Photo credit: Dave DiCello

17-803 Empirical Methods

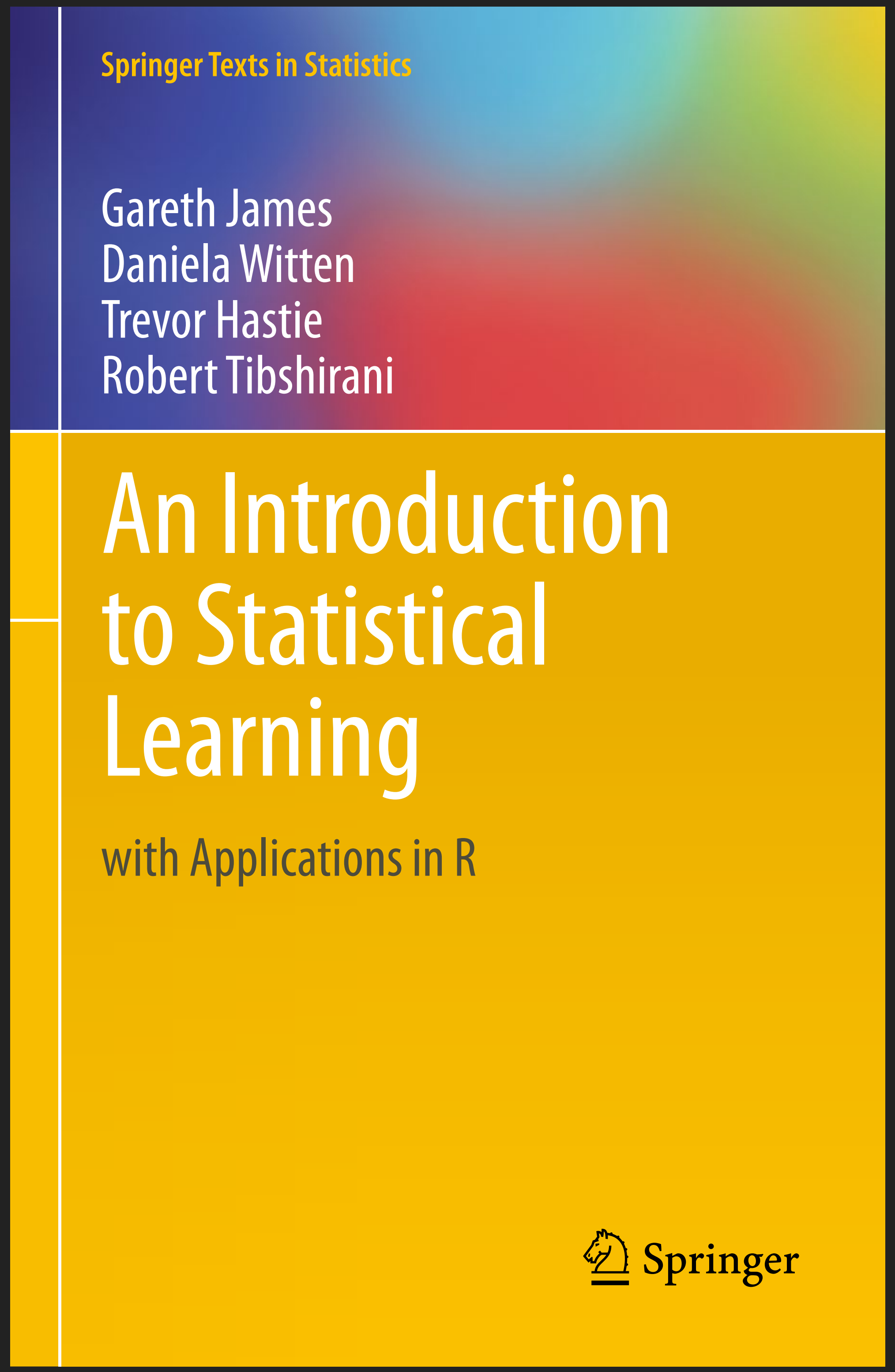
Bogdan Vasilescu, Institute for Software Research

Regression Modeling (Part 1)

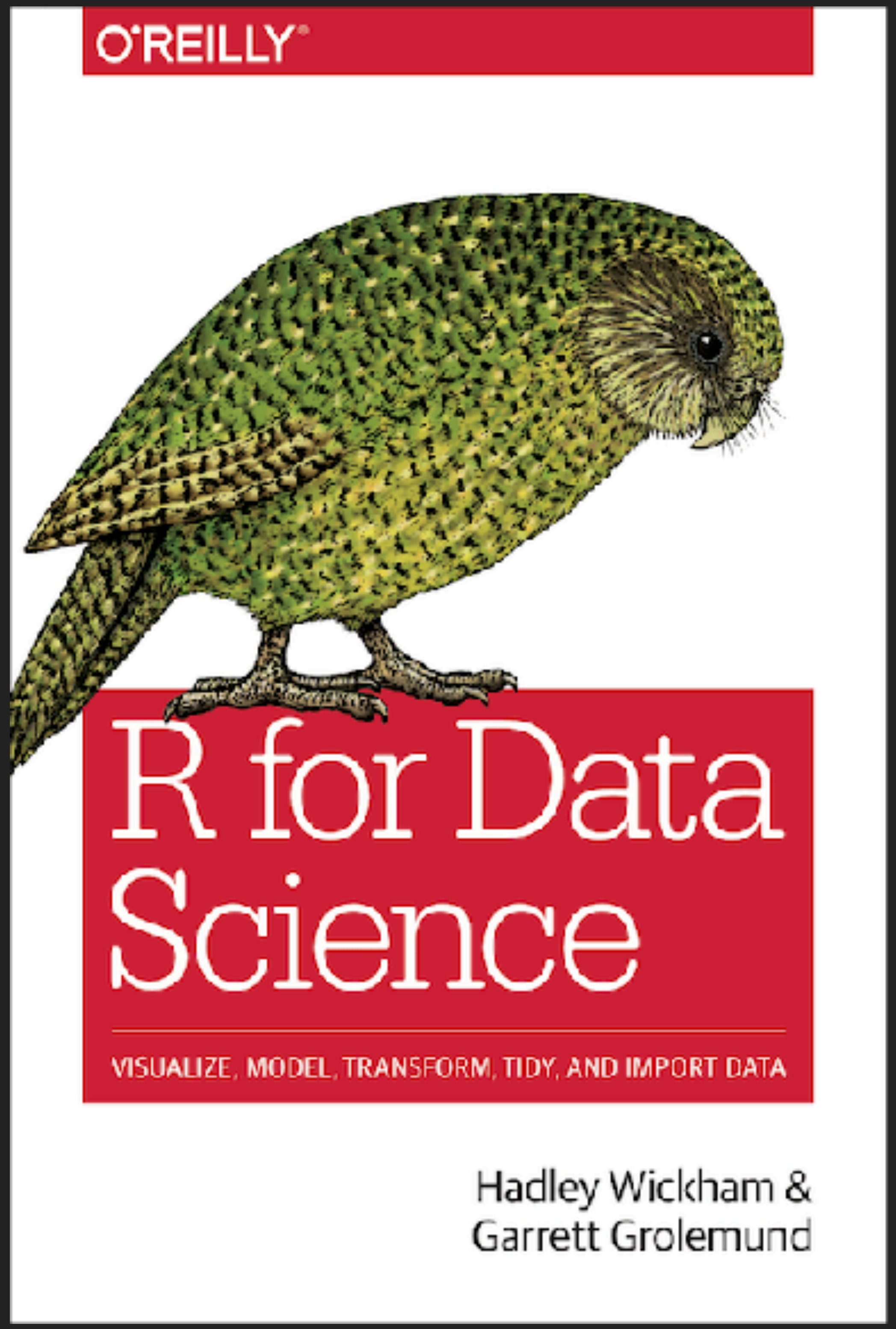
Tuesday, March 23, 2021

Outline for Today

- A few leftovers
- Linear regression



Ch 3 (Linear regression)



Ch 22-24 (Modeling)

- Remember:
  - https://bvasiles.github.io/empirical-methods/

Leftovers from last lecture: Type I and Type II Errors

Hypothesis Tests

- Aka “significance tests”
- Purpose:
  - Could random chance be responsible for an observed effect?
- **Null hypothesis** (H_0):
  - The hypothesis that chance is to blame.
  - e.g., “There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch.”
- **Alternative hypothesis** (H_a):
  - Counterpoint to the null (what you hope to prove).
  - e.g., “It takes less time on average to complete a task using NL2Code rather than by writing code from scratch.”

Type I and Type II Errors

		Study conclusion	
		No difference	Using NL2Code is faster
Reality	No difference	✓	Type I error
	Using NL2Code is faster	Type II error	✓

Type I and Type II Errors

- In assessing statistical significance, two types of error are possible:
  - Type I: you mistakenly conclude an effect is real, when it is really just due to chance
    - False positives
  - Type II: you mistakenly conclude that an effect is due to chance, when it actually is real
    - False negatives
- The basic function of hypothesis tests is to protect against being fooled by random chance; thus they are typically structured to minimize Type I errors.

Controlling the Risks of Type I and Type II Errors

- The probability of making a Type I error is called alpha.
  - (or “significance level”, “P-value”)
- The probability of making a Type II error is called beta.
- The statistical power of a test, defined as $1 - \beta$, refers to the probability of successfully rejecting a null hypothesis when it is false and should be rejected.
- To reduce errors:
  - Type I: $P < 0.05$
  - Type II: large sample size

Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the probability of Type I errors (false positives)?

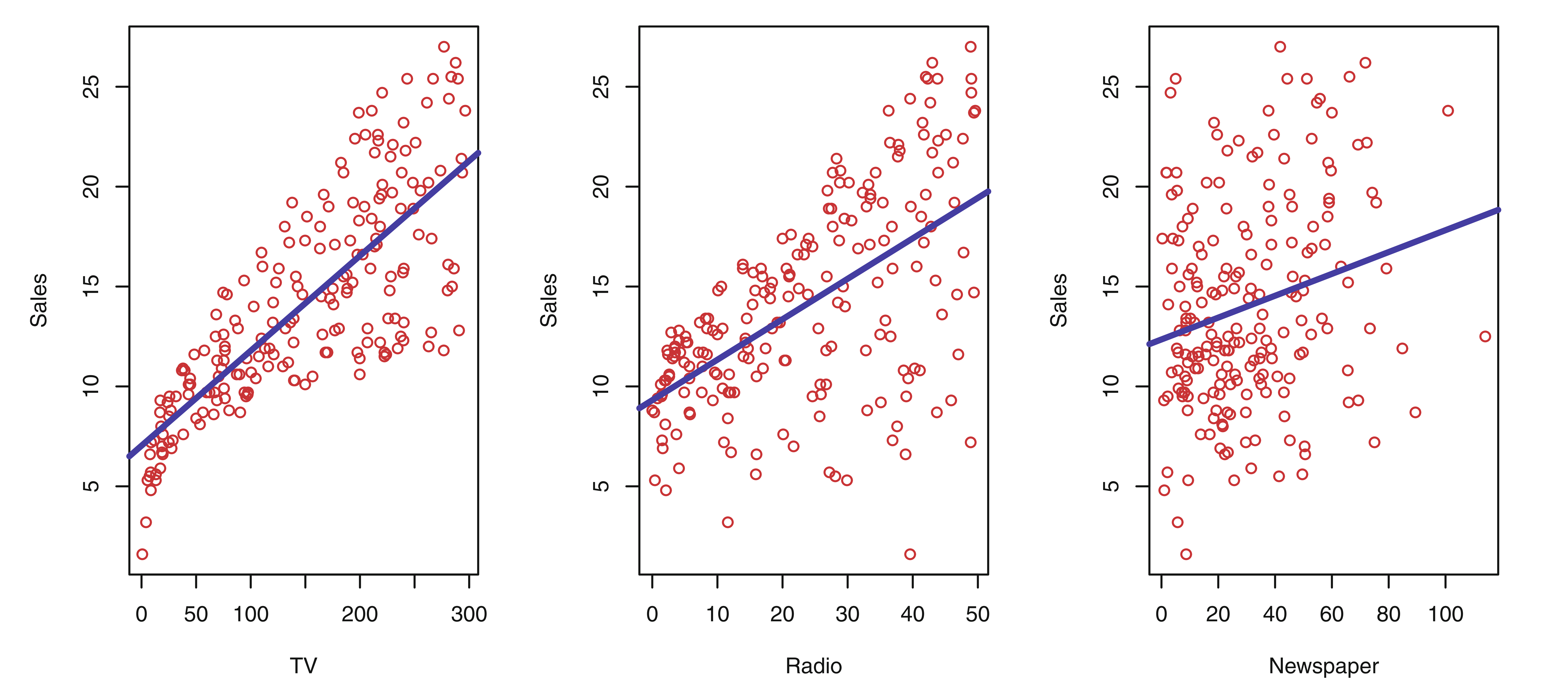
Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the probability of Type I errors (false positives)?

- The probability that one will correctly test nonsignificant is 0.95
- The probability that all 20 will correctly test nonsignificant is:
  - $0.95 \times 0.95 \times 0.95\ldots$, or $0.95^{20} = 0.36$
- The probability that at least one predictor will (falsely) test significant:
  - 1 – (probability that all will be nonsignificant) = 0.64 😱

Main topic for today:
Let's start with a case study.

Sales (in thousands of dollars) as a function of TV, radio, and newspaper advertising budgets (in thousands of dollars), for 200 cities.

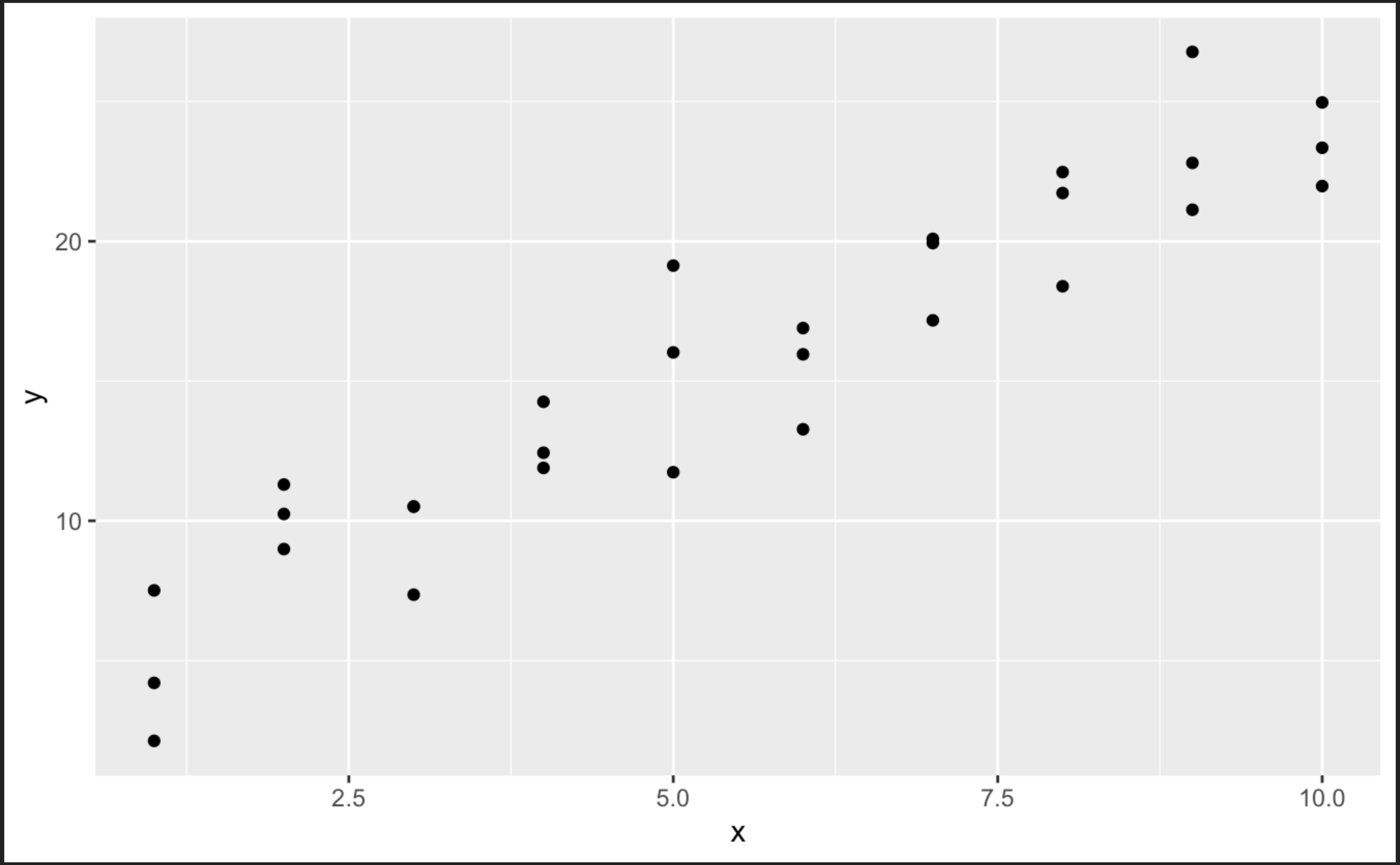


A Few Important Questions That We Might Seek To Address

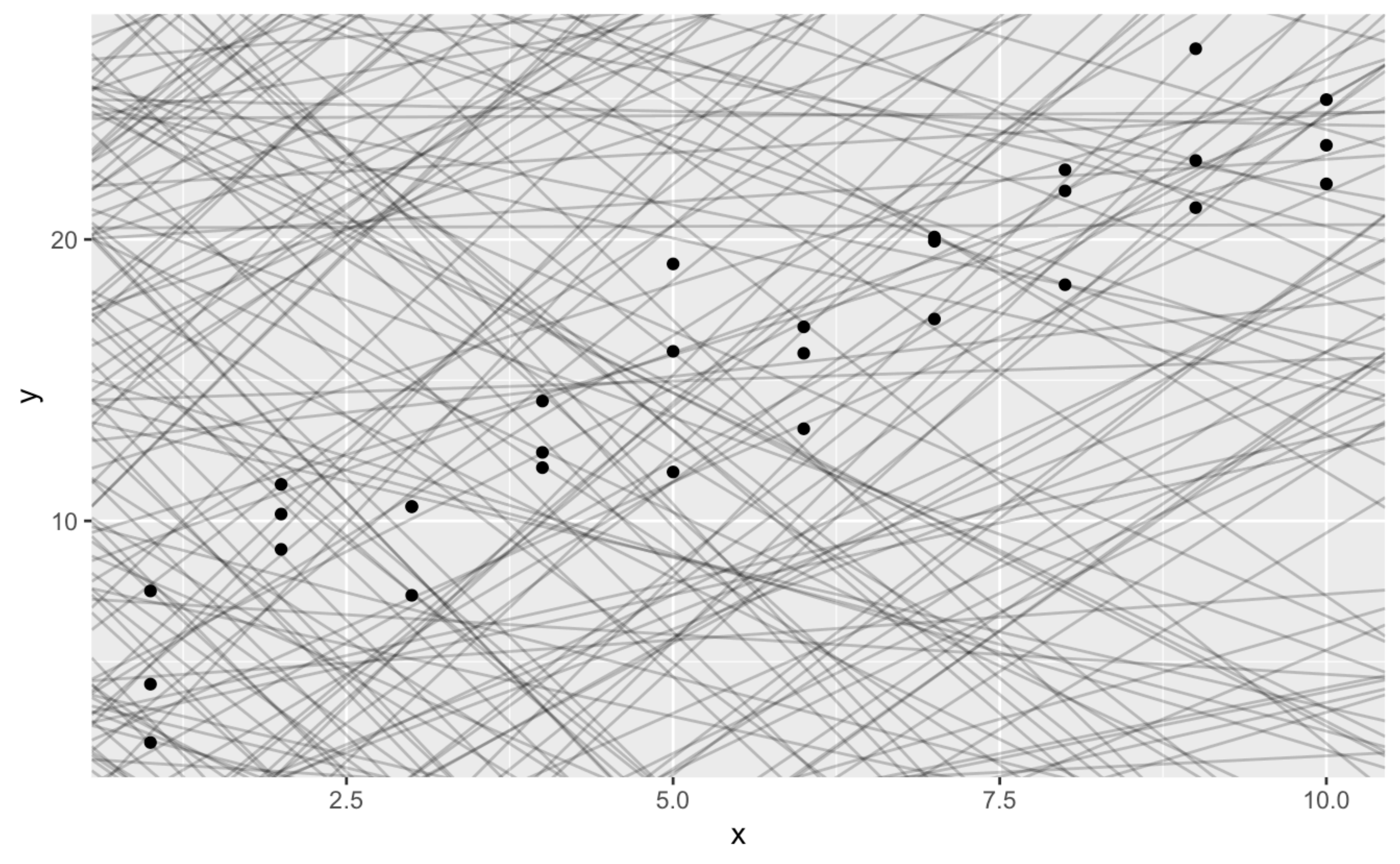
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Simple Linear Regression

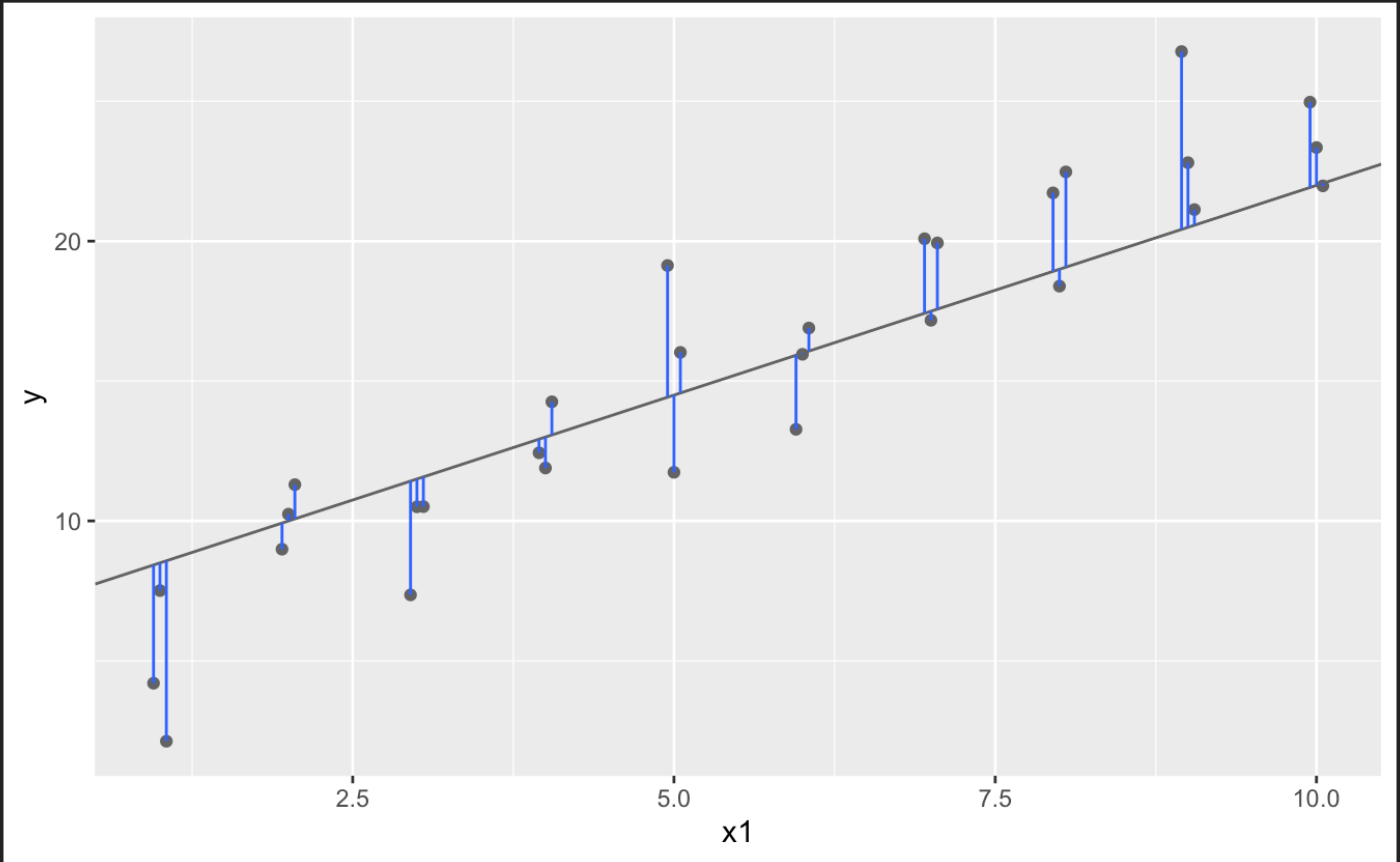
$$Y \approx \beta_0 + \beta_1 X.$$



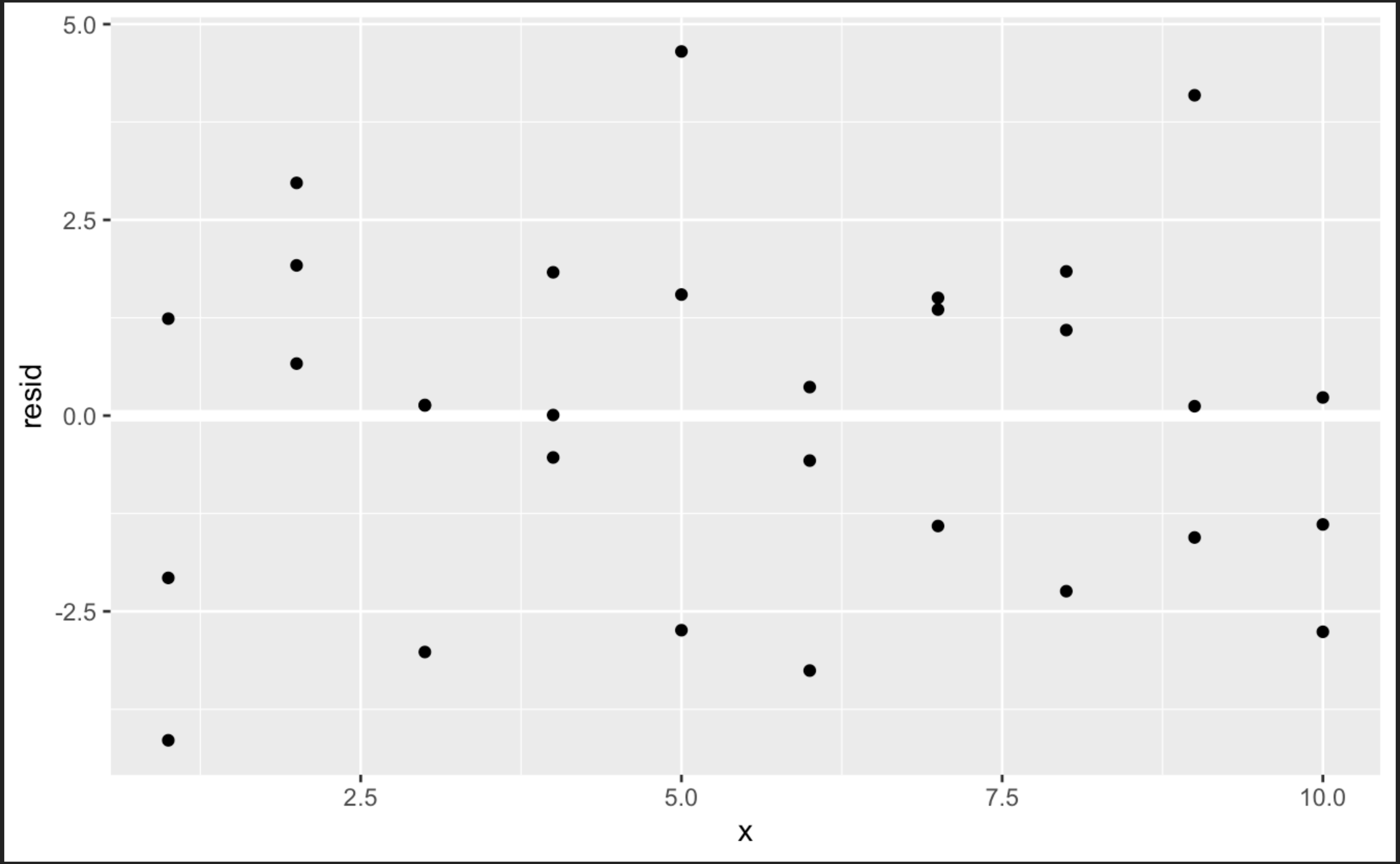
Many Possible Linear Models



Best Model? Minimize Error



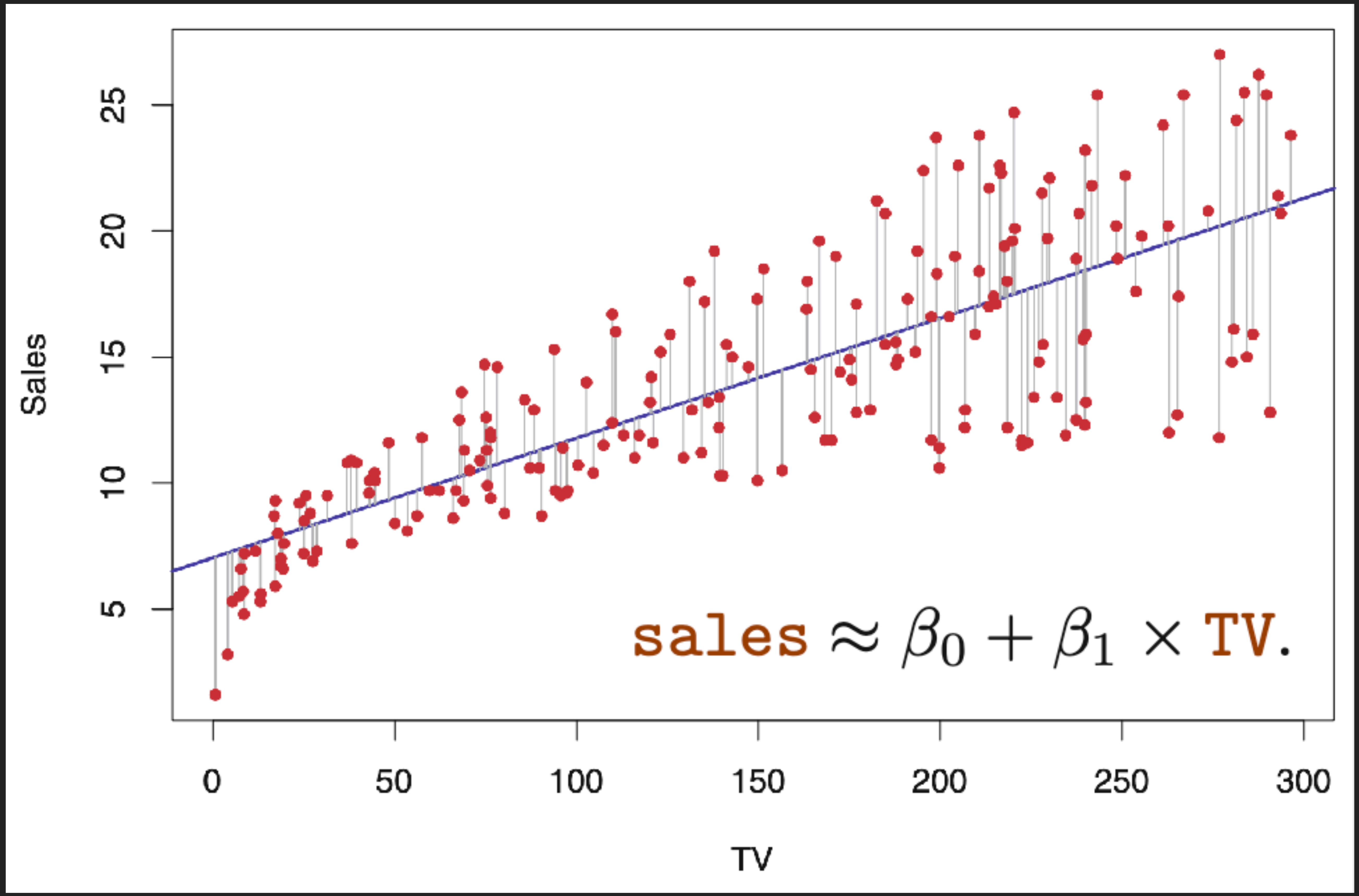
Residuals



Simple Linear Regression

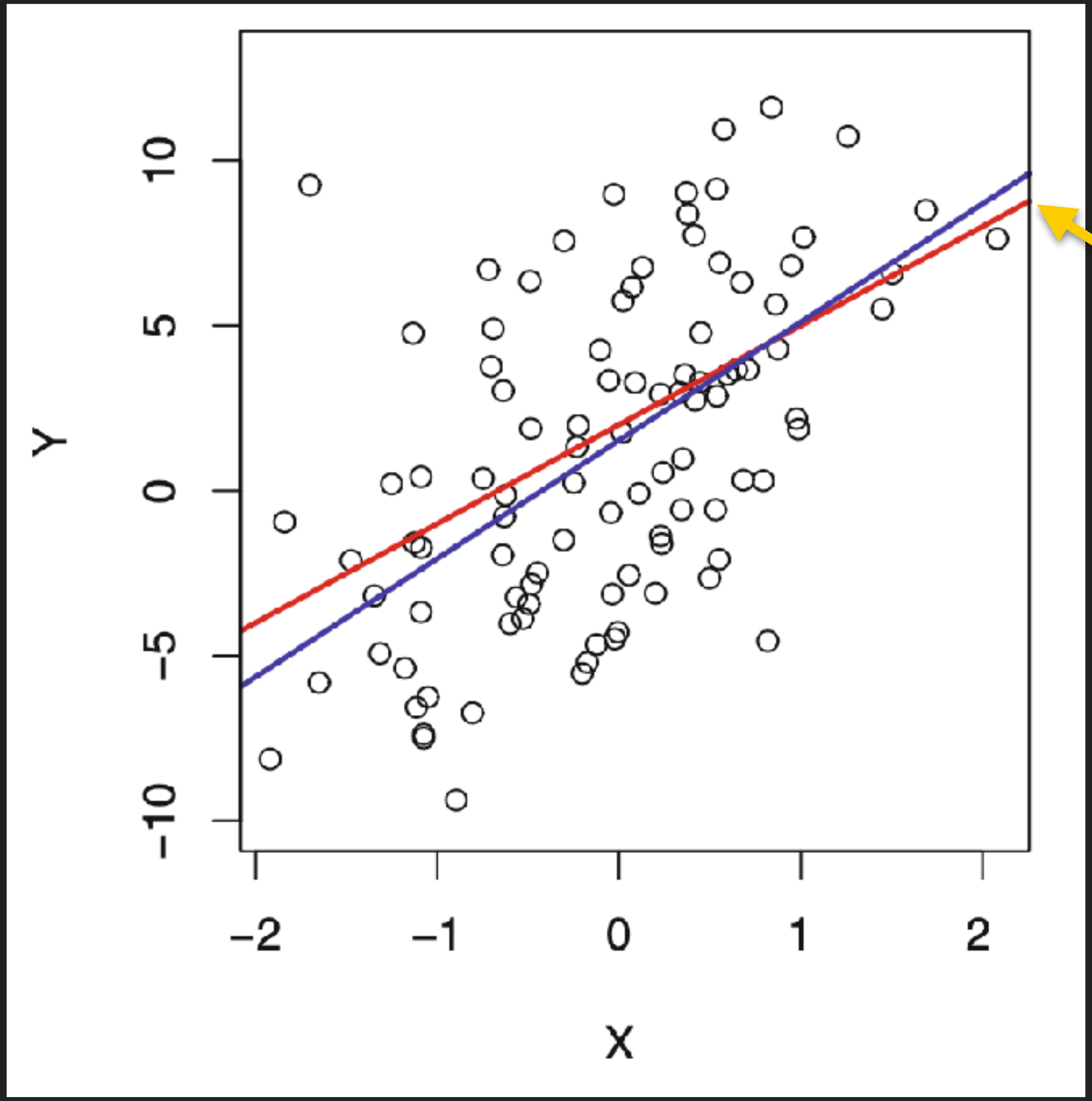
$$Y \approx \beta_0 + \beta_1 X.$$

The least squares fit for the regression of sales onto TV



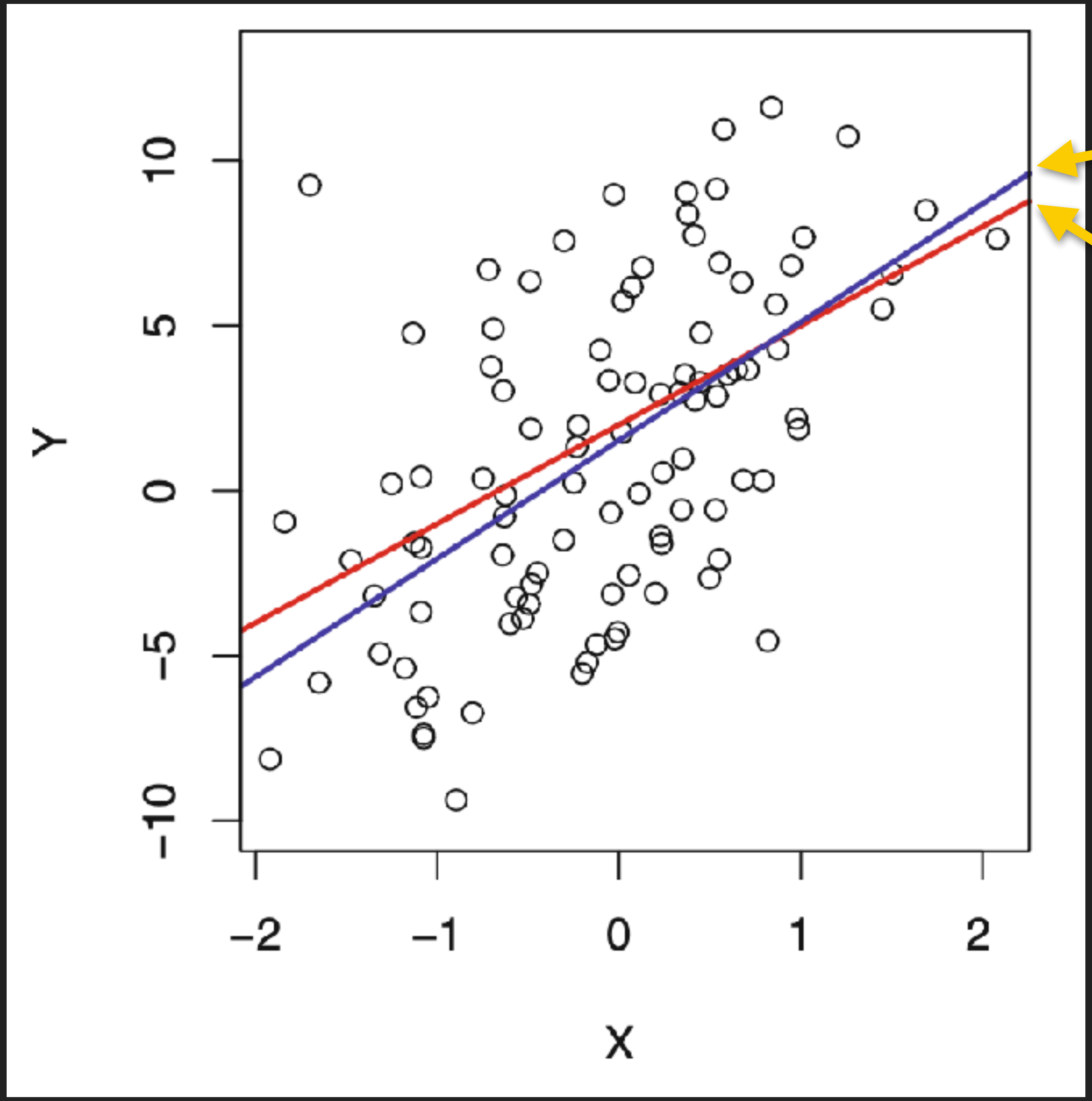
- The least squares fit for the regression of sales onto TV is found by minimizing the sum of squared errors.

Assessing the Accuracy of the Coefficient Estimates



The true relationship: $f(X) = 2 + 3X$

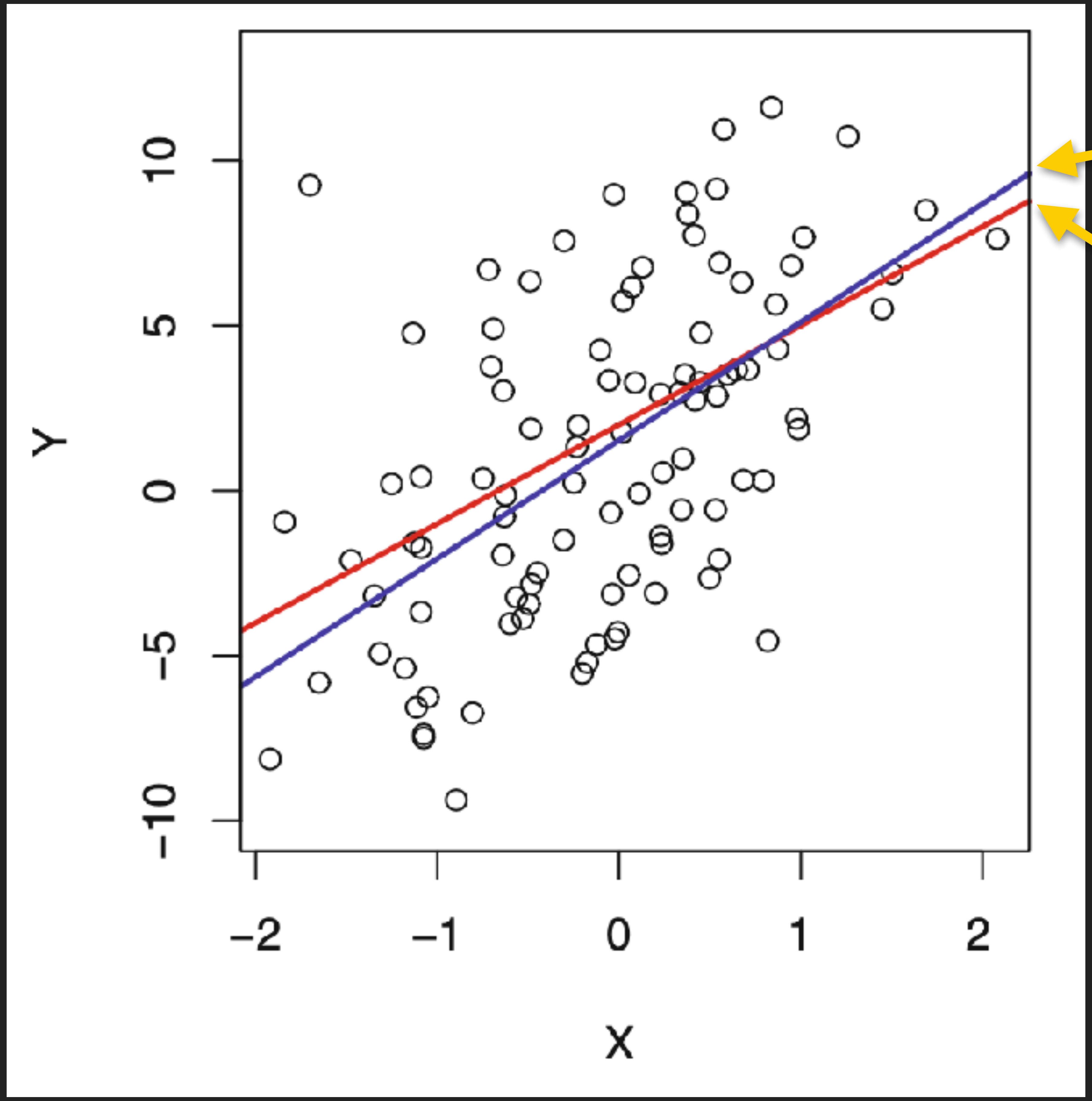
Assessing the Accuracy of the Coefficient Estimates



The least squares estimate for f(X) based on the observed data.

The true relationship: $f(X) = 2 + 3X$

Assessing the Accuracy of the Coefficient Estimates



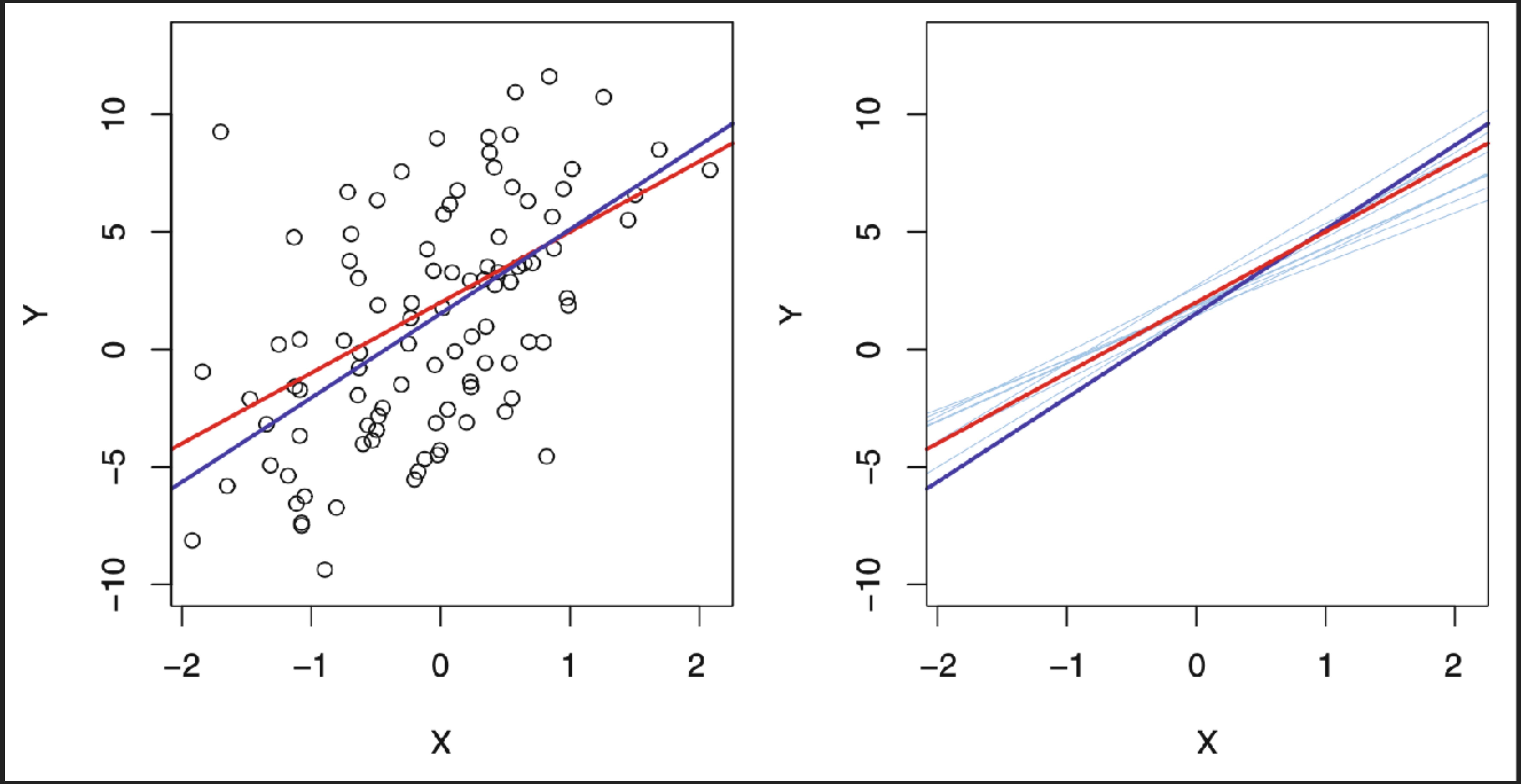
The least squares estimate for f(X) based on the observed data.

The true relationship: $f(X) = 2 + 3X$

In real applications, the population regression line is unobserved.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

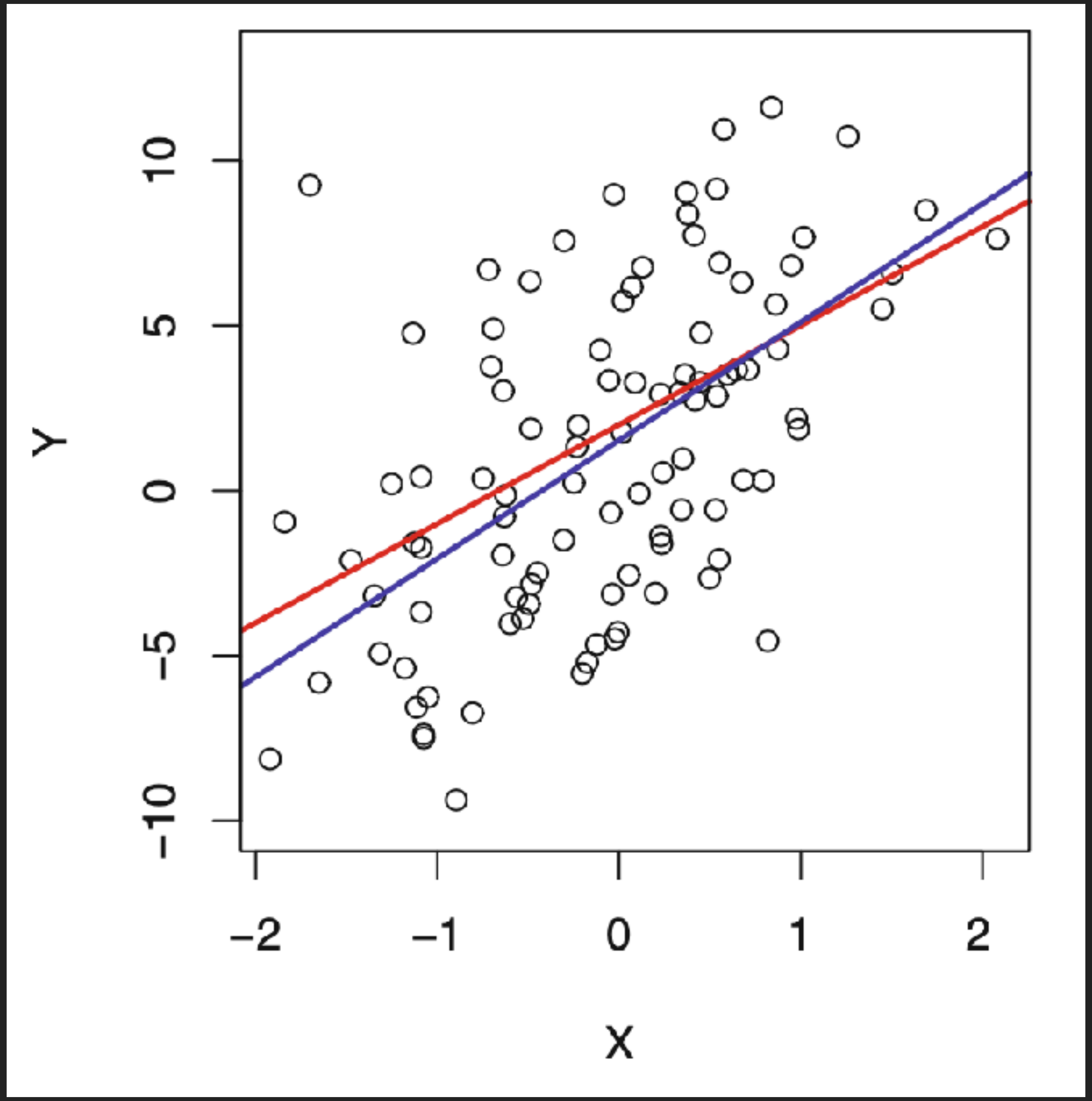
Assessing the Accuracy of the Coefficient Estimates



Ten least squares lines, each computed on the basis of a separate random set of observations.

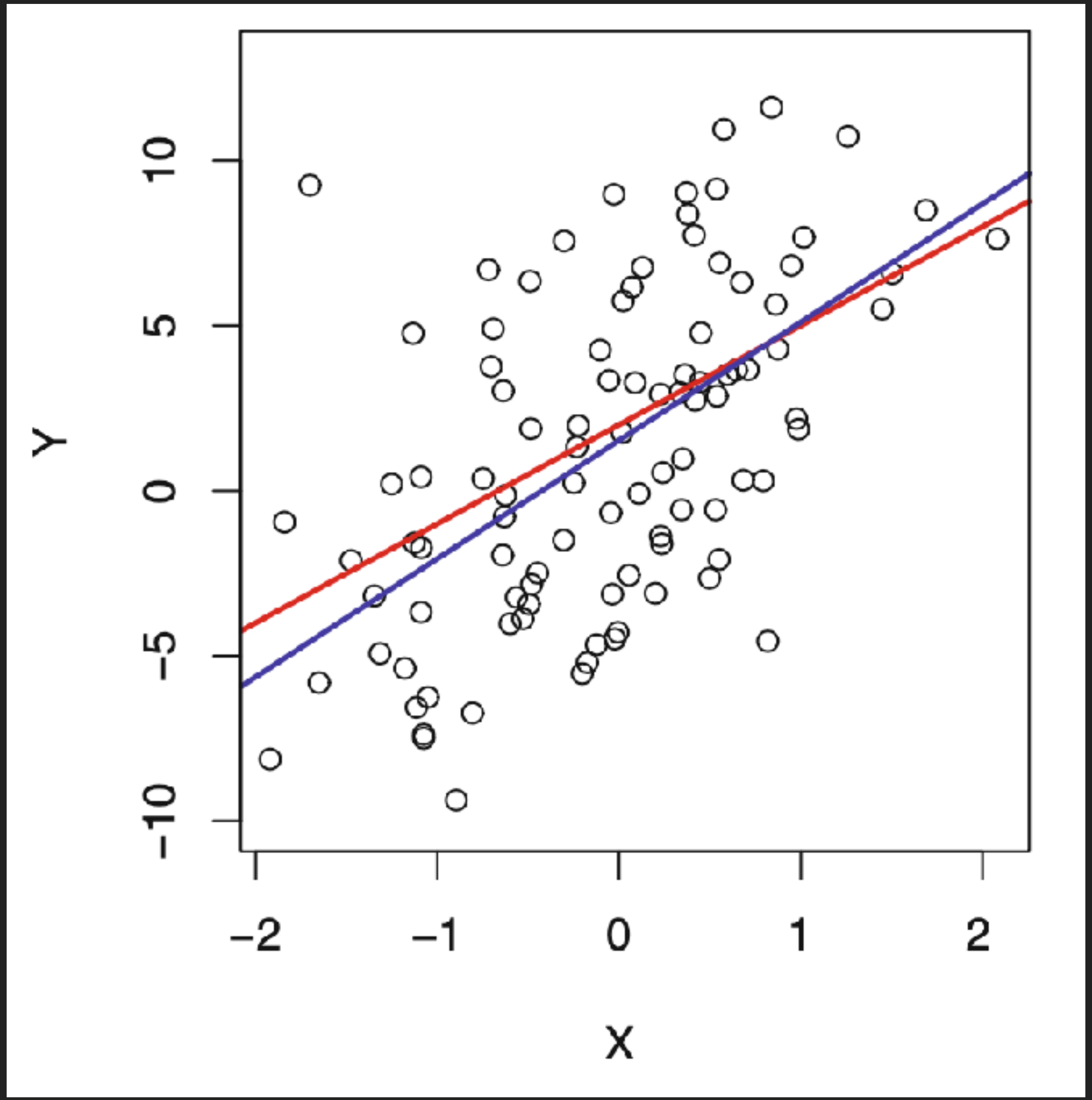
The average of many least squares lines is pretty close to the true population regression line.

Analogy with the estimation of the population mean μ of a random variable Y



- A natural question: how accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
  - Standard error
- Standard errors can be used to compute confidence intervals.
  - A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Analogy with the estimation of the population mean μ of a random variable Y



- For linear regression, the 95% confidence interval for β1 approximately takes the form

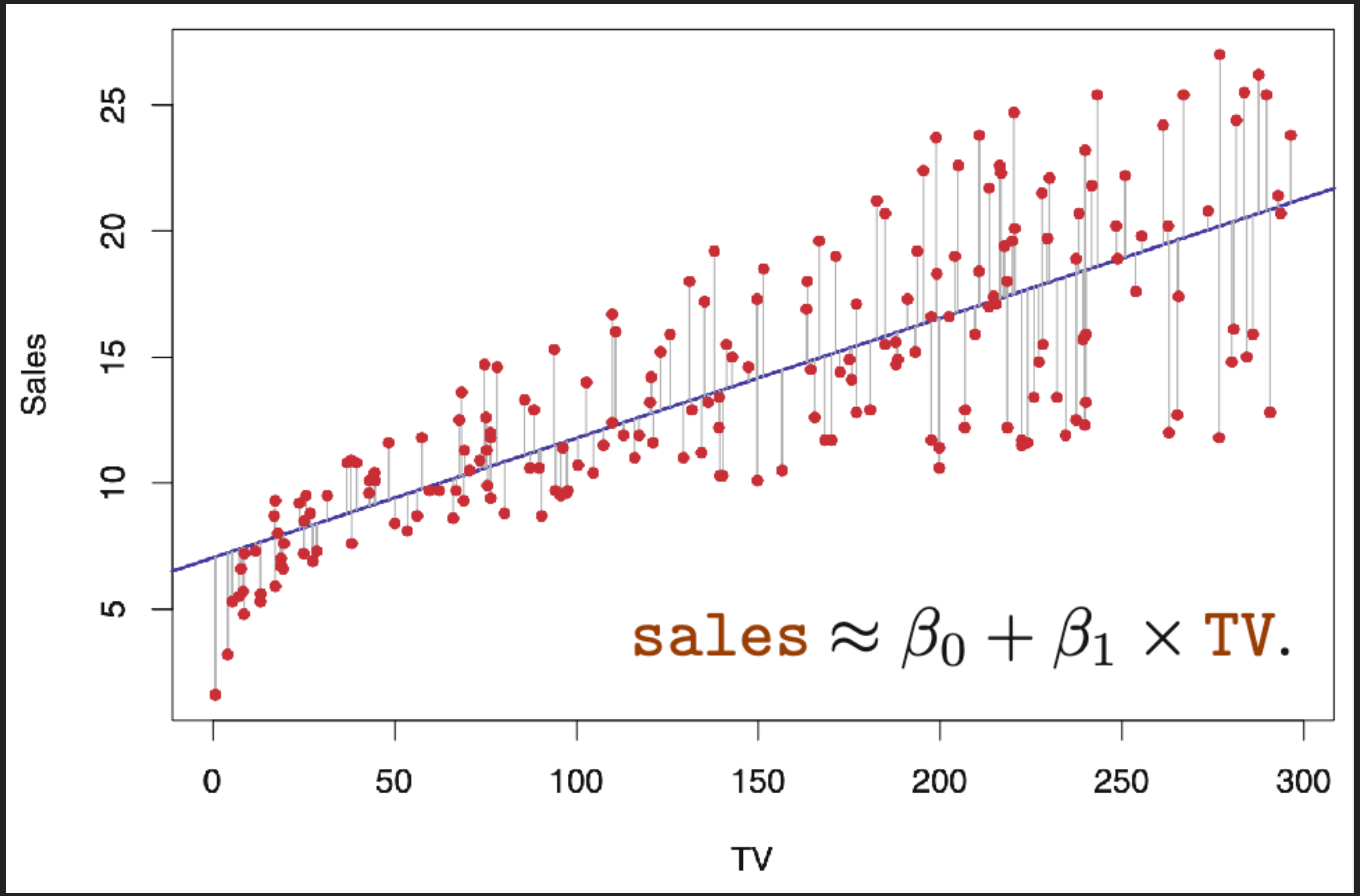
$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

- Similarly, a confidence interval for β0 approximately takes the form

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0).$$

Back to our example

The least squares fit for the regression of sales onto TV



- The 95 % CI for β0: [6.130, 7.935]
  The 95 % CI for β1: [0.042, 0.053]
- In the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,940 units.
- For each $1,000 increase in TV advertising, there will be an average increase in sales of between 42 and 53 units.

Key idea for empirical research

Standard Errors Can Also Be Used To Perform Hypothesis Tests on the Coefficients.

- Testing the null hypothesis:
  - H0 : There is no relationship between X and Y
- vs the alternative hypothesis
  - Ha : There is some relationship between X and Y

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Standard Errors Can Also Be Used To Perform Hypothesis Tests on the Coefficients.

- Testing the null hypothesis:
  - H0 : There is no relationship between X and Y
- vs the alternative hypothesis
  - Ha : There is some relationship between X and Y

- Corresponds to testing

$$H_0 : \beta_1 = 0$$

- vs

$$H_a : \beta_1 \neq 0,$$

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

=> Compute a t-statistic and associated p-value

Standard Errors Can Also Be Used To Perform Hypothesis Tests on the Coefficients.

- Testing the null hypothesis:
  - H0 : There is no relationship between X and Y
- vs the alternative hypothesis
  - Ha : There is some relationship between X and Y

- Corresponds to testing

$$H_0 : \beta_1 = 0$$

- vs

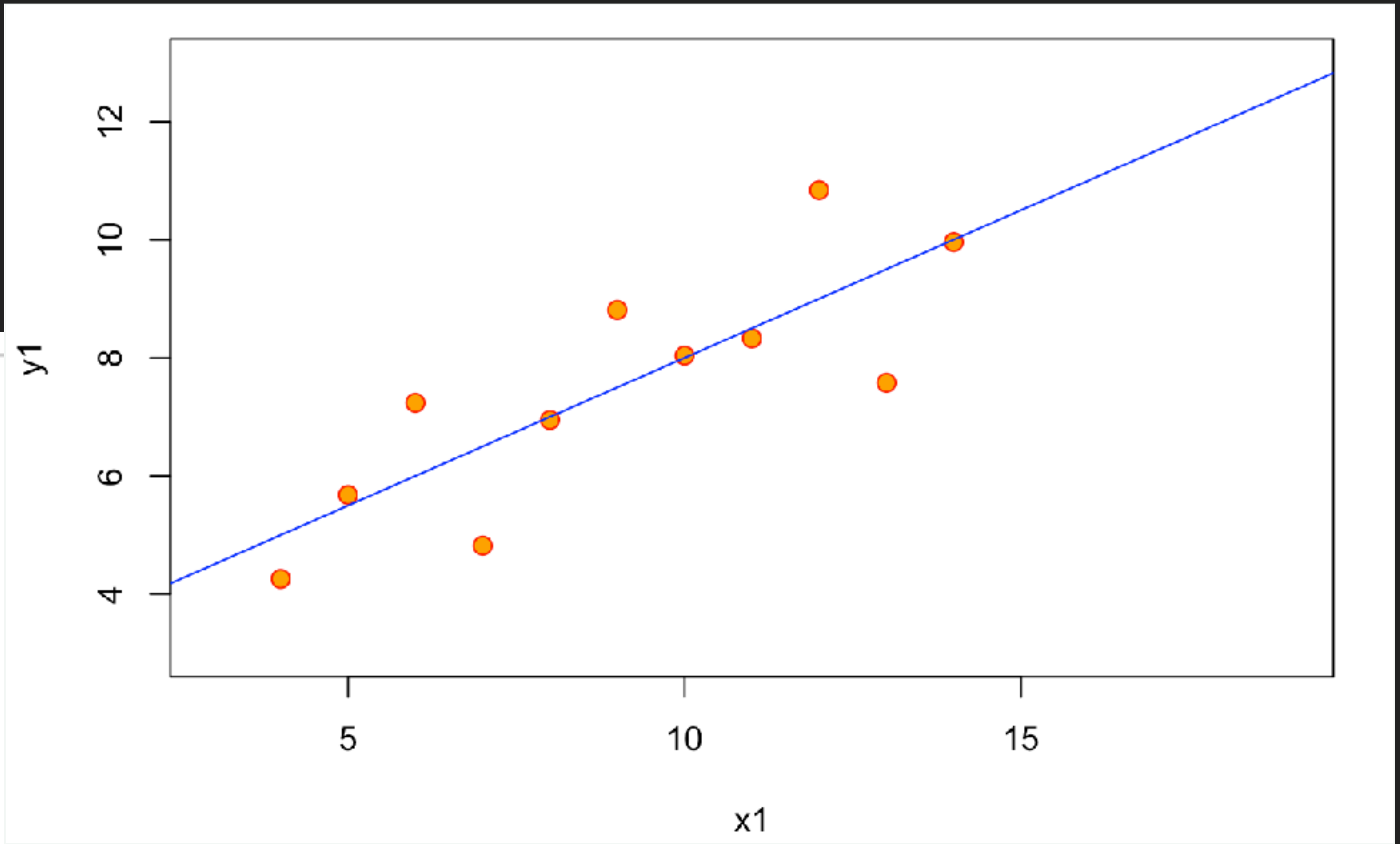
$$H_a : \beta_1 \neq 0,$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units.

Another Example

```
##
## Call:
## lm(formula = y1 ~ x1, data = anscombe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## x1            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```



Let´s make it more realistic

How To Extend our Analysis To Accommodate all Predictors?

- One option is to run three separate simple linear regressions.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

How To Extend our Analysis To Accommodate all Predictors?

- A better option is to give each predictor a separate slope coefficient in a single model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

- We interpret βj as the average effect on Y of a one unit increase in Xj, *holding all other predictors fixed*.

Aside: Ingredients for Establishing a Causal Relationship

The cause preceded the effect

The cause was related to the effect

We can find no plausible alternative explanation for the effect other than the cause

Back to our Advertising Example

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

. . . to be continued

Credits

- Graphics: Dave DiCello photography (cover)
- Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In Seminars in Hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Grolemund, G., & Wickham, H. (2018). R for data science.