Hôpital général juif
Jewish General Hospital

McGill

INSTITUT LADY DAVIS DE RECHERCHES MÉDICALES / LADY DAVIS INSTITUTE FOR MEDICAL RESEARCH

LDI Seminar Series in Biostatistics: Lecture 3

Kevin McGregor

October 2, 2017

Info

- Reminder: No seminar next week (Oct 9th). Next one is Monday, Oct 16th.
- Slides and tutorial available on my Github page
- `http://github.com/kevinmcgregor/LDI-Biostatistics-Seminar`
- Email: kevin.mcgregor@mail.mcgill.ca

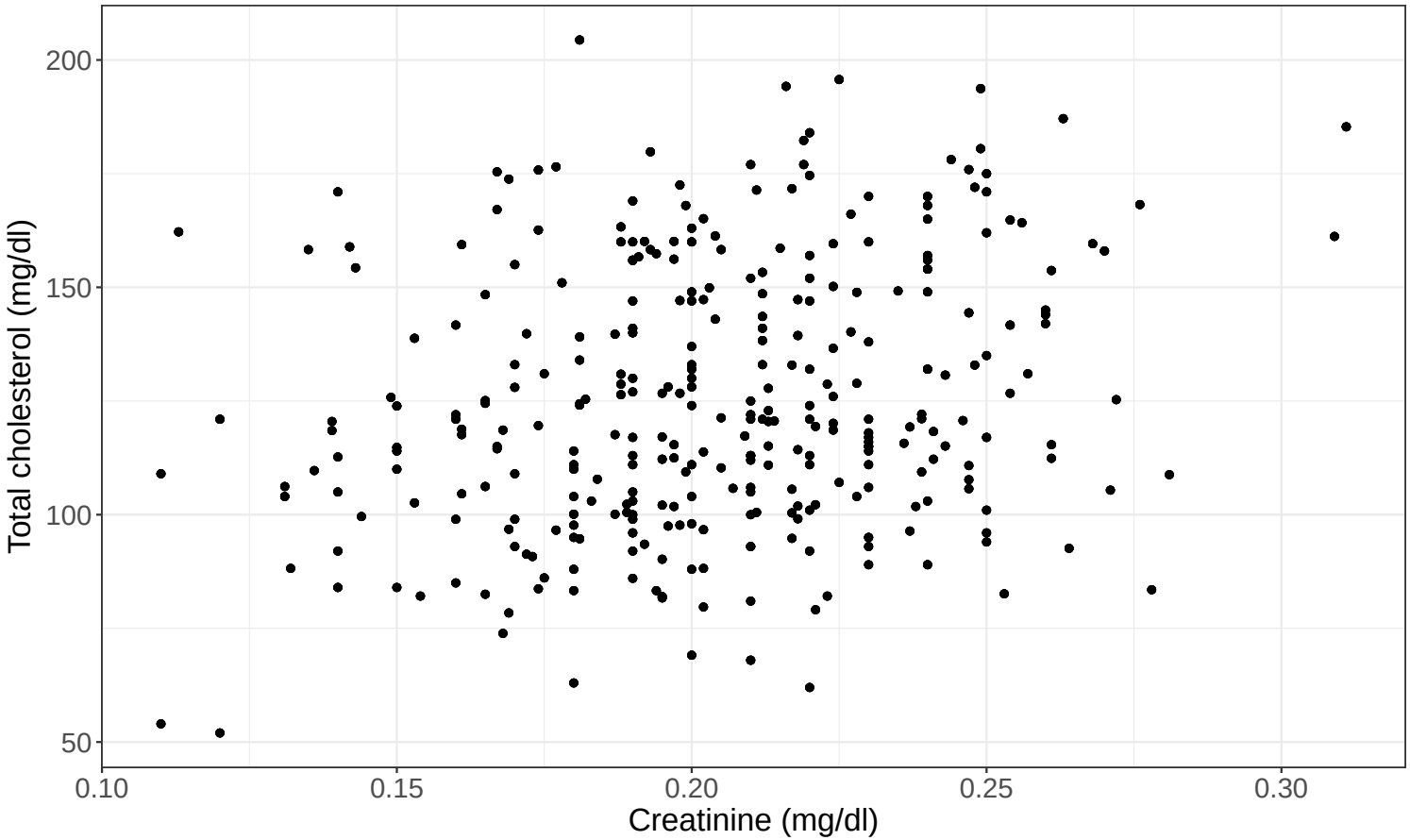
Last time

- Defined hypothesis testing
- Looked at the distribution of the sample mean
  - Interpretation of the standard error of the mean $\frac{s}{\sqrt{n}}$
- One-sample t-test.
- Independent-samples t-test.

Linear regression

- Today: Linear regression
- Allows us to study the effect of one or more continuous or categorial variables on the *mean* of a continuous variable.
- Many statistical methods are special cases of linear regression: one-sample t-test, independent-samples t-test, ANOVA.
- Needs a little bit more care when considering assumptions compared to the t-test.

Scatterplot



Correlation

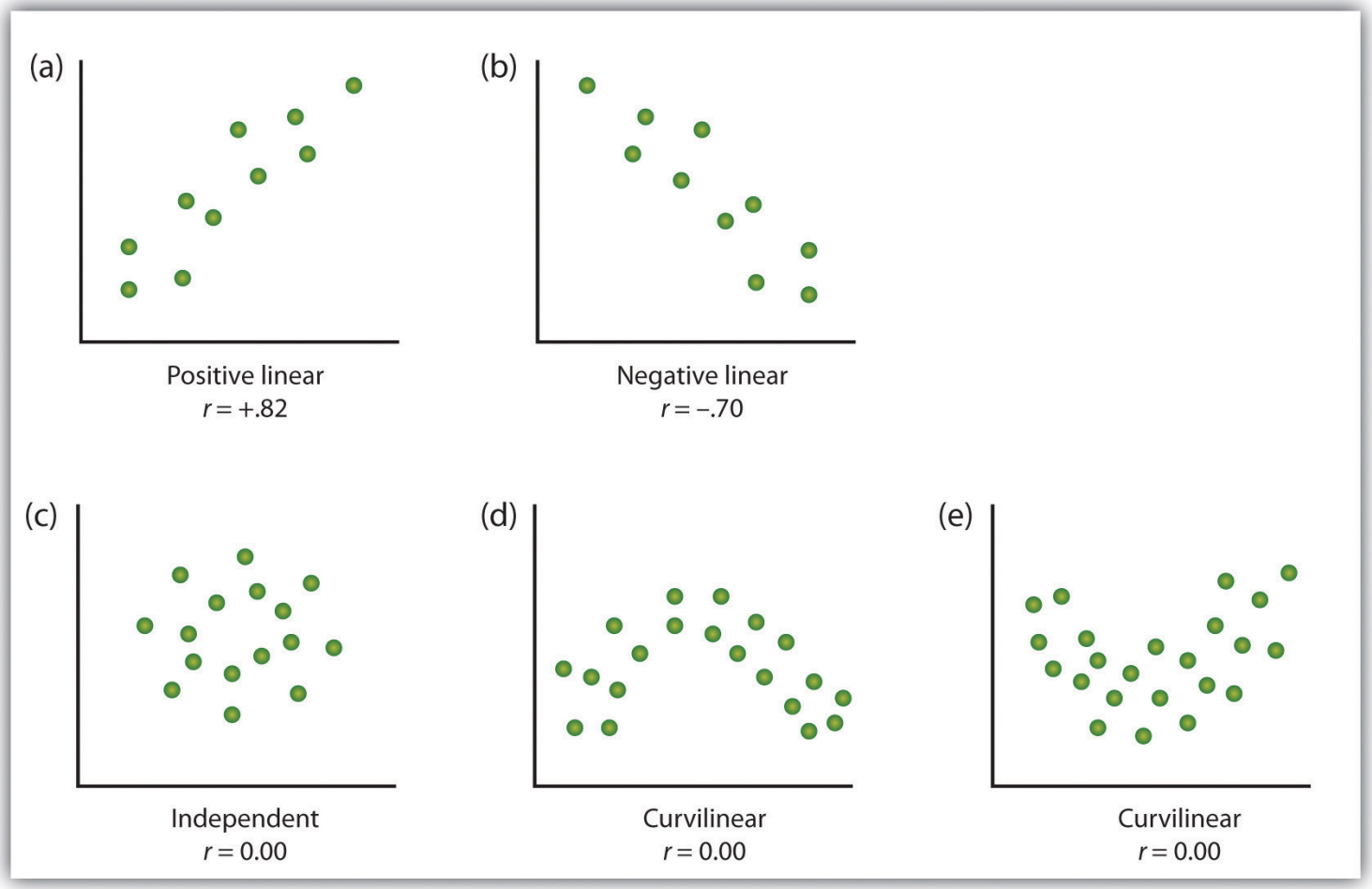
- Popular method of association between two continuous variables: *Pearson's correlation coefficient*.
- Assume n samples from two variables $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$.
- The sample correlation coefficient, denoted by r, is defined as:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Interpretation of correlation coefficient

- The correlation coefficient is bounded: $-1 \leq r \leq 1$.
  - If $r > 0$, then the variables are positively associated
  - If $r < 0$, then the variables are negatively associated
  - If $r = 0$, then there is no apparent linear relationship between the variables
- Correlation only explores whether there is a *linear* relationship between two variables. $r = 1$ and $r = -1$ imply a perfect positive or negative linear relationship between the two variables, respectively.

Zero correlation



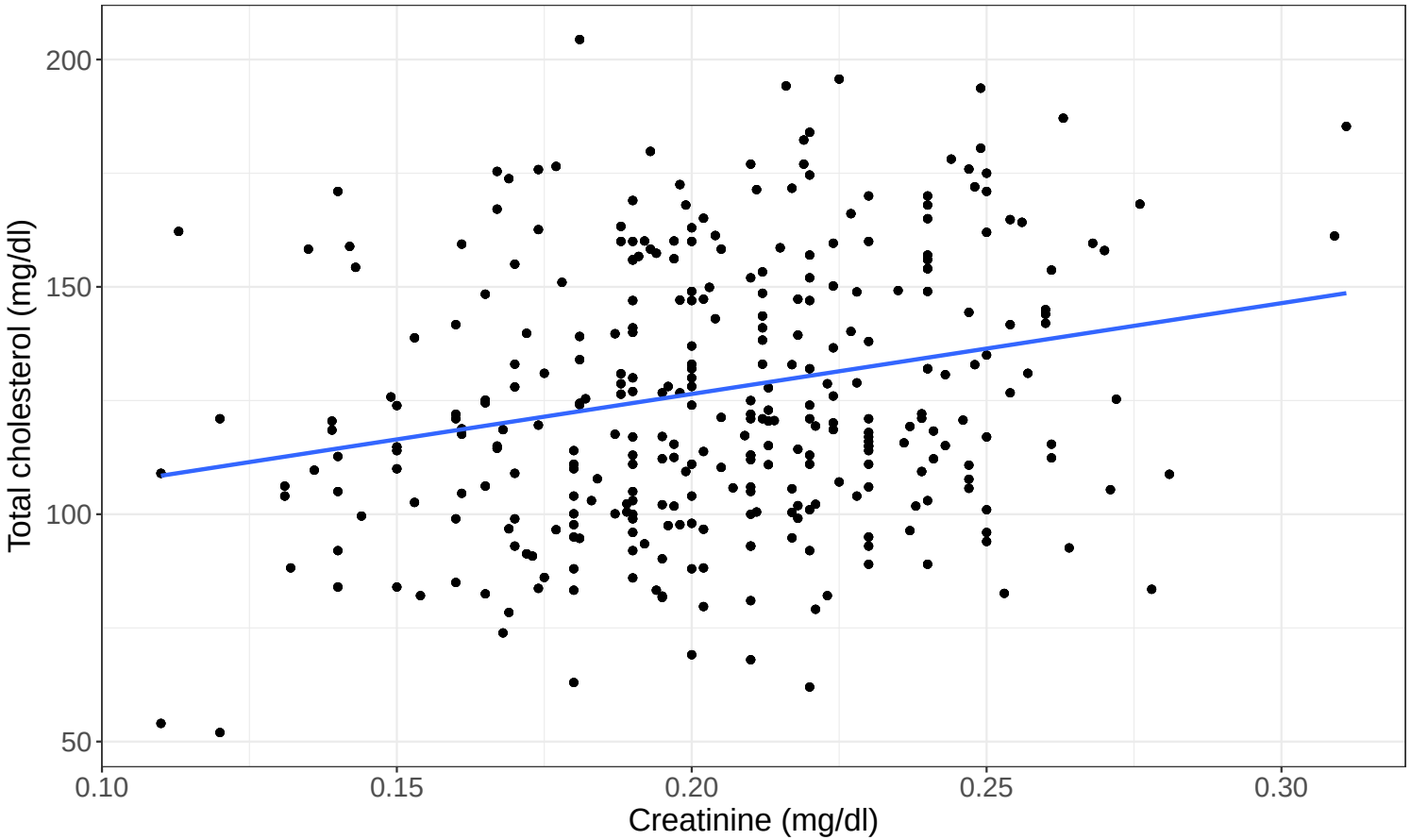
Zero correlation does not necessarily mean independence!

Why not always use correlation?

- Correlations can be compared between different pairs of variables (since it's bounded)
- Gives no information on effect size, i.e. the amount of change in one variable associated with change in the other.
- Simple linear regression allows us to estimate this quantity.

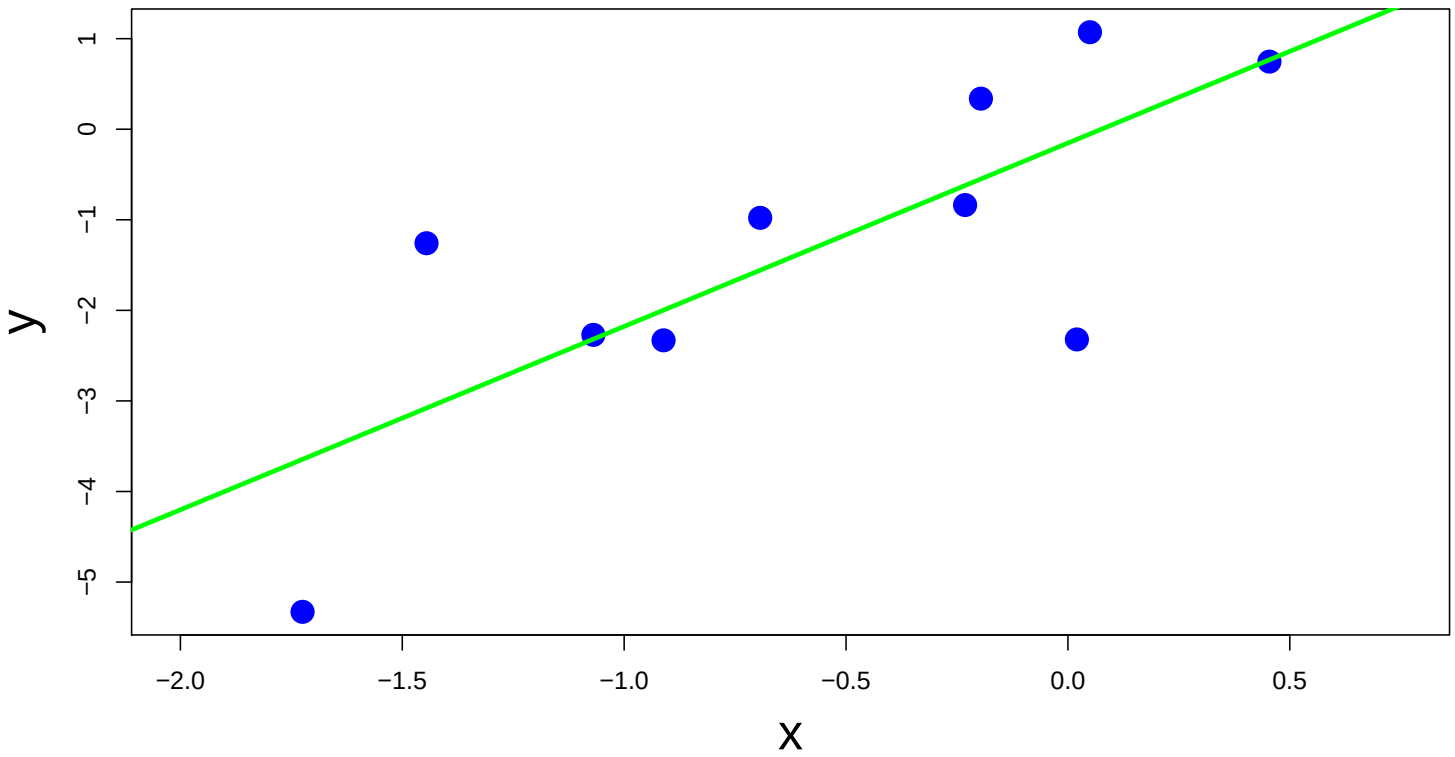
Simple linear regression

Scatterplot with best fit line



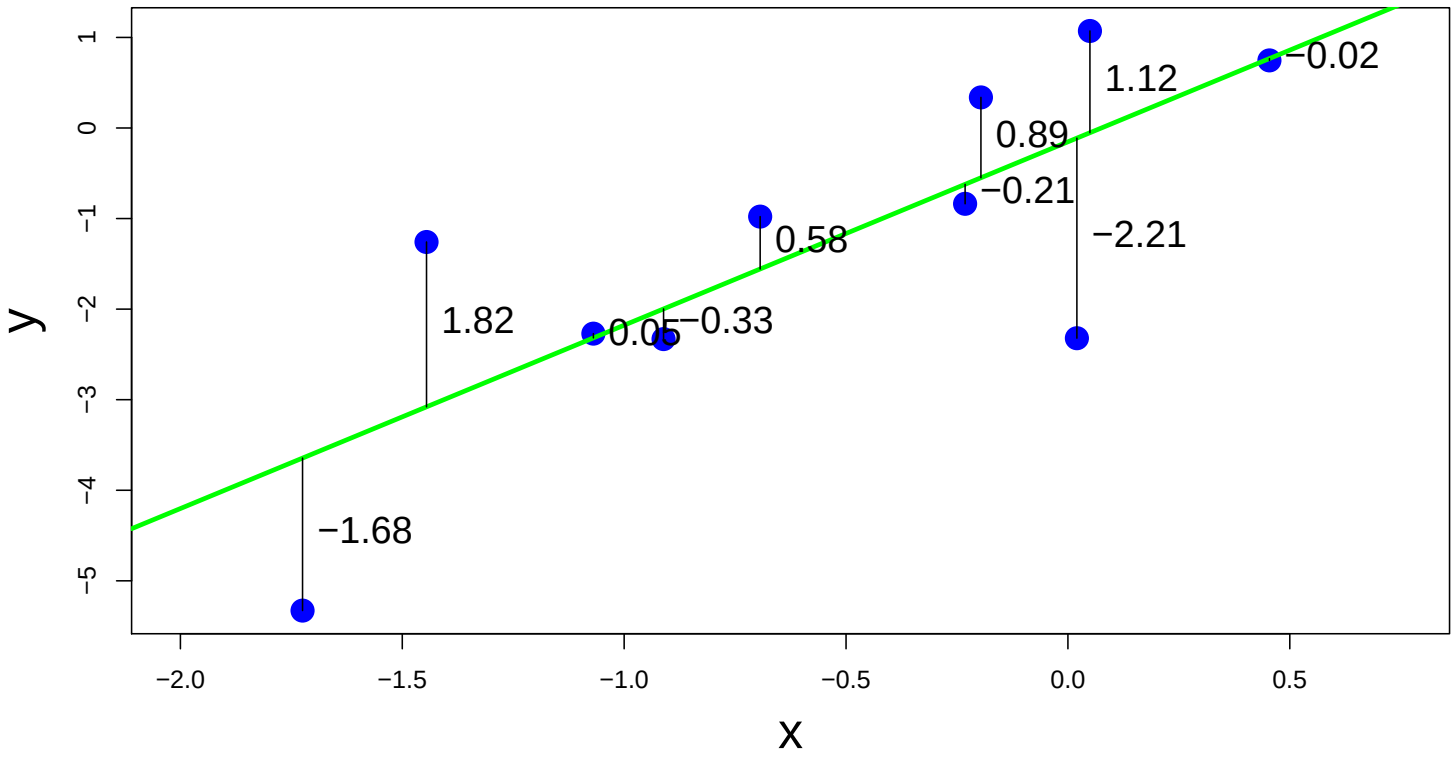
Want to find a "best fit" line to the data. The slope of the line is usually the parameter of interest.

How to find best fit line?



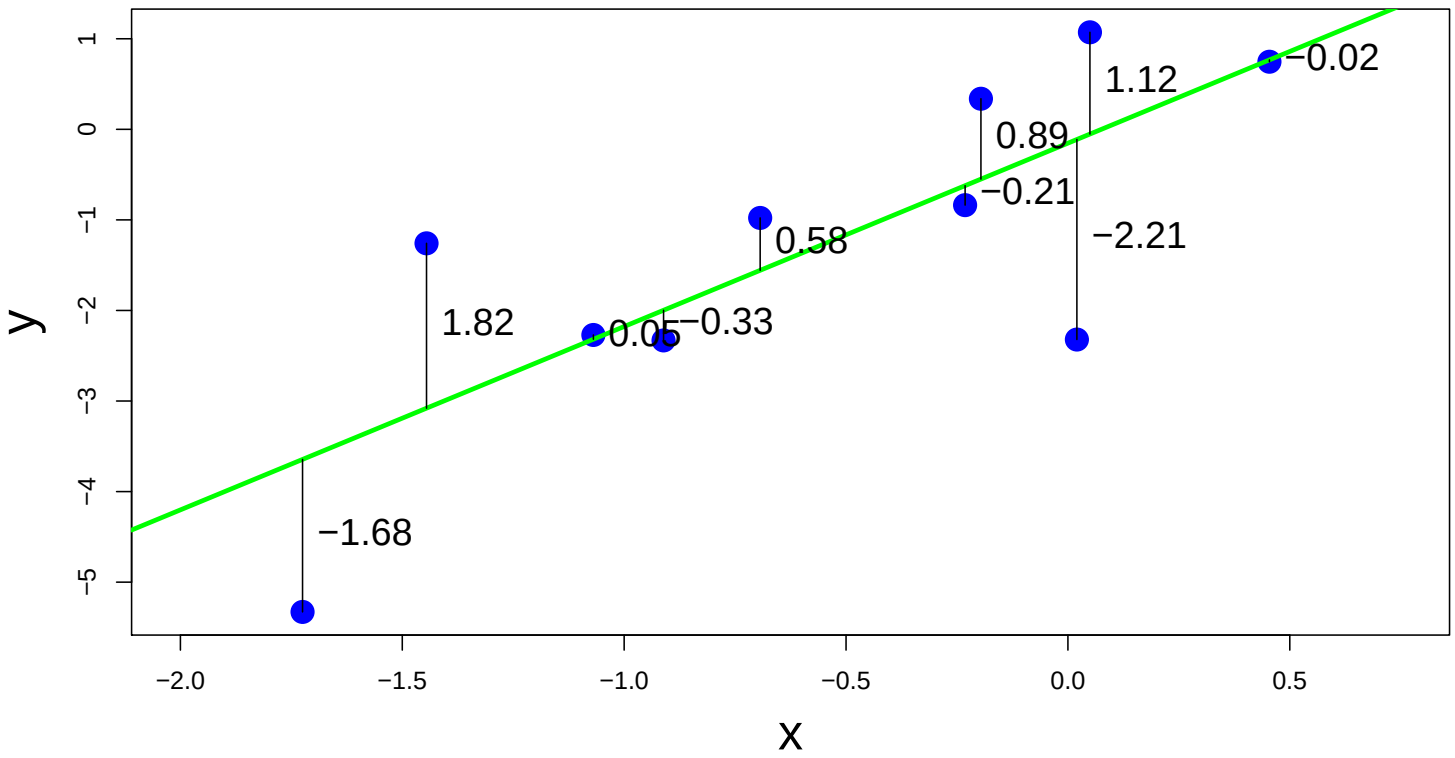
How do we find a best fit line? Depends on how we define "best" fit.

Residuals



Recall the residuals (vertical distance between the true point and the fitted line). Assume x values are fixed (no measurement error).

Squared residuals



Consider the sum squared residuals: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the value falling on the line directly above or below y_i.

Least-squares estimator

- Turns out there is a nice solution to this problem, called the *least-squares* estimator.
- Assume that the regression line takes on the form:

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

- b_1 estimates the slope of the line... or the effect of the *predictor* variable x on the *outcome* variable y.
- We also assume that each point has an error term ϵ_i which follows a normal distribution with mean 0 and standard deviation σ.
  - Contains additional variation in y not explained by the predictor variable x.

Least-squares estimator

- The least squares estimator for b_1 is then:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- The estimator for b_0 is:

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}$$

- We can calculate a "fitted" value for each individual:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

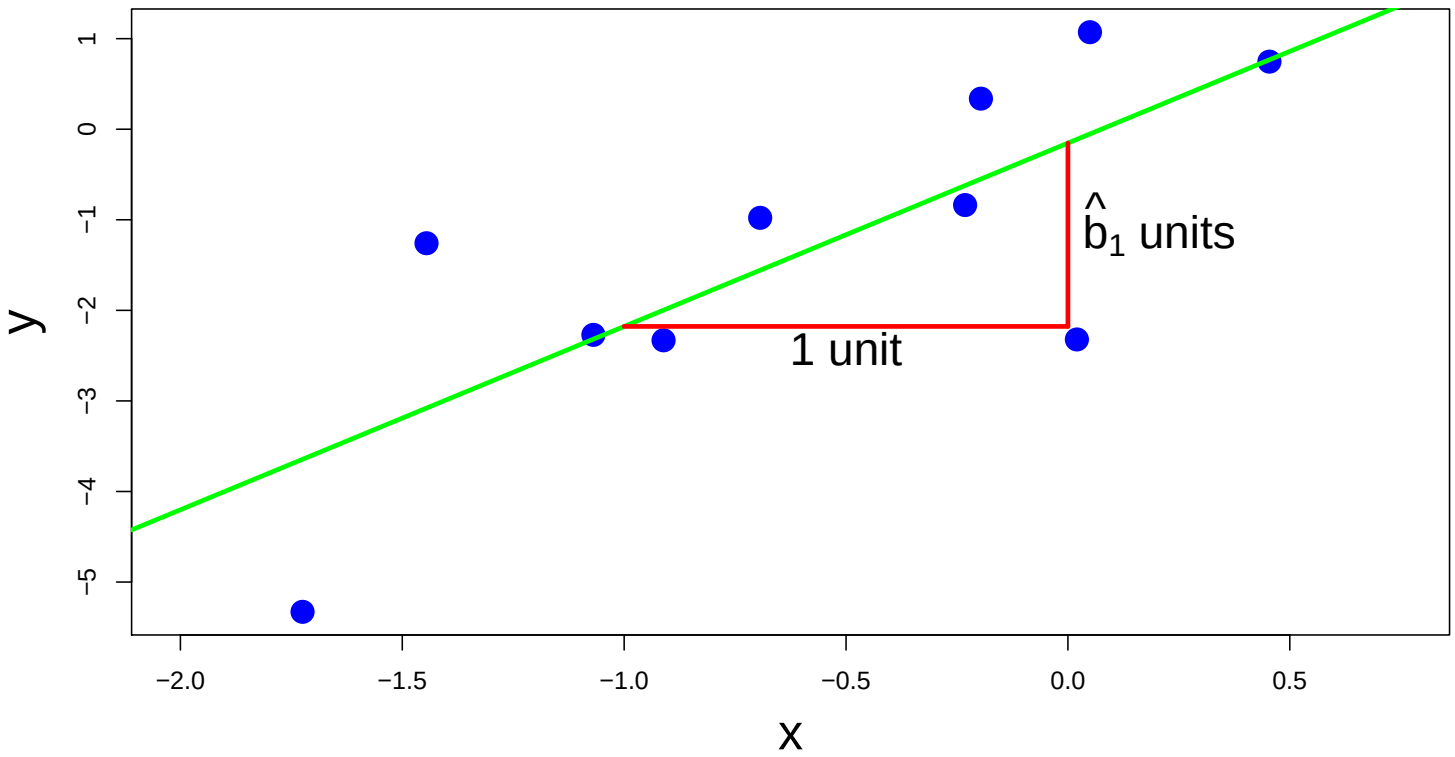
Least-squares estimator

- b_1 is usually the parameter of interest.
- Consider two individuals whose x values differ by exactly one unit (say, $x_1 = 5$ and $x_2 = 6$).
- Taking the difference of the two fitted values for these individuals gives:

$$\begin{aligned}\hat{y}_2 - \hat{y}_1 &= (\hat{b}_0 + \hat{b}_1 x_2) - (\hat{b}_0 + \hat{b}_1 x_1) \\ &= (\hat{b}_0 + \hat{b}_1 \cdot 6) - (\hat{b}_0 + \hat{b}_1 \cdot 5) \\ &= \hat{b}_0 - \hat{b}_0 + \hat{b}_1(6 - 5) \\ &= \hat{b}_1\end{aligned}$$

- The value $\hat{b}_1$ is the estimated change in y associated with an increase in *one unit* of x.

Coefficient visualization



Visual interpretation of $\hat{b}_1$.

Statistical test for b_1

- Can test b_1 to check for significant **linear** relationship between the two variables:

  $H_0 : b_1 = 0$
  $H_1 : b_1 \neq 0$

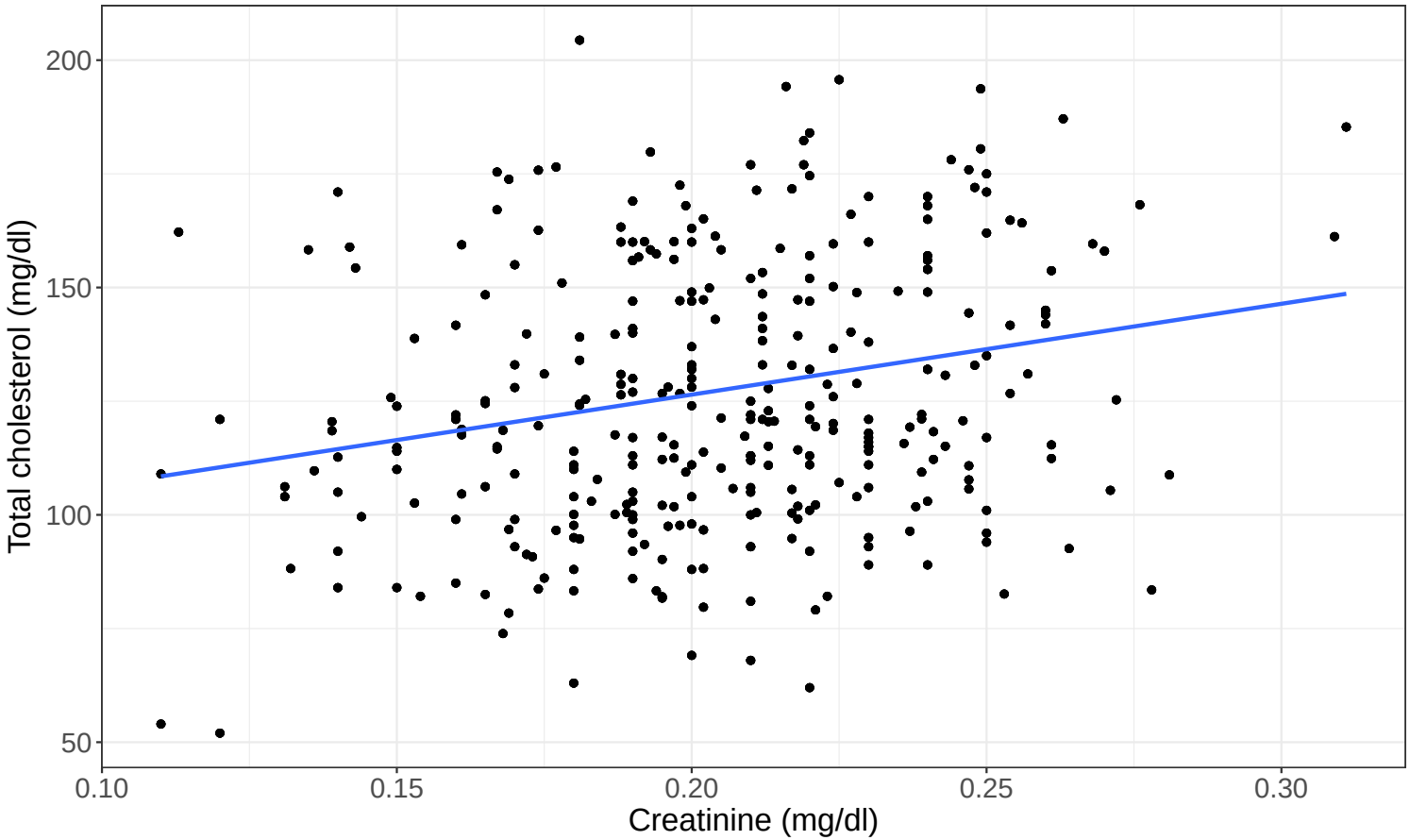
- Have to consider the standard error of $\hat{b}_1$:

$$SE(\hat{b}_1) = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Statistical test for b_1

- Can also use a t-test for this (under normality assumption, will come back to this)
- The test statistic is $t = \frac{\hat{b}_1}{SE(\hat{b}_1)}$
- Testing at significance level α: compare $|t|$ to the critical value $t_{\alpha,n-2}$
  - If $|t| > t_{\alpha,n-2}$, then reject H_0.
  - Otherwise, do not reject H_0.
- Can calculate p-values and confidence intervals in the same way as we did in the one-sample t-test.

IMPC data example



Example: $\hat{b}_1 = 199.74$. Means increase in 0.1 mg/dl of creatinine corresponds to an *average* increase in 19.974 mg/dl of total cholesterol.

IMPC data example significance

- In our example $\hat{b}_1 = 199.74$, $SE(\hat{b}_1) = 21.69$. Sample size is $n = 1471$. Test at level $\alpha = 0.05$.
- Calculating the t-statistic gives:

$$\begin{aligned} t &= \frac{\hat{b}_1}{SE(\hat{b}_1)} \\ &= \frac{199.74}{21.69} \\ &= 9.21 \end{aligned}$$

- $t_{0.05,1471-2} = 1.65$. Since $|t| > t_{0.05,1471-2}$, we reject H_0.
- $p < 10^{-16}$, and 95% confidence interval is (157.18, 242.29).

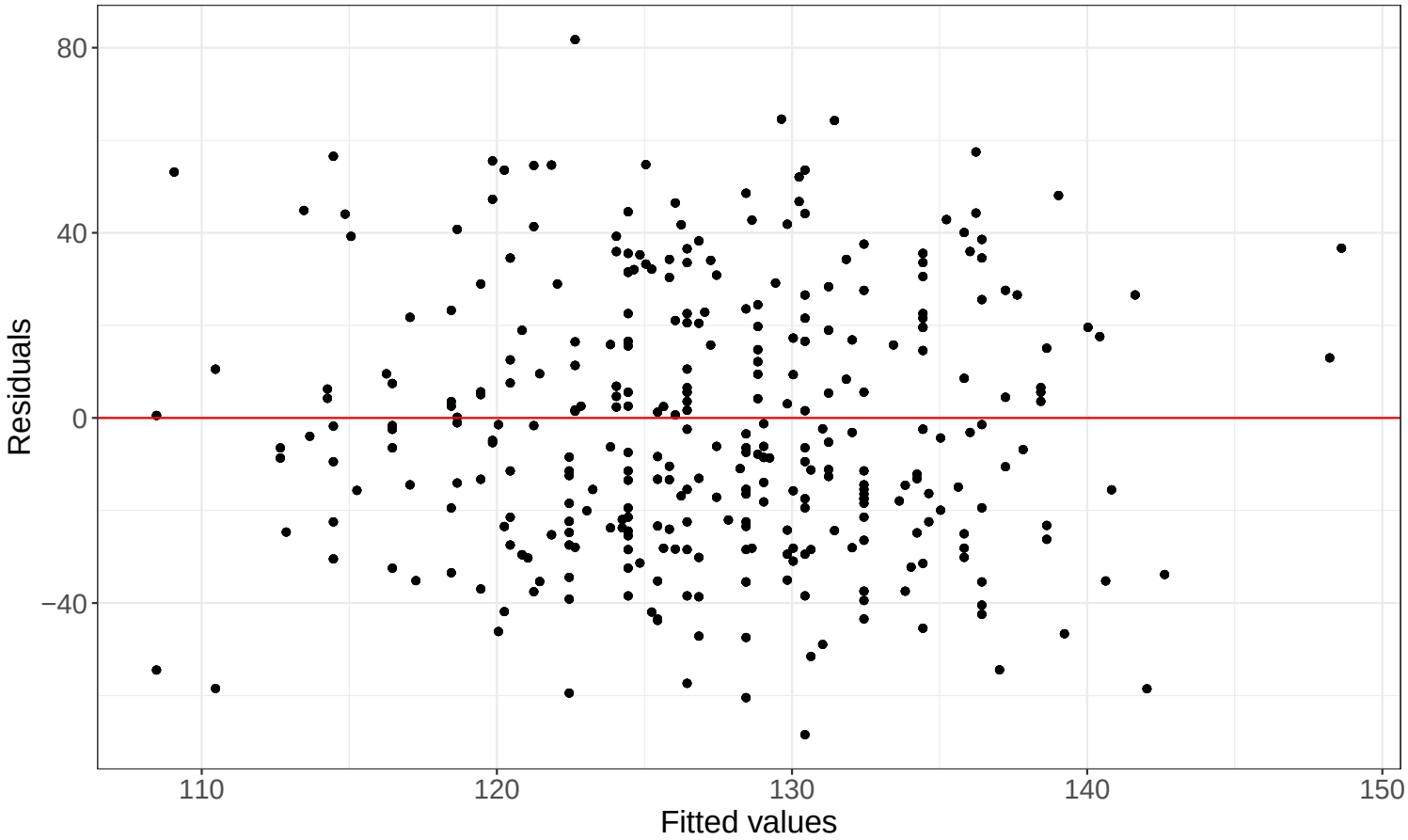
Assumptions in linear regression

- Assume an underlying linear relationship between the two variables x and y.
- Normality: Assume for a given value of x, that y follows a normal distribution.
- Independence of observations.
- Homoscedasticity: variance of y does not change over the values of x.

Residual plot

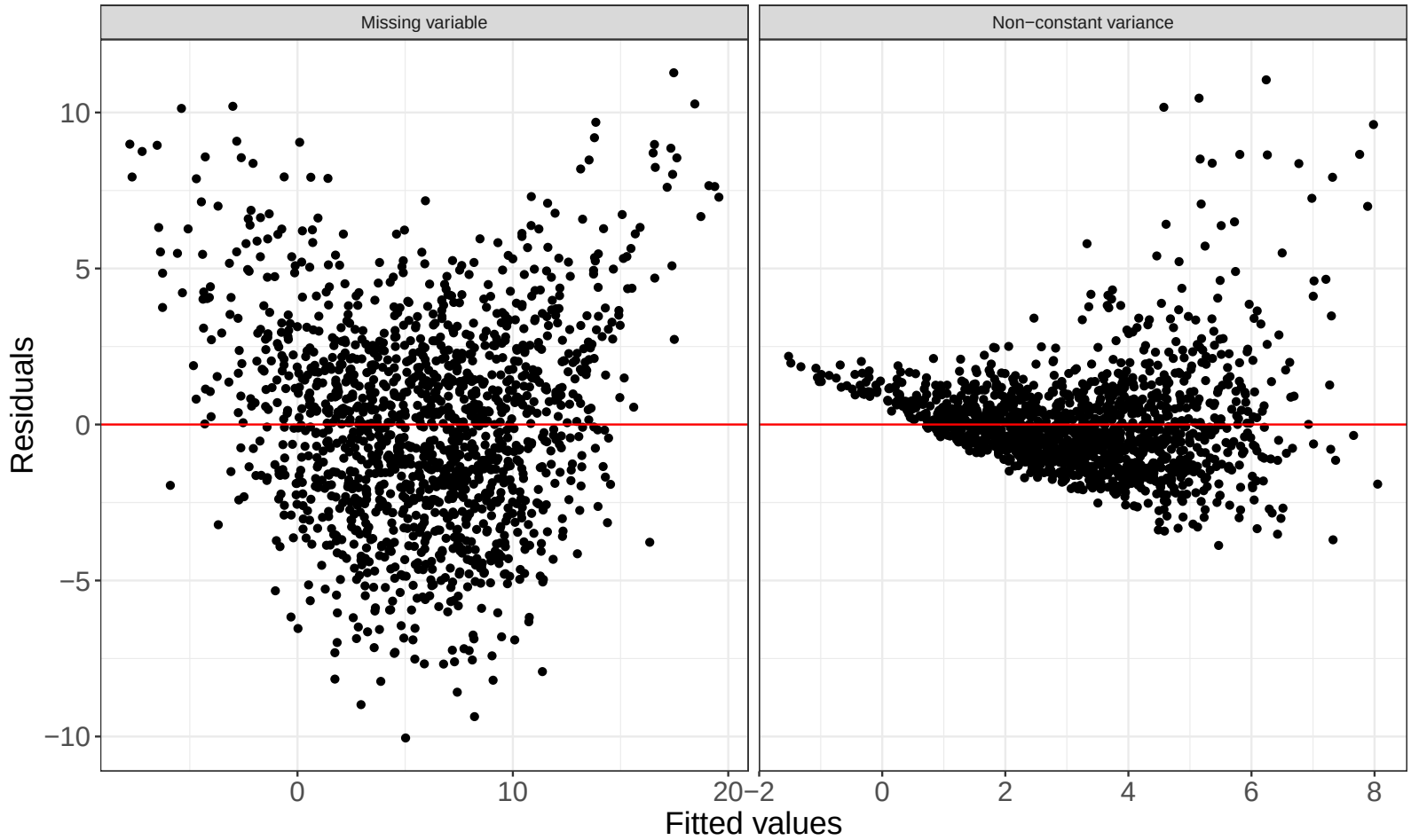
- Looking at residuals is an excellent way to check model assumptions
- Most basic tool: plotting the fitted values vs. the residuals
- Don't want to see any kind of discernible pattern in the residual plot. Otherwise:
  - Could have non-constant variance
  - Could have important variables missing
- Can also check the distribution of the residuals to see if normal distribution assumption is met.

IMPC data residual plot



Check variance of residuals over the fitted values. In this example, there is no discernible pattern.

Bad residual plots

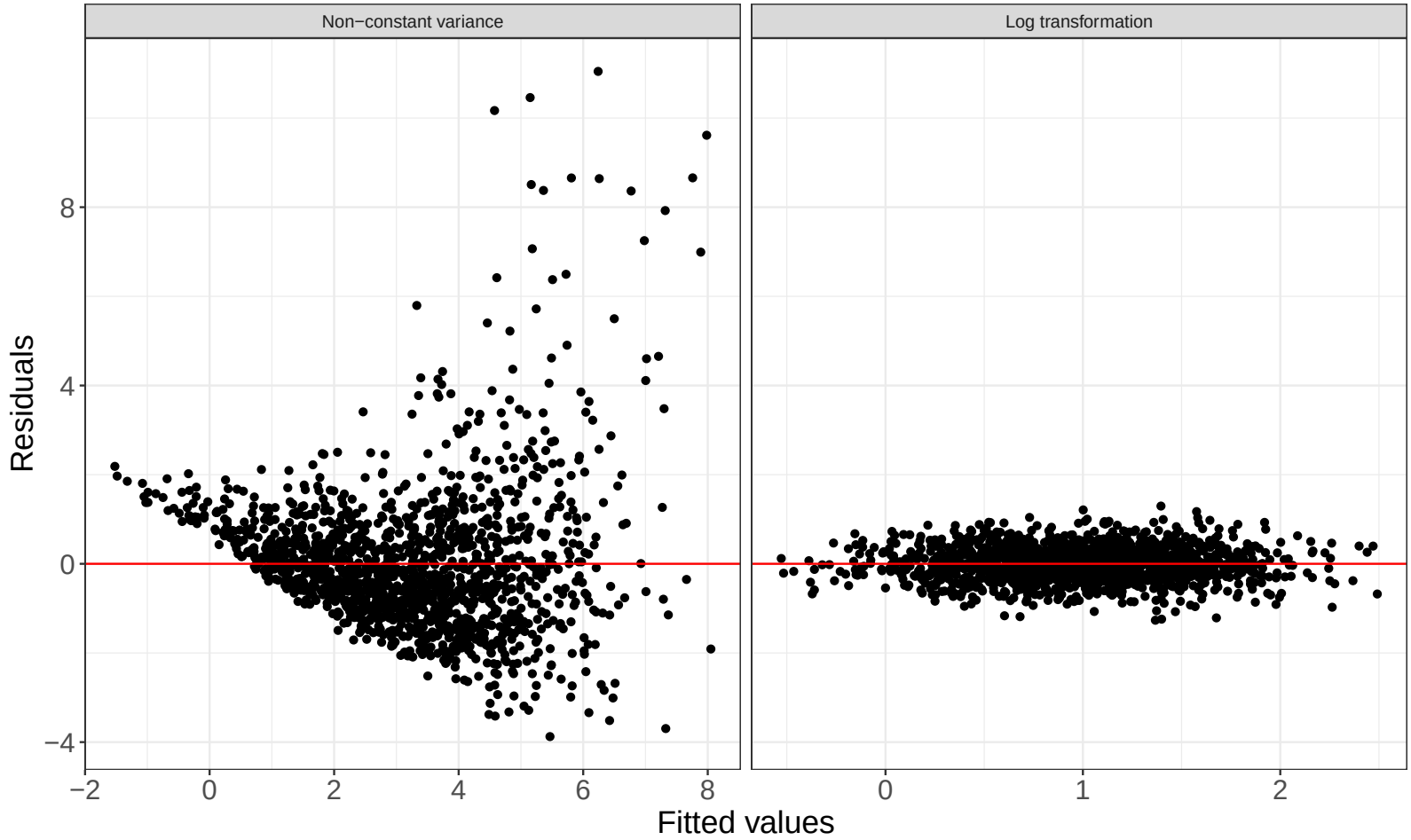


Two examples of bad residual plots (simulated data).

Action for bad residual plot

- Sometimes have to experiment a little.
- Can add extra variables to model (multiple linear regression).
- Could add additional higher order terms to model.
  - E.g. If age is in the predictor, could add age^2 as an additional predictor in the model.
- Could do transformations on response variable to get constant variance: log, square root.
  - Careful... this changes the interpretation of b_1.
- **Don't look at the p-value during this process!**

Fixed residual plot



Original residual plot on left. Fixed residual plot corresponding to log-transformed response variable on right.

Multiple regression

Multiple regression formulation

- Multiple regression is very similar to simple linear regression. But now there is more than one predictor variable (still a single response variable).
- E.g. if there were three predictor variables $x_1 = (x_{11}, x_{12}, \ldots, x_{1n})$, $x_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, $x_3 = (x_{31}, x_{32}, \ldots, x_{3n})$, then the regression model would be:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \epsilon_i$$

- Slope parameters b_1, b_2, and b_3 measure association between the predictors and y.
- All assumptions from before still present (linearity, normality, independence, constant variance)

Why use multiple regression?

- Could be interested in the joint effect of multiple variables on a single outcome variable.
  - Estimated effects are different than effects for a separate linear regression model run for each predictor.
- Even if only interested in one predictor and outcome, can included additional variables in model to "adjust" and therefore reduce potential bias.
  - E.g. in our total cholesterol vs. creatinine example, we could include mouse body weight as an additional variable in the model.
  - The estimated association between cholesterol and creatinine would then be adjusted for body weight.
- Including more variables in the model can often improve efficiency in estimates of association (smaller standard errors).

Interpretation of slopes

- Slope parameters b_1, b_2, and b_3 are of interest, but the interpretations are a bit different.
- $\hat{b}_1$ is the estimated increase in the response y associated with an increase in one unit of x_1 when *all other variables are held constant*.
- Likewise interpretations for b_2 and b_3
- Estimating the parameters in multiple linear regression is more complicated. Requires matrix algebra (therefore, I won't present the formulas).
- Standard errors are also a little bit more complicated and are omitted from this lecture.

Hypothesis testing

- Can do individual hypothesis tests for regression parameters. For each parameter $j \in \{1, 2, 3\}$:

  $H_0 : b_j = 0$
  $H_1 : b_j \neq 0$

- Testing the individual parameters once again results in t-tests. The test statstic is:

$$t = \frac{\hat{b}_j}{SE(\hat{b}_j)}$$

IMPC example

- Consider looking at the how total cholesterol changes with respect to creatinine, glucose, and body weight.
- Regression model:

$$y_i = b_0 + b_1 \text{creatinine} + b_2 \text{glucose} + b_3 \text{weight}$$

	Estimate	Std. Error	t-value	p-value
creatinine	67.91	19.71	3.45	5.86×10^{-4}
glucose	0.14	0.01	12.49	4.10×10^{-34}
weight	3.44	0.14	24.74	3.32×10^{-113}

Fitted model

- Can rewrite the fitted model as:

$$\hat{y}_i = -25.40 + 67.91 \times \text{creatinine}_i + 0.14 \times \text{glucose}_i + 3.44 \times \text{weight}_i$$

Take-home message

- Regression is a very powerful and versatile tool.
- Do a thorough investigation of model assumptions.
- Many assumptions to make, but lots of other models exist if assumptions are not met
  - Non-constant variance: Weighted least squares
  - Non-normal data: generalized linear models
  - Observations not independent: random effects models
- A lot of room for choosing models. Make your choice based on good statistical principles... not on the resulting p-values!

Thank you! - Merci!

Questions?