

Simple Linear Regression and the Method of Least Squares

Author: Nicholas G Reich, Jeff Goldsmith

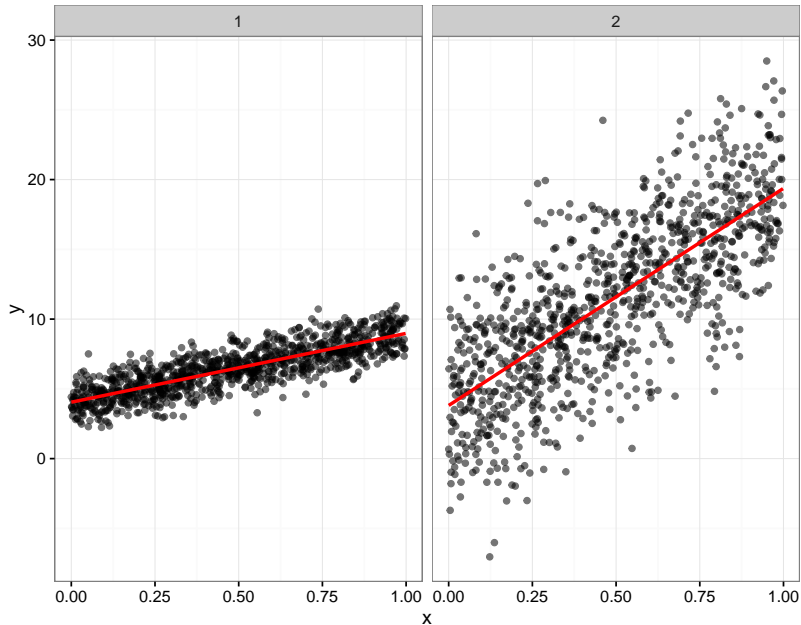
*This material is part of the **statsTeachR** project*

Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US



Figure acknowledgements to [Hadley Wickham](#).

Which data show a stronger association?



Goals for this class

You should be able to...

- interpret regression coefficients.
- derive estimators for SLR coefficients.
- implement a SLR from scratch (i.e. not using `lm()`).
- explain why some points have more influence than others on the fitted line.

Regression modeling

- Want to use predictors to learn about the outcome distribution, particularly conditional expected value.
- Formulate the problem parametrically

$$E(y \mid x) = f(x; \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

- (Note that other useful quantities, like covariance and correlation, tell you about the joint distribution of y and x)

Brief Detour: Covariance and Correlation

$$\text{cov}(x, y) = \mathbb{E}[(x - \mu_x)(y - \mu_y)]$$

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}$$

Simple linear regression

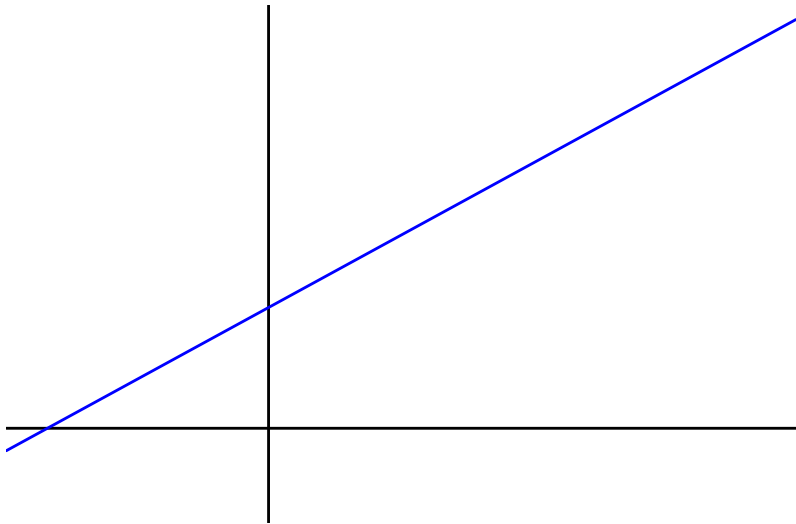
- Linear models are a special case of all regression models; simple linear regression is the simplest place to start
- Only one predictor:

$$E(y \mid x) = f(x; \beta) = \beta_0 + \beta_1 x_1$$

- Useful to note that $x_0 = 1$ (implicit definition)
- Somehow, estimate β_0, β_1 using observed data.

Coefficient interpretation

Coefficient interpretation



Step 1: Always look at the data!

- Plot the data using, e.g. the `plot()` or `qqplot()` functions
- Do the data look like the assumed model?
- Should you be concerned about outliers?
- Define what you expect to see before fitting any model.

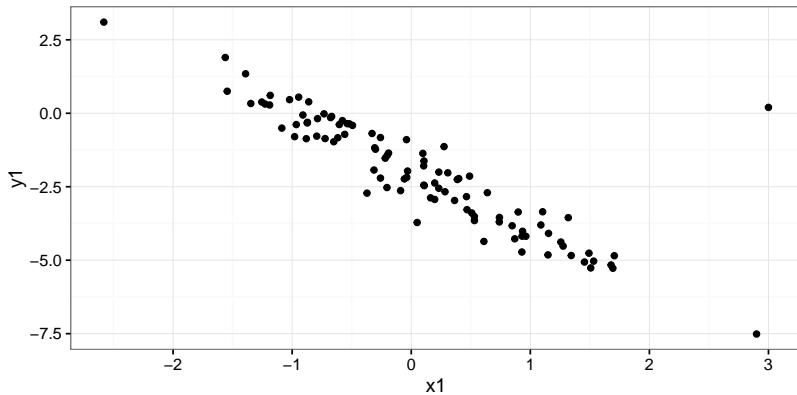
Least squares estimation

- Observe data (y_i, x_i) for subjects $1, \dots, I$. Want to estimate β_0, β_1 in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \stackrel{iid}{\sim} (0, \sigma^2)$$

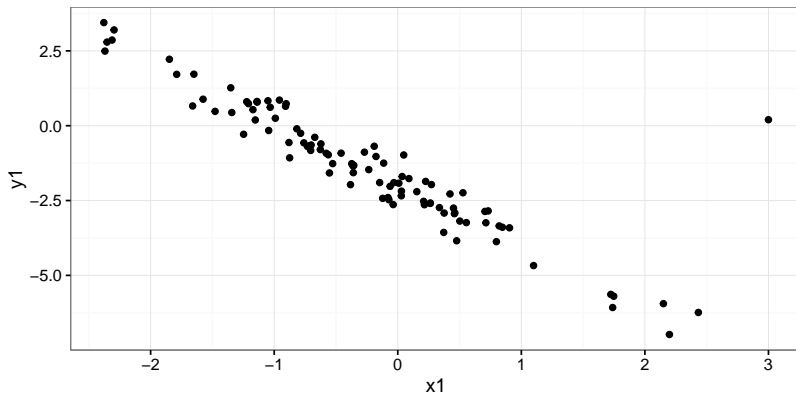
- Recall the assumptions:
 - A1: The model: e.g. $y_i = f(x_i; \beta) + \epsilon_i = \beta_0 + \beta_1 x_{i,1} + \epsilon_i$
 - A2: Unbiased errors: $\mathbb{E}[\epsilon_i | x_i] = \mathbb{E}[\epsilon_i] = 0$
 - A3: Uncorrelated errors: $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
 - A4: Constant variance: $\text{Var}[y_i | x_i] = \sigma^2$
 - A5: Probability distribution: e.g. $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$
[not needed for LS, is needed for inference].
 - A6: Representative sampling: generalize to population.

Any violations of assumptions? (Ex. 1)

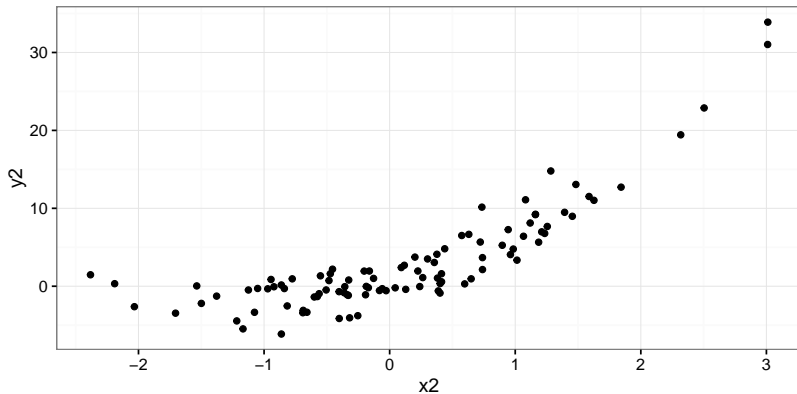


Any violations of assumptions? (Ex. 1)

```
x1 <- rnorm(100)
y1 = -2-2*x1 + rnorm(100, 0, .5)
x1[1] <- 3; y1[1] <- .2
qpplot(x1, y1)
```

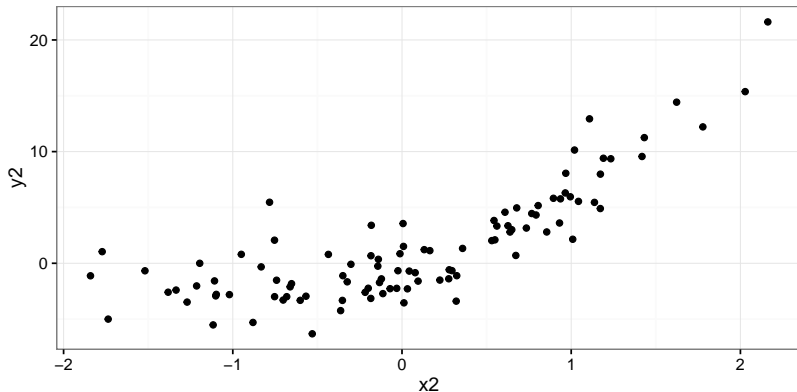


Any violations of assumptions? (Ex. 2)

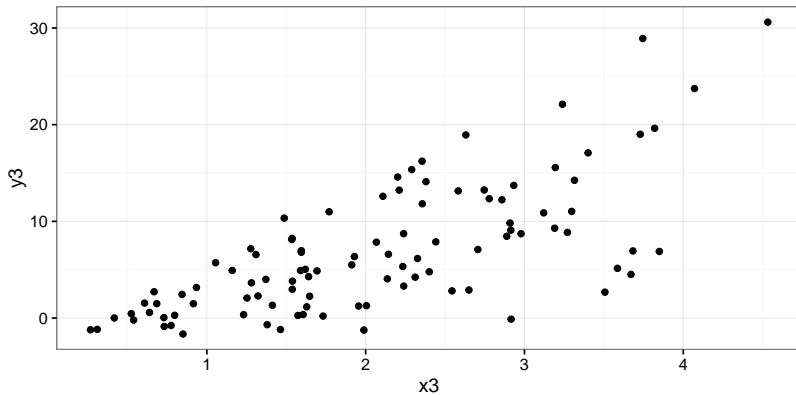


Any violations of assumptions? (Ex. 2)

```
x2 <- rnorm(100)
y2 = -2+2*(x2+1)^2 + rnorm(100, 0, 2)
qplot(x2, y2)
```

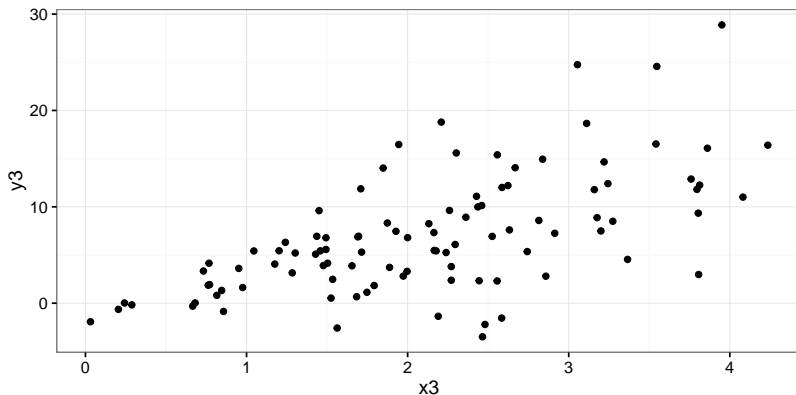


Any violations of assumptions? (Ex. 3)



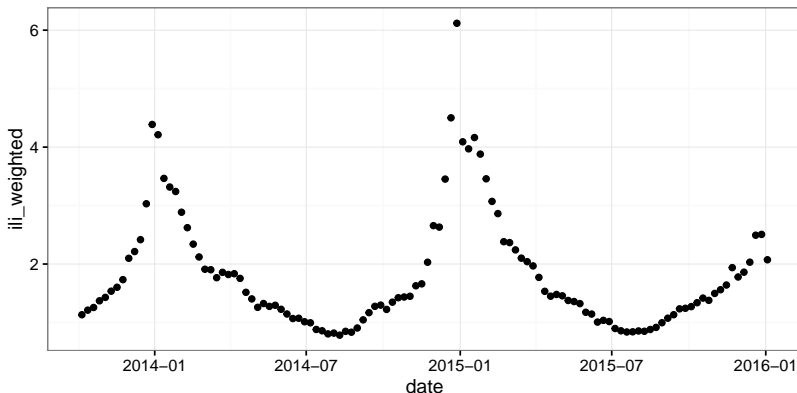
Any violations of assumptions? (Ex. 3)

```
x3 <- abs(rnorm(100, mean=2))  
y3 = -2+4*x3 + rnorm(100, 0, x3*2)  
qplot(x3, y3)
```



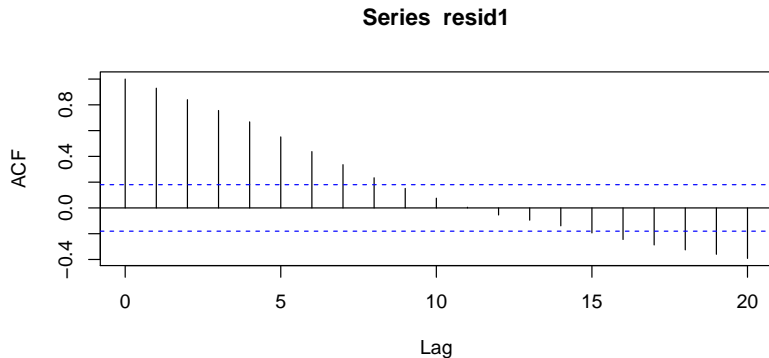
Any violations of assumptions? (Ex. 4)

```
library(cdcfluvview)
library(dplyr)
usflu <- get_flu_data("national", "ilinet", years=2013:2015)
usflu <- mutate(usflu,
                 date = as.Date(paste0(YEAR, sprintf("%02d", WEEK), "00"),
                                format="%Y%W%w"),
                 ili_weighted = X.UNWEIGHTED.ILI)
ggplot(usflu, aes(x=date, y=ili_weighted)) + geom_point()
```

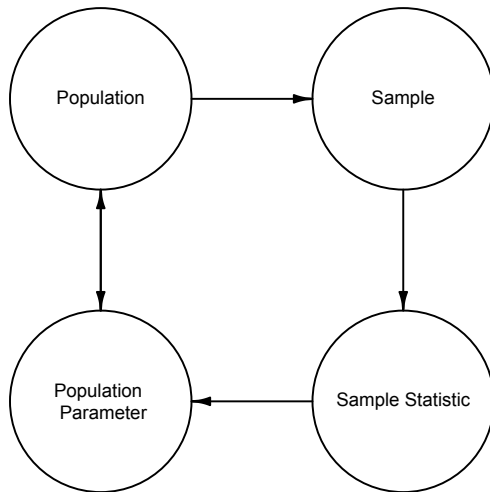


Any violations of assumptions? (Ex. 4)

```
fm1 <- lm(ili_weighted~date, data=usflu)  
resid1 <- resid(fm1)  
acf(resid1)
```



Circle of Life



Least squares estimation

- Recall that for a single sample $y_i, i \in 1, \dots, N$, the sample mean $\hat{\mu}_y$ minimizes the sum of squared deviations.

$$RSS(\mu_y) = \sum_{i=1}^N (y_i - \mu_y)^2$$

Least squares estimation

Find $\hat{\beta}_0$ and β_1 . By minimizing RSS relative to each parameter.

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^N (y_i - \mathbb{E}[y_i|x_i])^2$$

We obtain

$$\begin{aligned}\hat{\beta}_0 &= b_0 = \bar{y} - b_1 \bar{x} \\ \hat{\beta}_1 &= b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\end{aligned}$$

Notes about LSE

Relationship between correlation and slope

$$\rho = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}; \quad \beta_1 = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

Why we need to keep watch for outliers

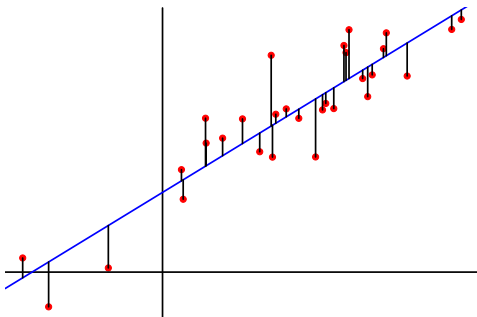
$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sum \frac{y_i - \bar{y}}{x_i - \bar{x}} (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \\ &= \sum \frac{y_i - \bar{y}}{x_i - \bar{x}} \omega_i\end{aligned}$$

Note that weight ω_i increases as x_i gets further away from \bar{x} .

Geometric interpretation of least squares

Least squares minimizes the sum of squared vertical distances between observed and estimated y 's:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^I (y_i - (\beta_0 + \beta_1 x_i))^2$$



Least squares foreshadowing

- Didn't have to choose to minimize squares – could minimize absolute value, for instance.
- Least squares estimates turn out to be a “good idea” – unbiased, BLUE (Best Linear Unbiased Estimator).
- Later we'll see about maximum likelihood as well.

Lab exercise: computing $\hat{\beta}$ on your own

- Load the heights data from lecture 1.
- Run a linear model using the R function `lm()`, with daughter height as the outcome.
- Compare the results of that regression with hand-calculated $\hat{\beta}_0$ and $\hat{\beta}_1$ coefficients.

```
# sample code  
install.packages("alr3")  
library(alr3)  
data(heights)  
fm1 <- lm(Dheight ~ Mheight, data=heights)  
summary(fm1)
```