Linear regression - recap

E. Pastucha

November 2024

Linear regression - recap

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad y = b_0 + b_1 x + \varepsilon$$



Linear regression - recap

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 - recap

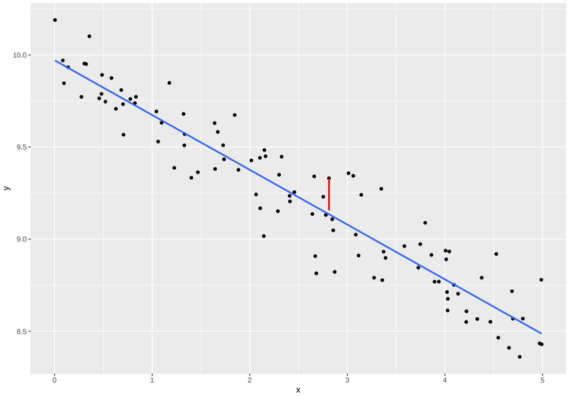
R squared - R^2

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

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{\mathrm{Var}\left(e_i\right)}{\mathrm{Var}\left(y_i\right)}$$

Linear regression - recap

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



Linear regression - recap

Conditions:

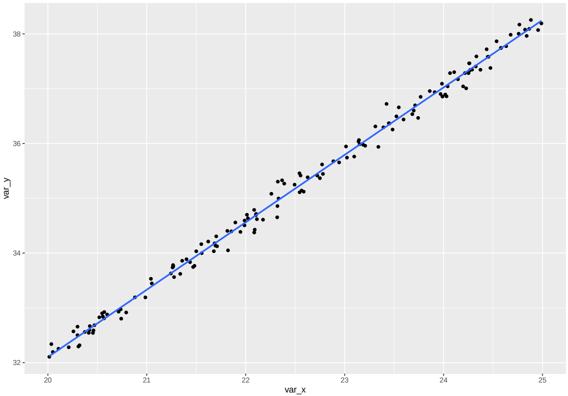
- linearity
- nearly normal residuals
- constant variability
- independent observations

Linear regression - recap

What should you pay attention to? - Outliers

Linear regression - recap

Implementation

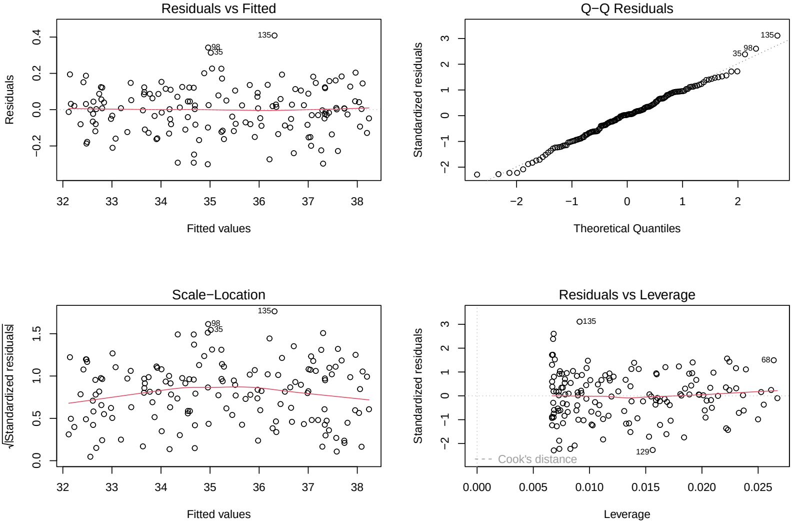


Linear regression - recap

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

Linear regression - recap

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



Linear regression - recap

R - corelation coefficient

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

```
## [1] 0.9973375
```

Linear regression - recap

```
summary(fit_linear_data)
```

```
##
## Call:
## lm(formula = var_y ~ var_x, data = linear_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30079 -0.08625  0.00465  0.09126  0.40854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.494165   0.167029   44.87   <2e-16 ***
## var_x       1.230346   0.007395  166.38   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1319 on 148 degrees of freedom
## Multiple R-squared:  0.9947, Adjusted R-squared:  0.9946
## F-statistic: 2.768e+04 on 1 and 148 DF,  p-value: < 2.2e-16
```

Linear regression - recap

$$var\_y = 7.494165 + var\_x \cdot 1.2303459$$

Linear regression - recap

Prediction:

$var\_y = 7.494165 + var\_x \cdot 1.2303459$

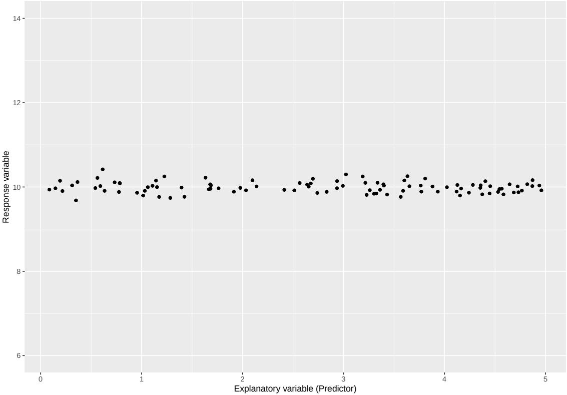
New value x = 24.15

y = ?

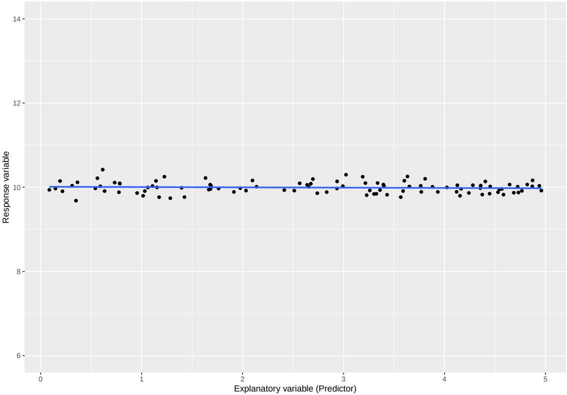
```r
summary(fit_linear_data)$coefficients[1] +
  24.15 * summary(fit_linear_data)$coefficients[2]
```

```
## [1] 37.20702
```

Linear regression - recap - impossible case



Linear regression - recap - impossible case



Linear regression - recap - impossible case

Correlation Coefficient

```r
cor(messy$x, messy$y)
```

```
## [1] -0.07878843
```

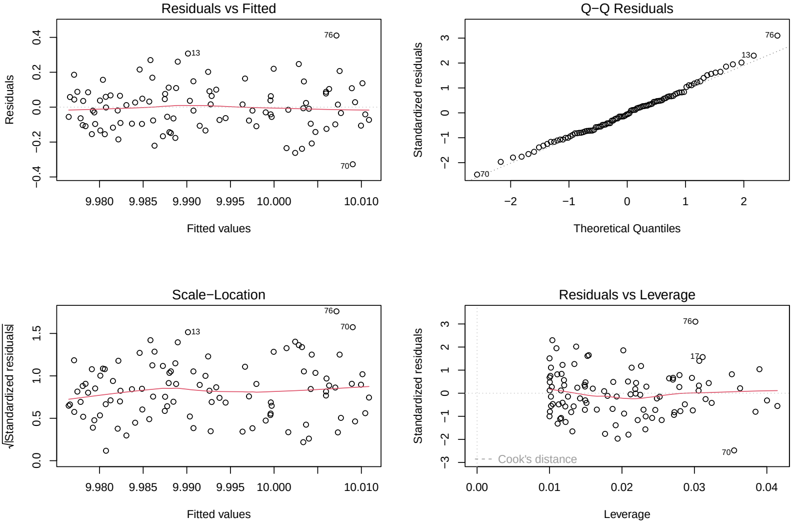
Linear regression - recap - impossible case

```r
summary(lm(y~x, data = messy))
```

```
##
## Call:
## lm(formula = y ~ x, data = messy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32724 -0.09610 -0.00768  0.08007  0.41059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.011483   0.028062 356.769   <2e-16 ***
## x           -0.007060   0.009024  -0.782    0.436
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1345 on 98 degrees of freedom
## Multiple R-squared:  0.006208,   Adjusted R-squared:  -0.003933
## F-statistic: 0.6121 on 1 and 98 DF,  p-value: 0.4359
```

Linear regression - recap - impossible case

```r
par(mfrow = c(2, 2))
plot(lm(y~x, data = messy))
```



Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
76
13
70
9.980
9.985
9.990
9.995
10.000
10.010
Fitted values
Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
76
13
70
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
76
70
13
9.980
9.985
9.990
9.995
10.000
10.010
Fitted values
Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
−2
−3
76
17
70
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage