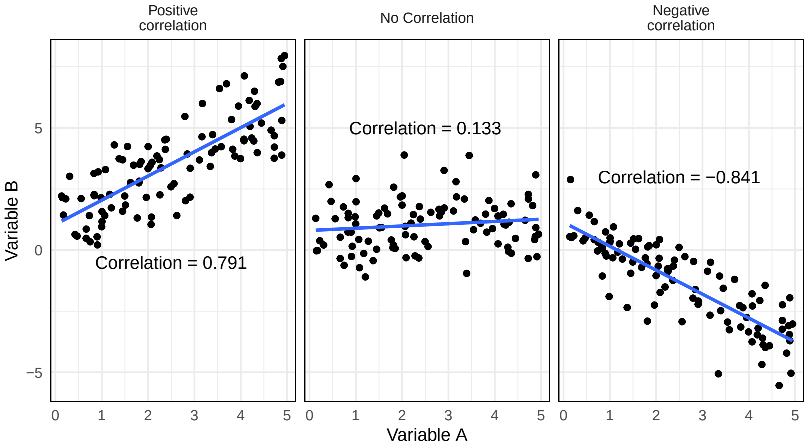
Corrrelation and Linear regression

Daniel Hammarström

2020-10-30

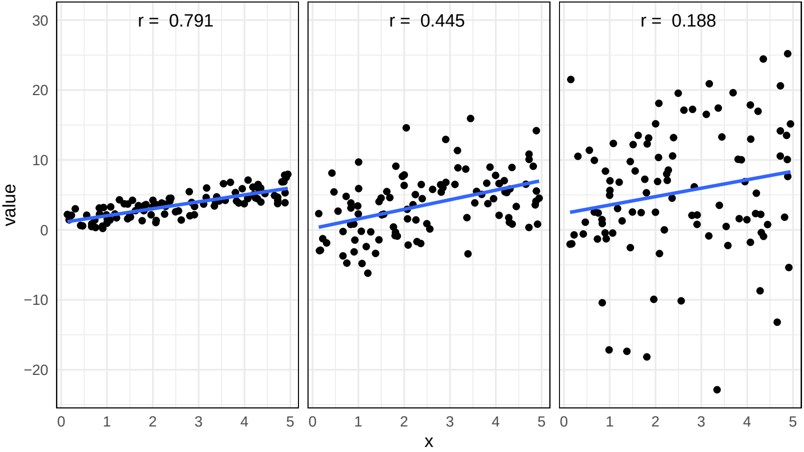
Association between variables

- A measure of association between continuous variables is the correlation (Pearson's correlation coefficient).



Correlation gives a unitless "strength of association"

- Estimates of association (r) is limited to $-1 \leq r \leq 1$.
- When r approaches ± 1, the association is stronger, estimates close to 0 suggest no association.



Assumptions in correlation

- Continuous variables, paired observations
- Bivariate normal distribution(?) – Both variables should be bell shaped.
- Linear relationship between variables
- Be careful when there are outliers, examine the effect of extreme data points.

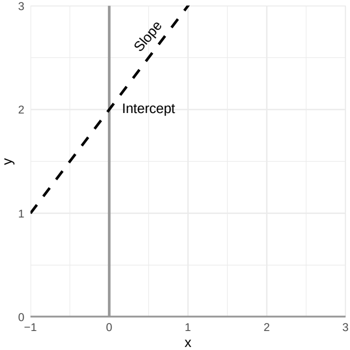
Correlation in R

```r
Yj <- c(25.2, 26.9, 21.7, 15.8,
        26.0, 20.4, 18.5, 15.5, 15.6, 16.0)

Yk <- c(21.9, 25.7, 23.6, 29.6,
        24.9, 23.4, 23.5, 25.1, 24.0, 21.5)

# The Pearson's product moment
# correlation coefficient
cor(Yk, Yj, method = "pearson")

# The Pearson's product moment
# correlation coefficient with
# test statistic
cor.test(Yk, Yj, method = "pearson")
```

Regression models the relationship between variables

- The regression model describe more aspects of the relationship between variables than the correlation.
- The equation for the straight line:

$$y = mx + c \qquad y = slope \times x + intercept$$



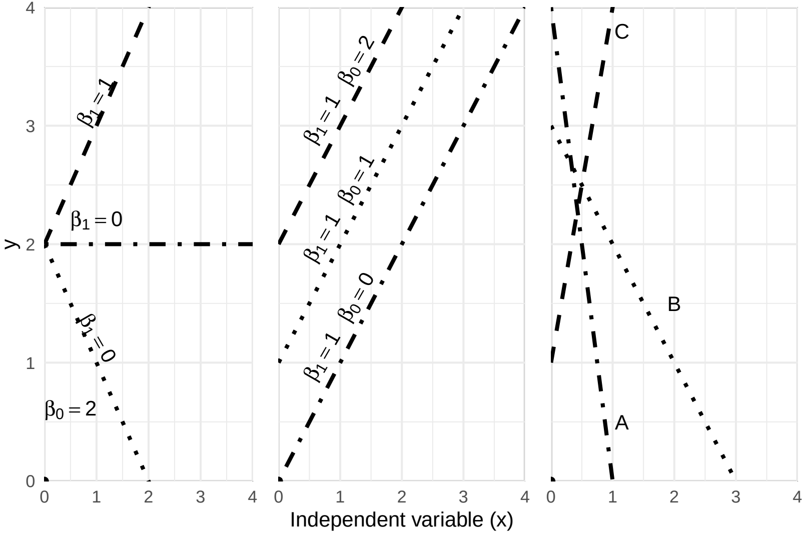
Regression estimates the line that best fits the data

- The basic univariate regression model

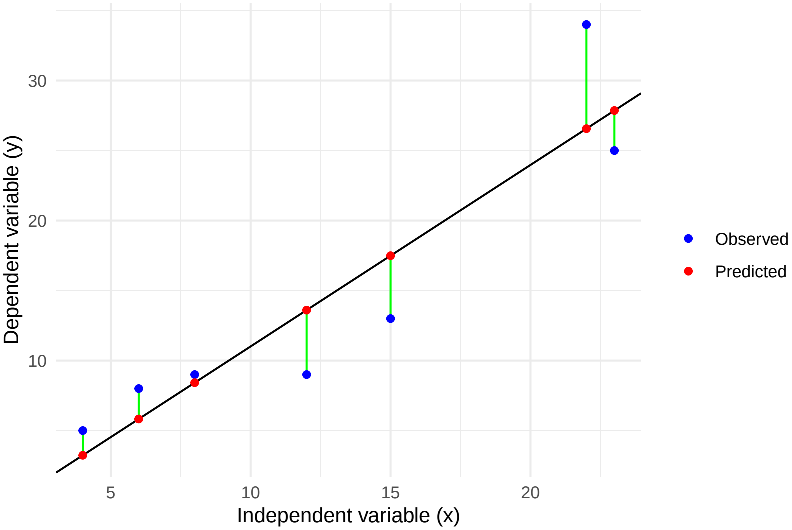
$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 is the model intercept (or constant)
- β_1 is the slope of the straight line
- ϵ is the unexplained error
- Model parameters (β_0, β_1) are estimated using sample data

Interpret slopes and intercepts



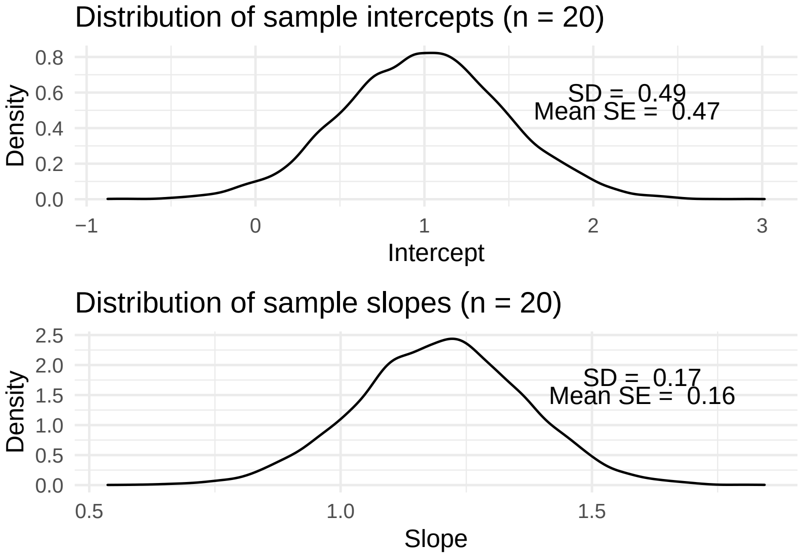
Estimating the best fit line



Estimating the best fit line

- The best fit line can be estimated by minimizing the vertical distances between **observed** and **predicted** values.
- The distance between observed and predicted values are called **residuals**, these can help us *diagnose* the regression.
- The **residuals** are also used to *estimate* the standard errors of the parameters in the model.

The standard error of the regression parameter is an estimate of the SD of the sampling distribution



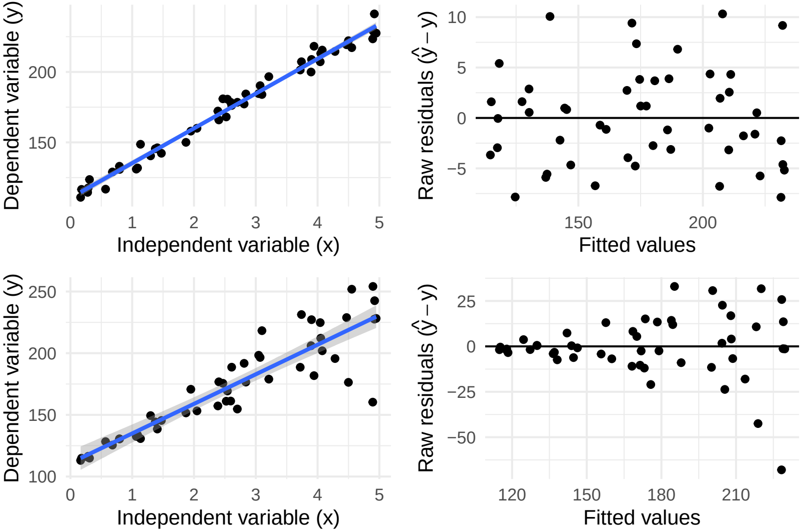
Assumptions in linear regression

- There is a linear relationship between x and y
- Residuals are normally distributed (with mean $= 0$)
- Residuals have an equal spread along the the fitted range (homoskedasticity)
- Observations are independent

Why are assumptions important

- We assume that errors in our model $(\hat{y} - y)$ are well behaved
- The errors are used to calculate standard errors
- If the assumptions are wrong our standard errors are **biased**
- *Biased* standard errors will lead to bad **inference**

Model diagnostics

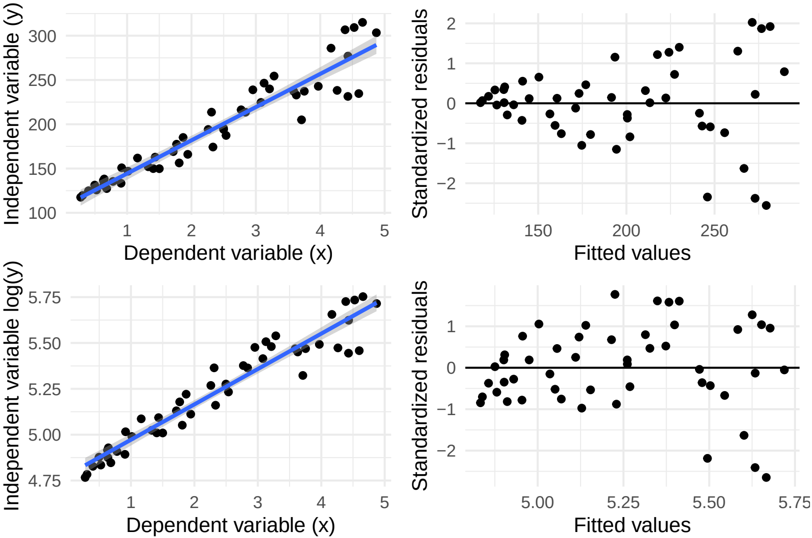


What can be done with heteroscedasticity?

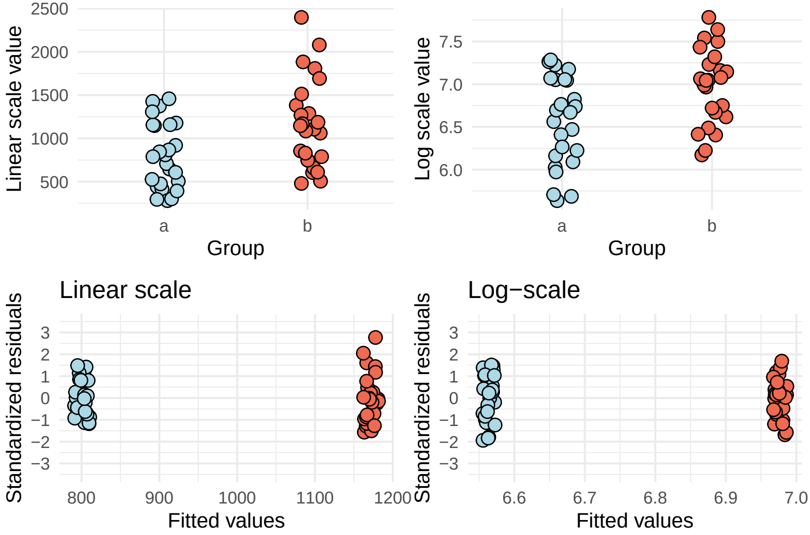
- Transformation of the data can reduce increased variation with increased values
- The most common transformation is the log
- log-transformed data

```
df$y <- log(df$y)
```

Log-transfomed data



Interpreting log-transfomed data in a regression



Interpreting log-transfomed data in a regression

Paramter	Estimate	SE	t-value	p-value	Model
(Intercept)	800.90	89.396	8.96	0.000	Linear scale
groupb	370.73	126.425	2.93	0.005	Linear scale
(Intercept)	6.56	0.097	68.01	0.000	Log-transformed
groupb	0.41	0.136	3.03	0.004	Log-transformed

$$log(a) + log(b) = log(a \times b)$$

$$log(a) - log(b) = log(a/b)$$

$$e^{log(x)} = x$$

Categorical data can be used as predictor variables

- We can use categorical data as the independent variable
- Categories are (automatically in R) converted to "dummy varaibles"
- If we have two groups (e.g. men and women), in the univariate model, men will be represented by the intercept and women by the slope.

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

$$y = MEN + \beta_1 \times WOMEN + \epsilon$$

- If there are more levels, additional dummy variables are added to the model