Simple linear regression

Nora Wickelmaier

October 24, 2022

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 $i = 1, \ldots, n$

Simple linear regression

- From the properties of the error variables, we conclude

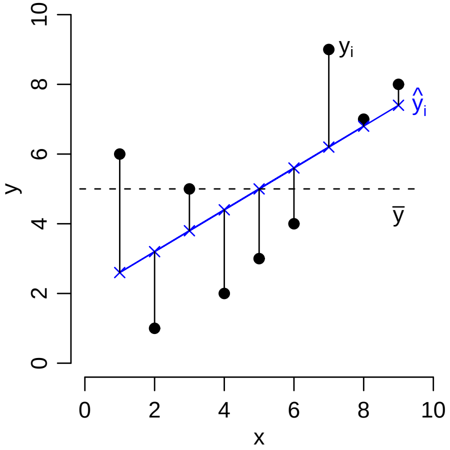
$$E(y_i) = E(\beta_0 + \beta_1 \cdot x_i + \varepsilon_i) = \beta_0 + \beta_1 \cdot x_i = \bar{y}$$

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

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

- 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, 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)
```

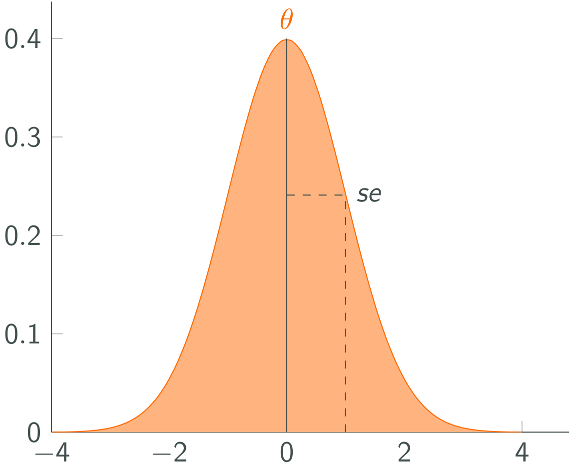
Re-cover parameters

```
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

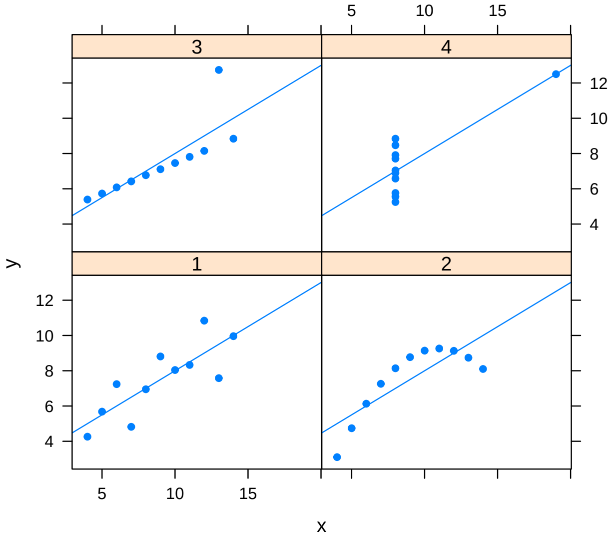


Exercise

- 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

```r
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)
```

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	$y = \beta_0 + \beta_1 x_1 + \beta_2 time + \upsilon_0 + \upsilon_1 time + \varepsilon$

...

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.