Simple and multiple linear regression

Nora Wickelmaier

Last modified: October 7, 2024

Outline

1. Basic concepts
2. Assumptions
3. Multiple linear regression

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 $i = 1, \ldots, n$

Simple linear regression

- From the properties of the error variables, we conclude

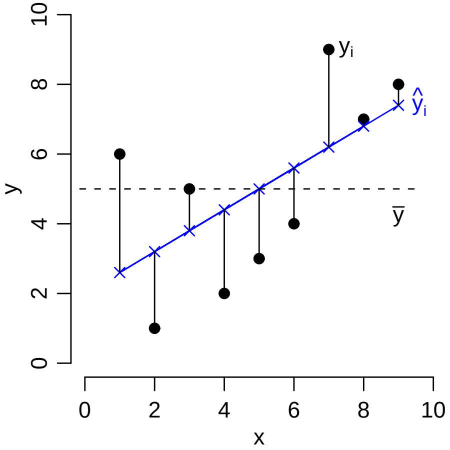
$$E(y_i) = E(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) = \beta_0 + \beta_1 \cdot x_i = \bar{y}$$

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 ε_i transfers to y_i

Simple linear regression



$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

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

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

Exercise

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

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

- 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, dat)
```

Fit regression model

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

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

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

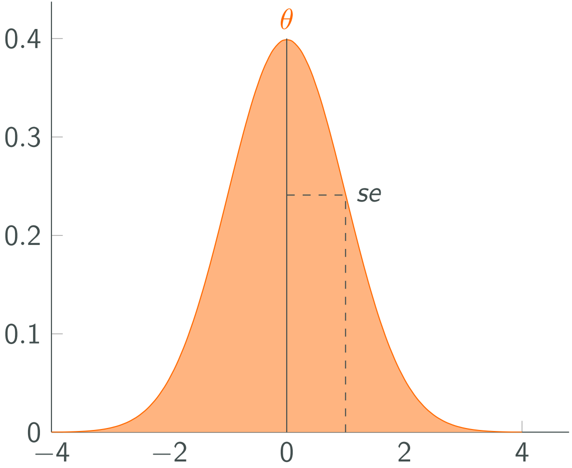
Re-cover parameters

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

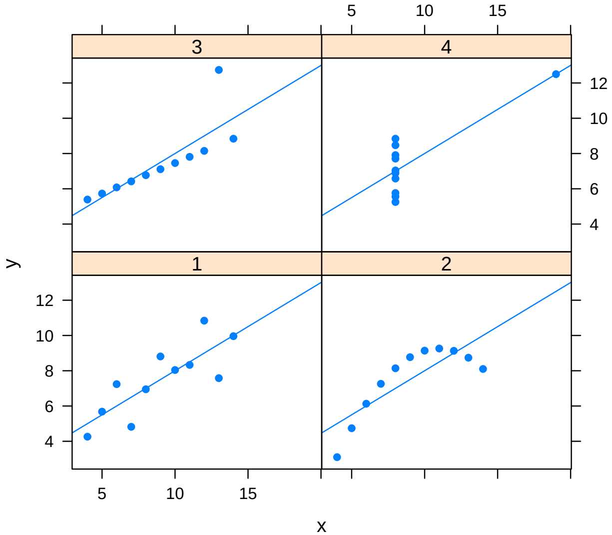


Exercise

- 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?

2 Assumptions

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

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

  and we get the stochastical model

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

  which transfers to

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

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

Example: Multiple linear regression

- We are fitting the following model

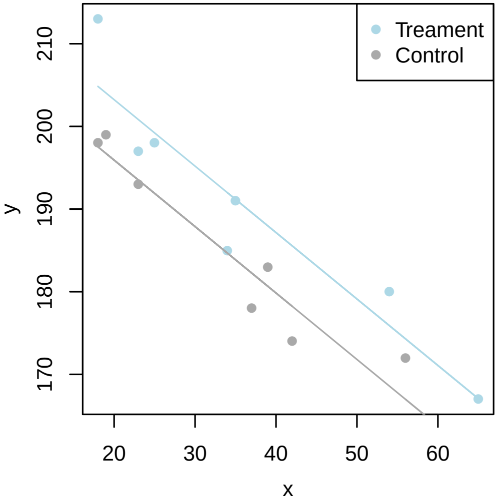
$$y_{ij} = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot z_j + \varepsilon_{ij}$$

with $i = 1 \dots N$ 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.