# Simple and multiple linear regression

Nora Wickelmaier

Last modified: October 15, 2024

# Outline

# What is regression?

# What is regression?

Set of statistical processes for estimating the relationships between a dependent variable (often called the ‘outcome variable’) and one or more independent variables (often called ‘predictors’, ‘covariates’, or ‘features’)

https://en.wikipedia.org/wiki/Regression_analysis

# What is regression?

Set of statistical processes for estimating the relationships between a dependent variable (often called the ‘outcome variable’) and one or more independent variables (often called ‘predictors’, ‘covariates’, or ‘features’)

https://en.wikipedia.org/wiki/Regression_analysis

- Predict an outcome variable
- Compare predictions for different groups
- “Find the line that most closely fits the data”
- Continuous outcome $y$

# 1 Basic concepts

# Simple linear regression

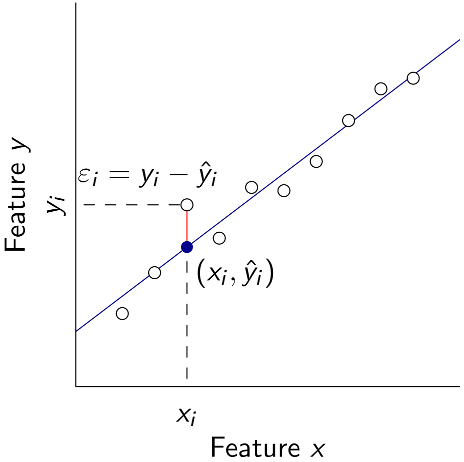
- For the pairs

$$(x_1, y_1), \ldots, (x_n, y_n),$$

  we get the stochastical model

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

  for all $i = 1, \ldots, n$
- Errors are independent identically distributed (i.i.d.)



# Simple linear regression

- From the properties of the error variables, we conclude

$$E(y_i) = E(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) = \beta_0 + \beta_1 \cdot x_i = \bar{y}$$

and

$$Var(y_i) = Var(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) = \sigma^2$$

- For a given $x_i$, the stochastical independence of $\varepsilon_i$ transfers to $y_i$

# Parameter estimation

- The parameters $\beta_0$ and $\beta_1$ are estimated with the method of least squares
- Hereby, the sum of squares of the residuals is minimized

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \min$$

- The minimum is obtained by setting the partial derivatives for $\beta_0$ and $\beta_1$ to 0

$$\frac{\partial \left( \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right)}{\partial \beta_0} \qquad \frac{\partial \left( \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right)}{\partial \beta_1}$$

- Solving these equations results in

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2}$$

where $\sigma_{xy}$ is the covariance between $x$ and $y$

# Correlation coefficient

- The correlation coefficient $r$ is defined as the standardized covariance

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Hence, we get

$$\hat{\beta}_1 = \frac{\sigma_y}{\sigma_x} r$$

- When $x$ and $y$ are $z$ standardized with $\bar{x} = \bar{y} = 0$ and $\sigma_x = \sigma_y = 1$, we get

$$\hat{\beta}_0 = 0 \quad \text{and} \quad \beta_1 = r$$

# Determination coefficient

- With the assumptions made so far, it can be shown that the variance of the residuals

$$\sigma_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

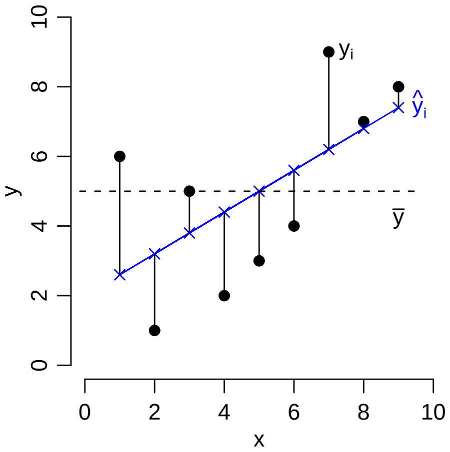
can be rewritten as

$$\sigma_\varepsilon^2 = (1 - r^2)\sigma_y^2$$

- The factor $(1 - r^2)$ determines the proportion of the variance of $y$ that cannot be explained by the regression of $x$
- Hence, he determination coefficient $r^2$ gives the proportion of the variance of $y$ explained by $x$

$$r^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

# Variance decomposition




$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

## Exercise

- Simulate a data set based on a simple regression model with

$$\begin{aligned}\beta_0 &= 0.2\\ \beta_1 &= 0.3\\ \sigma &= 0.5\\ x &\in [1, 20] \text{ in steps of } 1\end{aligned}$$

- What functions in $R$ do we need?

# Simulate data set

```
x     <- 1:20
n     <- length(x)
a     <- 0.2
b     <- 0.3
sigma <- 0.5
y     <- 0.2 + 0.3 * x + rnorm(n, sd = sigma)

dat <- data.frame(x, y)

# clean up workspace
rm(x, y)

# plot data
plot(y ~ x, data = dat)
```

# Fit regression model

```
lm1 <- lm(y ~ x, data = dat)
summary(lm1)

mean(resid(lm1))
sd(resid(lm1))
hist(resid(lm1), breaks = 15)

# plot data
plot(y ~ x, data = dat)
abline(lm1)
```

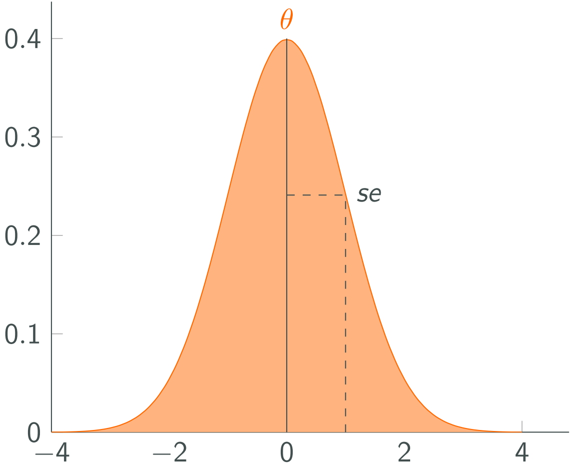
# Re-cover parameters

```
pars <- replicate(2000, {
  ysim <- 0.2 + 0.3 * x + rnorm(n, sd = sigma)
  lm1  <- lm(ysim ~ x, data = dat)
  c(coef(lm1), sigma(lm1))
})

rowMeans(pars)
# standard errors
apply(pars, 1, sd)

hist(pars[1, ])
hist(pars[2, ])
hist(pars[3, ])
```

# Sample distribution



## Exercise

- Simulate data with the parameters from slide 8
- Do not assume that we have one subject per value for $x$, but more than one subject
- Simulate data for $n = 40$ and $n = 100$
  Hint: Use `sample(x, n, replace = TRUE)`
- Re-cover your parameters as done on slide 11
- What happens to your standard errors?

# Confidence intervals

- We get the $(1-\alpha)$ confidence intervals for the estimates with

$$\left[\hat{\beta}_0 - \hat{\sigma}_{\hat{\beta}_0}\, t_{1-\beta_0/2}(n-2),\ \hat{\beta}_0 + \hat{\sigma}_{\hat{\beta}_0}\, t_{1-\beta_0/2}(n-2)\right]$$

and

$$\left[\hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1}\, t_{1-\alpha/2}(n-2),\ \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1}\, t_{1-\alpha/2}(n-2)\right]$$

- For $n > 30$ the $t$ quantiles of the $t(n-2)$ distribution can be replaced by quantiles of the $N(0,1)$ distributiom
- For a sufficient sample size, even when the normality assumption is violated, the least square estimators is approximately $t$ or normally distributed

# Hypothesis tests

## Wald test

- The estimates for $\beta_0$ and $\beta_1$ are unbiased, sufficient, consistent, and efficient
- The normality assumption implies

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}) \quad \text{and} \quad \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

and with that for some hypothetical value $\gamma_0$

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \gamma_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2) \quad \text{und} \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \gamma_0}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2)$$

with estimates $\hat{\sigma}^2_{\hat{\beta}_0}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$ for the variances

- We are usually interested in the hypotheses $\beta_0 = 0$ and $\beta_1 = 0$, hence $\gamma_0 = 0$ and

$$T_{\beta_0} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t(n-2) \quad \text{und} \quad T_{\beta_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t(n-2)$$

# Overall $F$ test

- For simple linear regression, we can construct an equivalent test for testing $\beta_1 = 0$ using variance decomposition
- This test can conceptionally be considered to test, if predictor $x$ explains a significant proportion of the variance of $y$

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \frac{R^2}{1 - R^2}(n-2)$$

- It can be shown that $F = T^2 - \gamma_0$ for $\gamma_0 = 0$ with $F \sim F(1, n-2)$
- With this more general test, we can also test the assumption that any variance of $y$ is explained by the predictors in a multiple regression

## 2 Assumptions

# Assumptions

- We have the stochstic model $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.
- The error variables $\varepsilon_i$ are considered to be non-observable and comprise influence that cannot be controlled and is unsystematic (or random)
- Hence, it makes sense to assume

$$E(\varepsilon_i) = 0, \text{ for all } i = 1, \ldots, n$$

and

$$Var(\varepsilon_i) = \sigma^2, \text{ for all } i = 1, \ldots, n$$

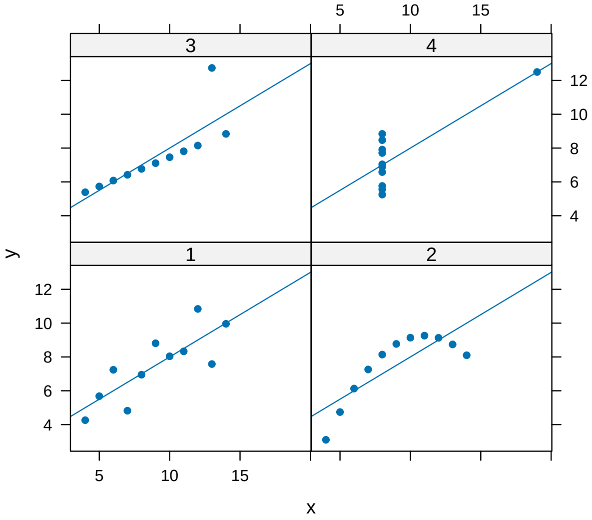
- From this follows

$$E(y_i) = E(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) = \beta_0 + \beta_1 \cdot x_i$$

and

$$Var(y_i) = Var(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) = \sigma^2$$

# Assumptions




- Four data sets by Anscombe (1973) with the same traditional statistical properties (mean, variance, correlation, regression line, etc.)
- Available in R with `data(anscombe)`

# Assumptions

```
data(anscombe)

lm1 <- lm(y1 ~ x1, anscombe)
lm2 <- lm(y2 ~ x2, anscombe)
lm3 <- lm(y3 ~ x3, anscombe)
lm4 <- lm(y4 ~ x4, anscombe)

rbind(coef(lm1), coef(lm2), coef(lm3), coef(lm4))

par(mfrow = c(2, 2))
plot(lm1)
plot(lm2)
plot(lm3)
plot(lm4)
```

## Exercise

- Create two vectors $x$ and $y$ with 100 observations each and $X \sim N(1,1)$ and $Y \sim N(2,1)$
- Create a data frame with variables `id`, `group` and `score`. $x$ and $y$ are your score values
- Conduct a $t$ test assuming that $X$ and $Y$ are independent having the same variances
- Then use the function `aov()` to compute an analysis of variance for these data
- Use then function `lm()` for a linear regression with predictor `group` and dependent variable `score`
- Compare your results

# Extending simple linear regression

| | |
|---|---|
| Additional predictors | $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$ |
| Nonlinear models | $\log y = \beta_0 + \beta_1 \log x + \varepsilon$ |
| Nonadditive models | $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$ |
| Generalized linear models | $g(E(y)) = \beta_0 + \beta_1 x$ |
| Mixed-effects models | $y = \beta_0 + \beta_1 x_1 + \beta_2 \mathit{time} + \upsilon_0 + \upsilon_1 \mathit{time} + \varepsilon$ |

...

# 3 Multiple linear regression

# Multiple linear regression

- Empirical observations consist of tuples for each observation unit

$$(y_i, x_{i1}, \dots, x_{ip}) \;\text{ with }\; i = 1, \dots, n$$

and we get the stochastical model

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2) \text{ i.i.d.}$$

which transfers to

$$y_i \sim N(\mu_i, \sigma^2) \;\text{ with }\; \mu_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip}$$

- The criterion variable $y$ is always a metric variable, whereas the predictor variables $x_1, \dots, x_p$ can be either metric or categorical variables, or both

# Overall $F$ test

- Hypotheses

$$\begin{aligned} H_0 &: \ \beta_1 = \ldots = \beta_p = 0 \\ H_1 &: \ \beta_j \neq 0 \ \text{ for at least one } j \in \{1, \ldots, p\} \end{aligned}$$

- Test statistic

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p}$$

- Distribution of test statistic assuming $H_0$ is true

$$F \sim F(p, n - p - 1)$$

- Rejection region

$$F > F_{1-\alpha}(p, n - p - 1)$$

# Incremental $F$ test

- We have two nested models $M_1$ and $M_0$, meaning that $M_0$ is a special case of $M_1$ where some parameters $\beta_{M_0,j} = 0$ and $\beta_{M_1,j} \neq 0$ and want to test if
$$R_1^2 > R_0^2$$
- Test statistic
$$F = \frac{R_1^2 - R_0^2}{1 - R_1^2} \cdot \frac{n - q_1}{q_1 - q_0}$$
- Distribution of test statistic assuming $H_0$ is true
$$F \sim F(q_1 - q_0, n - q_1)$$
- Rejection region
$$F > F_{1-\alpha}(q_1 - q_0, n - q_1)$$

# Likelihood ratio test

- The incremental $F$ test is a special case of the more general likelihood ratio test

$$G^2 = 2\log\frac{L_1}{L_0}$$

  where $L_1$ is the likelihood of the more general model and $L_0$ the likelihood of the smaller model
- The models need to be nested
- The test statistic $G^2$ is $\chi^2$ distributed

$$G^2 \sim \chi^2(q_1 - q_0)$$

  where $q_1$ and $q_0$ are the number of parameters for the bigger and the smaller model, respectively

# Example: Multiple linear regression

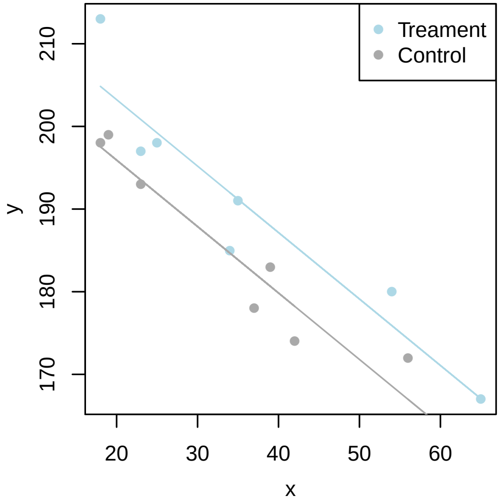
- We are fitting the following model

$$y_{ij} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot z_j + \varepsilon_{ij}$$

  with $i = 1 \ldots N$ and $j = 1, 2$ for two groups
- This means that we have one dummy variable for $z$ which takes the values 0 and 1
- Hence, we get the two models

$$y_{i1} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot 0 + \varepsilon_{ij} = \beta_0 + \beta_1 \cdot x_i + \varepsilon_{ij}$$
$$y_{i2} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot 1 + \varepsilon_{ij} = (\beta_0 + \beta_2) + \beta_1 \cdot x_i + \varepsilon_{ij}$$

# Example: Multiple linear regression



# Example: Multiple linear regression

```
dat <- data.frame(
  x = c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37),
  y = c(213, 197, 198, 191, 167, 180, 185, 172, 153, 199, 193, 174,
        198, 183, 178),
  z = rep(c("treatment", "control"), c(7, 8))
)

aggregate(y ~ z, dat, mean)
```

# Example: Multiple linear regression

- We can now use the parameters to calculate adjusted means for the two groups
- The observed means are $\bar{x}_{treat} = 190.14$ and $\bar{x}_{contr} = 181.25$
- The adjusted means correspond to

$$\bar{x}_{contr} = \beta_0$$
$$\bar{x}_{treat} = \beta_0 + \beta_2$$

  These are the means for a value of $x = 0$ which should have a meaningful interpretation
- Hence, it might be indicated to center $x$

# Example: Multiple linear regression

```
dat$xc <- dat$x - mean(dat$x)

lm2 <- lm(y ~ xc + z, dat)
summary(lm2)

# adjusted means
coef(lm2)[1]
coef(lm2)[1] + coef(lm2)[3]
```

## Exercise

- The data set `cars` contains speed and stopping distances of 50 cars
- Estimate the regression model

$$dist_i = \beta_0 + \beta_1 speed_i + \varepsilon_i$$

- How much variance of the stopping distances is explained by speed?
- Look at the residuals of the model. Are there any systematic deviances?
- Now estimate the model

$$dist_i = \beta_0 + \beta_1 speed_i + \beta_2 speed_i^2 + \varepsilon_i$$

  Hint: Use `I(speed^2)` in the model formula in `R`
- Which model fits the data better?

# References

Anscombe, F. J. (1973).Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.