Fall 2023

BIF524/CSC463 Data Mining

Linear Regression

Eileen Marie Hanna, *PhD*

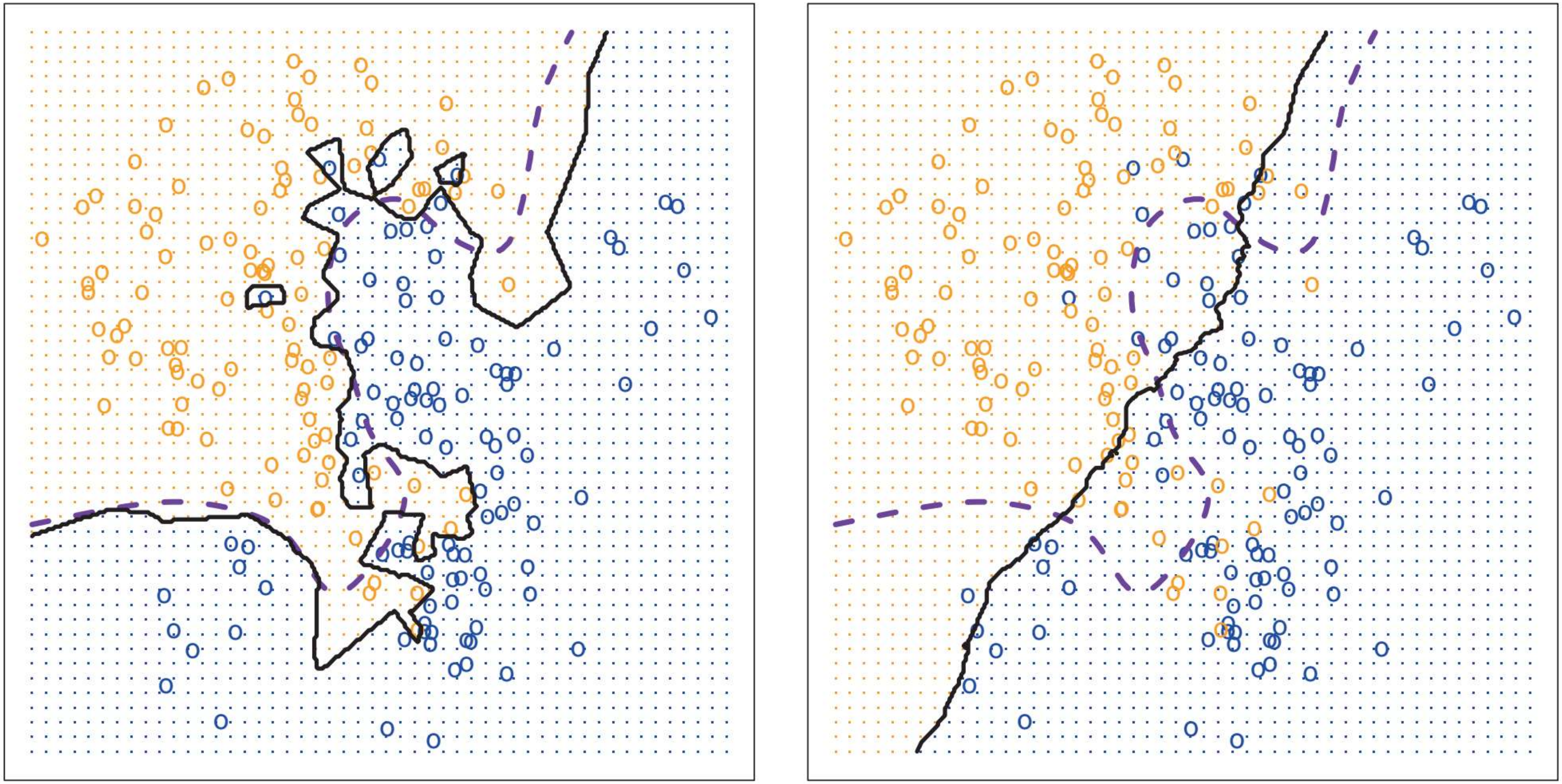
19/09/2023

cran dplyr | ggplot2

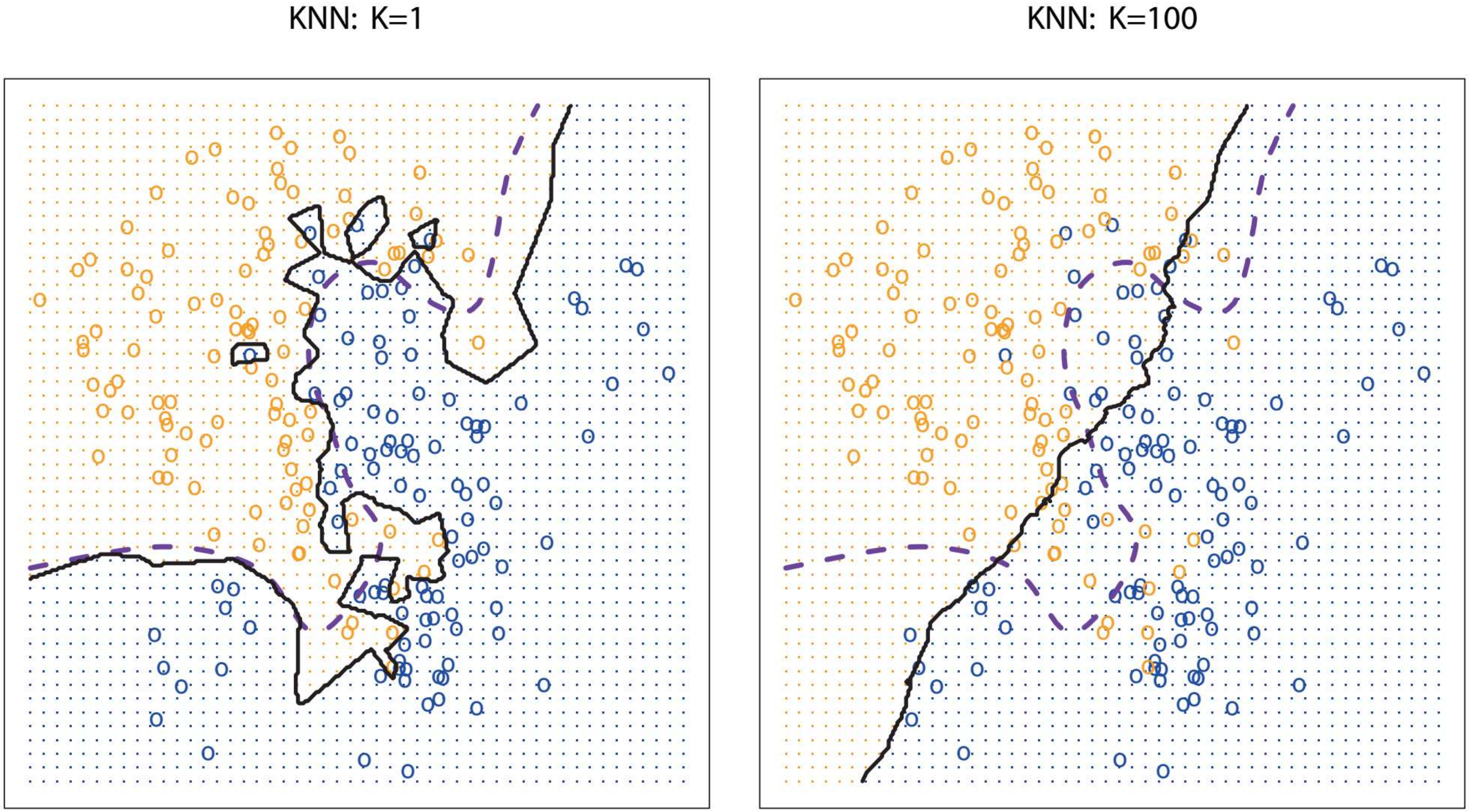
↳ Reference manual

Vignettes

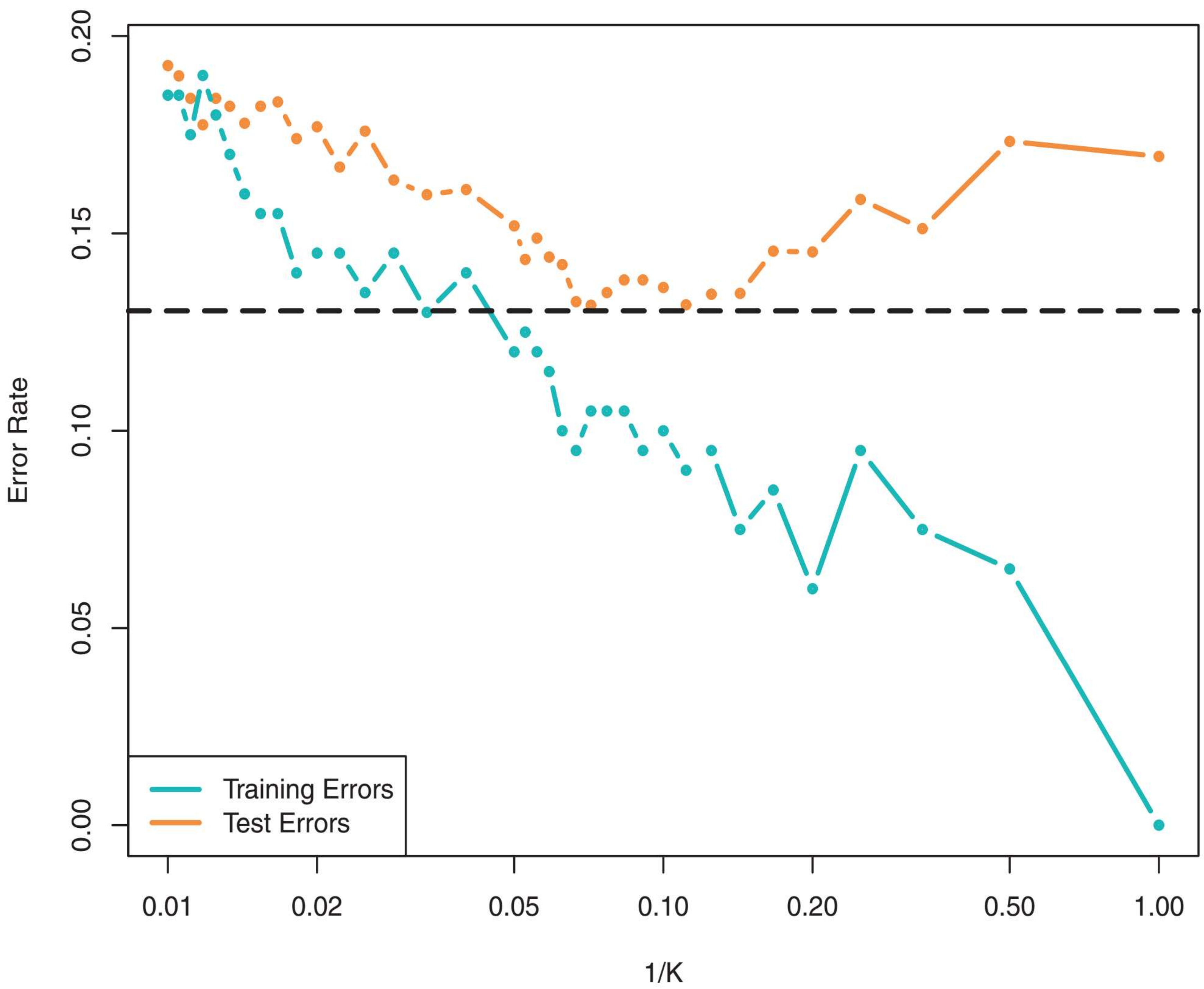
Identify KNN with $K = 1$ and $K = 100$



K-nearest neighbors classifier



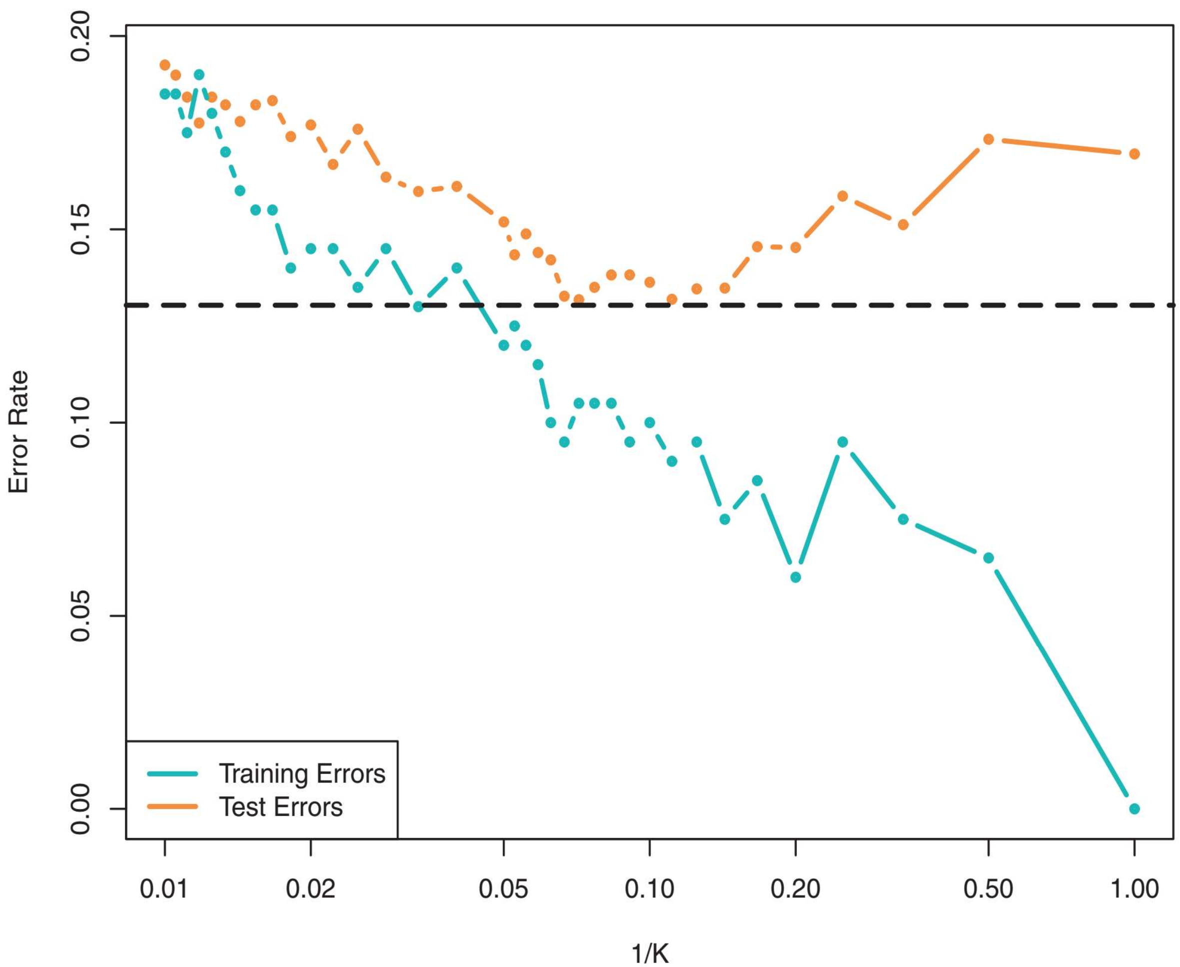
- The **choice of K** highly influences the classification.
- The **decision boundary is very flexible** when $K = 1$ -> low bias and high variance.
- **As K increases**, the flexibility decreases, and the **decision boundary becomes closer to linear** -> low variance and high bias.
- Both cases don't lead to good predictions.



Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

K-nearest neighbors classifier

- No strong relationship between training and test error rates.
- $K = 1$ -> training error rate 0 but test error rate potentially high.
- **flexible classification methods -> lower training error but no guarantee on test error.**



- training and test errors as a function of $1/K$.
- as $1/K$ increases -> higher flexibility
- **the training error rate declines when flexibility decreases**.
- **U-shape of the test error** which declines at the first (to a min. approx. $K = 10$) and then increases when the method becomes more flexible.

Linear regression

- **A very simple supervised learning approach**, in comparison to modern approaches.

- Can be used **to predict quantitative outputs.**

- It is somehow the basis for other novel approaches.

Linear regression – “Advertising” dataset

- You need to come up with a **marketing plan to increase the product sales.**
  - What kind of information can be useful in this process?
    - **Is there a relationship between advertising budget and sales?**
      - If no evidence of relationship -> don’t spend money on advertising!
      - If yes, how strong is the relationship between advertising budget and sales?
        - Given a certain advertising budget, can we predict sales with a high level of accuracy?

Linear regression – "Advertising" dataset

- **Which media contributes to sales?**
  - Find a way to examine (**separate**) **individual effects** of each medium on sales when money is spent on all three.

- **How accurately can we estimate the effect of each medium on sales?**
  - accurately predict the amount of sales increase.

- **How accurately can we predict future sales?**
- **Is the relationship between advertising expenditure linear?**
- **Is there synergy (interaction effect) among the advertising media?**

Simple linear regression

- Predicting a **quantitative response** Y based on a **single predictor variable** X.

- **Assumes a linear relationship** between X and Y:

intercept

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

slope

e.g., $$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- After using the training data to compute those two coefficients, we can predict future sales based on TV advertising budget.

Estimating the coefficients

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

- Given a set of observations with measured outputs, we need to **obtain coefficient estimates $\widehat{\boldsymbol{\beta}}_0$ $\boldsymbol{and}$ $\widehat{\boldsymbol{\beta}}_1$ such as the resulting line is as close as possible to the data points** (closeness).

- We want the linear model to fit the data well, such that:

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \textbf{for } \boldsymbol{i = 1, \dots, n}$$

Estimating the coefficients – residual sum of squares

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- The **least squares approach seeks to** find $\hat{\beta}_0$ *and* $\hat{\beta}_1$ that **minimize the RSS**.

- The corresponding **least squares coefficient estimates** for simple linear regression are:

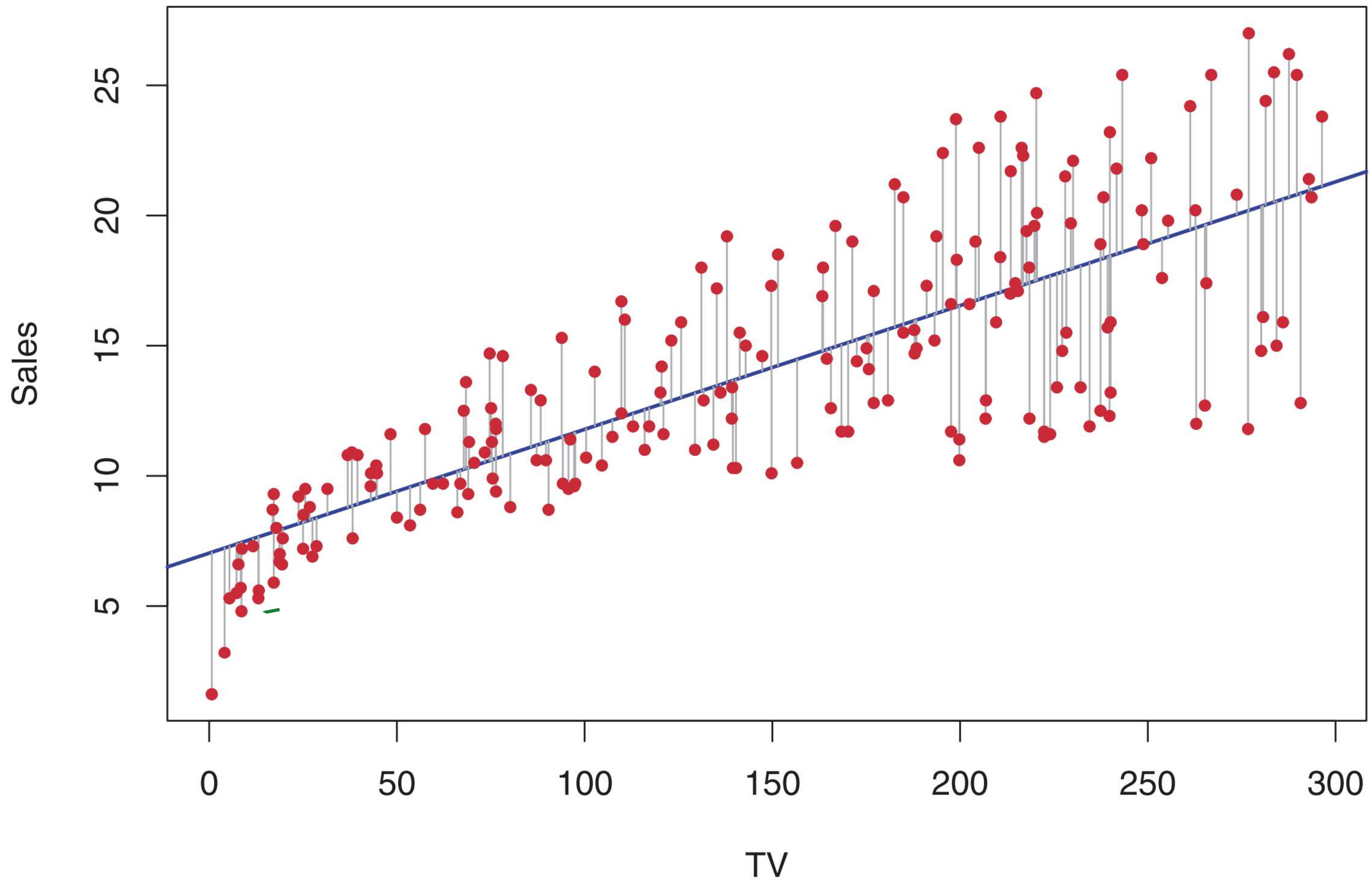
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

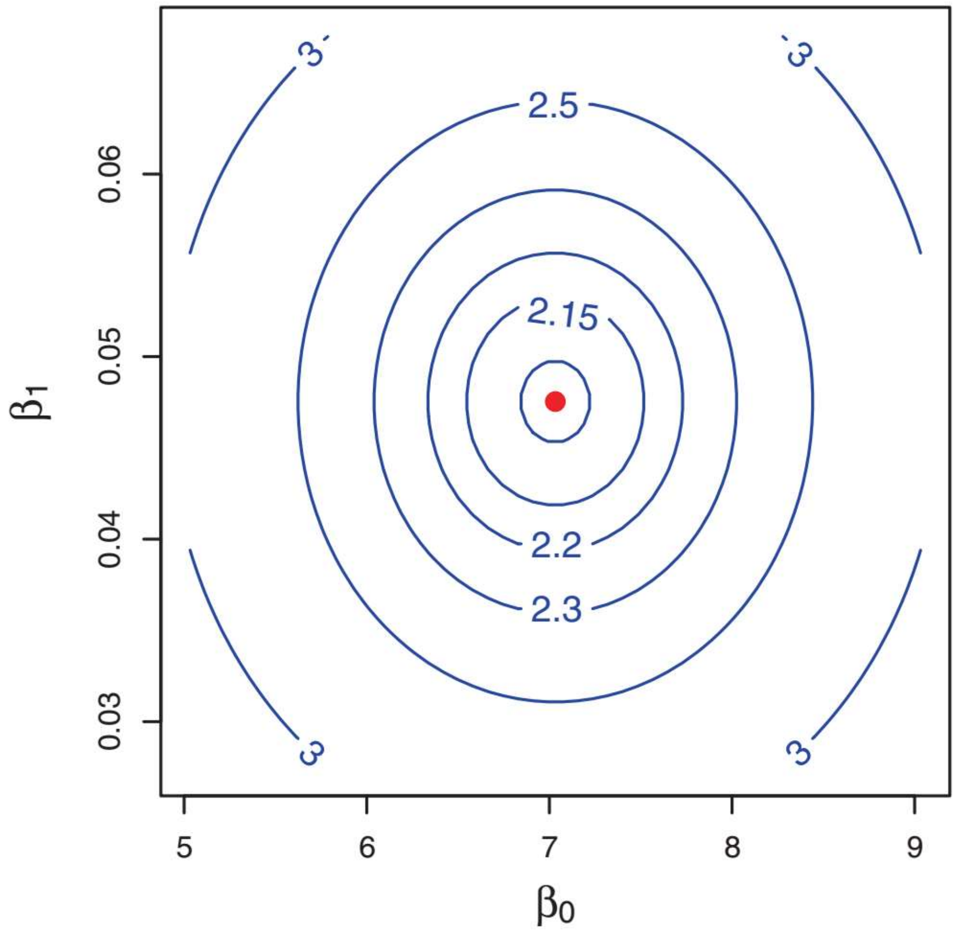
$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$$

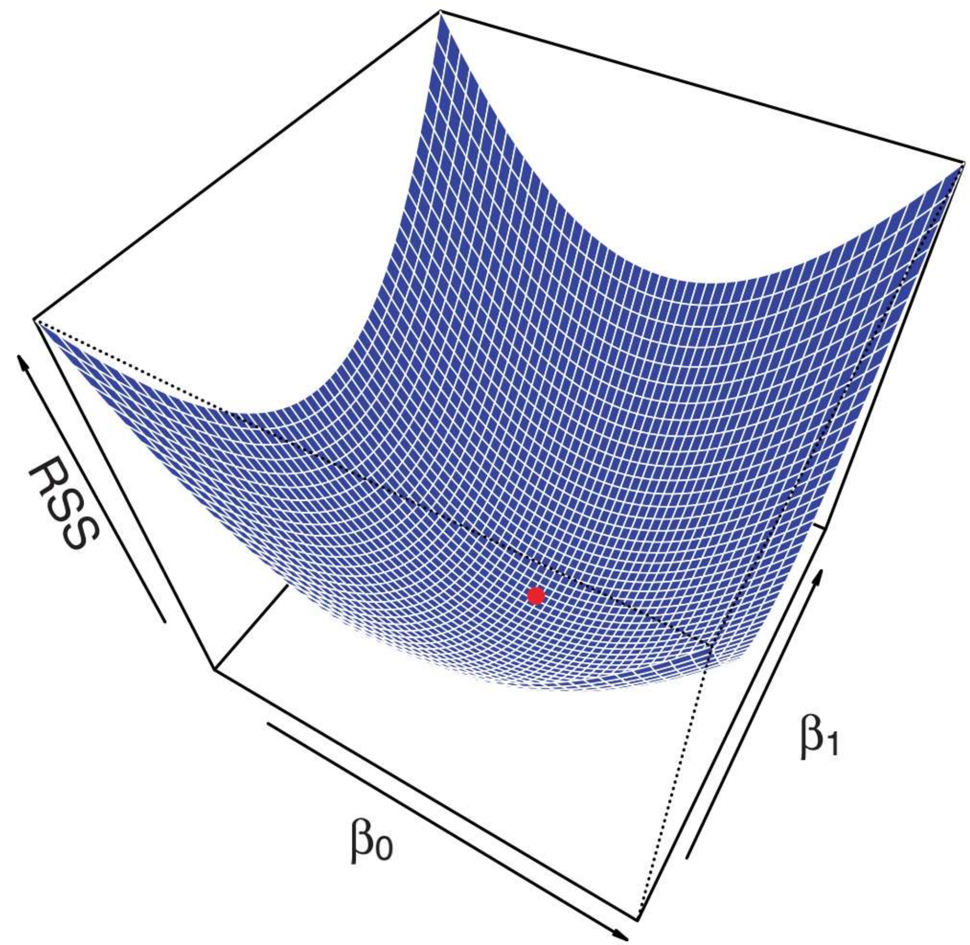
Least squares fit for the regression of sales onto TV



- It is found by **minimizing the sum of squared errors.**
- Grey lines represent errors, and the fit makes a compromise by averaging their squares.
- For most of the observations, except the first part, the line fits the data relatively well.

Least squares fit for the regression of sales onto TV





Contour and $3D$ plots of the RSS on the advertising data using TV as predictor and sales as response. Red dots correspond to the least squares estimates.

Assessing the accuracy of the coefficient estimates

- Initially, we assumed that the true relationship between X and Y is:

$$Y = f(X) + \epsilon$$

for an unknown function f, and ϵ is a mean-zero random error.

- If we approximate f with a linear function, the we can rewrite this relationship as:

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

intercept: expected value of Y when $X = 0$

slope: average increase in Y corresponding to a unit increase in X

Assessing the accuracy of the coefficient estimates

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

stands for all what we miss with this simple model.

e.g., the true relationship may not be linear, other variables may also cause variations in Y, possible measurement errors, ..etc.

ϵ is typically assumed to be independent of X.

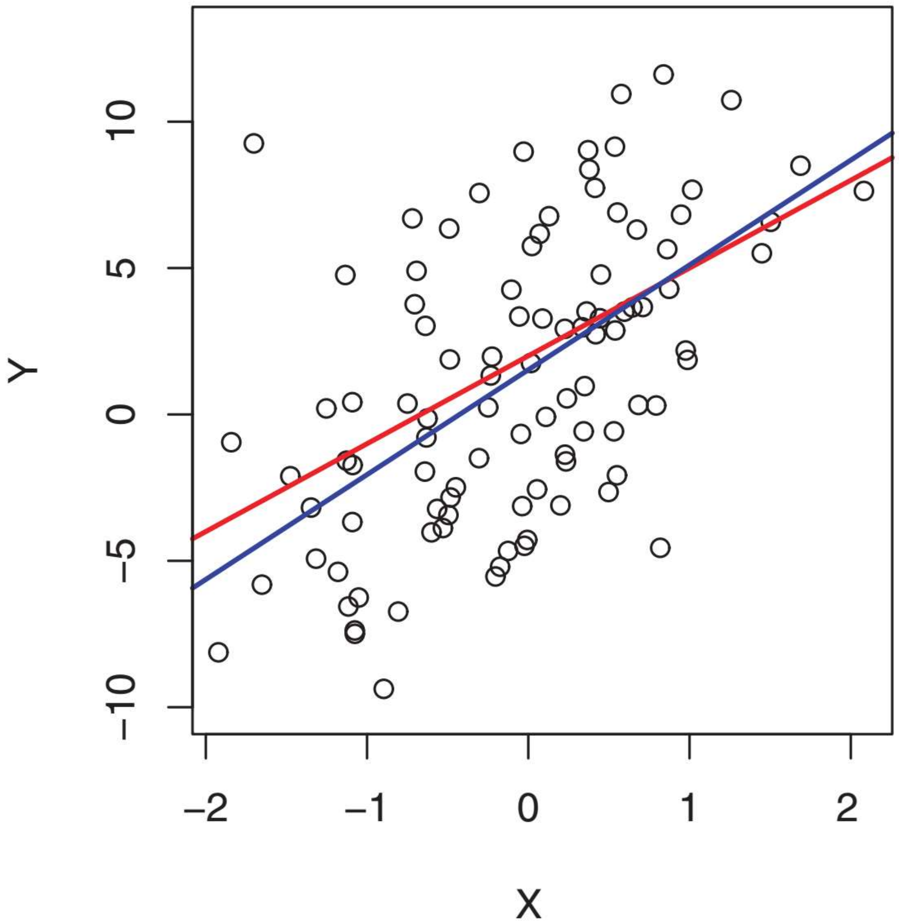
Assessing the accuracy of the coefficient estimates

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

This model defines the population regression line, i.e., the best linear approx. of the true relationship between input and output.

The resulting line is characterized by the estimated coefficients.

Assessing the accuracy of the coefficient estimates



the population regression line (true relationship), here known as $f(X) = 2 + 3X$

the least squares line (estimate of $f(X)$) based on the observed data, calculated using coefficient estimates presented in the previous equations.

- The two lines were created using 100 random Xs for which the corresponding Ys were computed, from the model:

$$Y = 2 + 3X + \epsilon$$

- The error was generated from a normal distribution with mean 0.

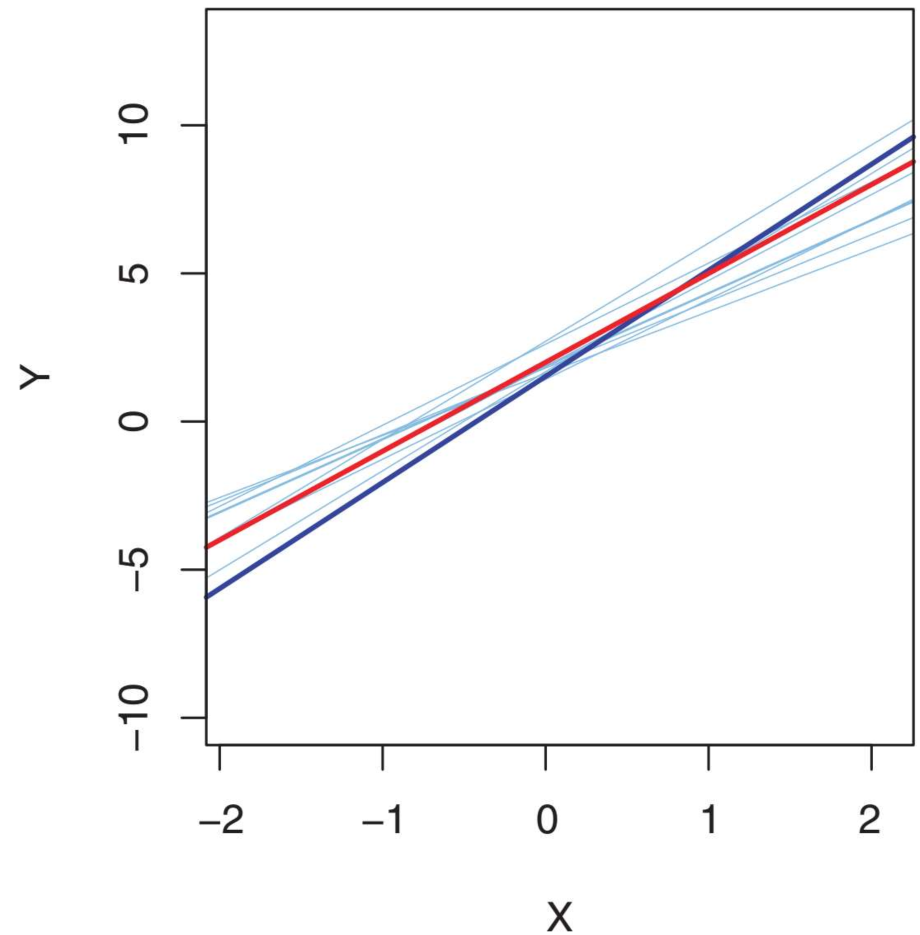
Assessing the accuracy of the coefficient estimates

- What does it mean to have two different lines (regression and least squares) to describe the relationship between predictor and response?
  - The idea is similar to using a sample to estimate the characteristics of a large population – **analogy with standard statistics**.
  - Suppose that we are interested in estimating the population mean μ of a random variable Y, but we do not have access to n observations from Y.

Assessing the accuracy of the coefficient estimates

- The sample mean and the population mean are different, but in general, the **sample mean will provide a good estimate of the population mean**.

- In analogy with this, the unknown coefficients β_0 and β_1 in linear regression define the population regression line.



Assessing the accuracy of the coefficient estimates

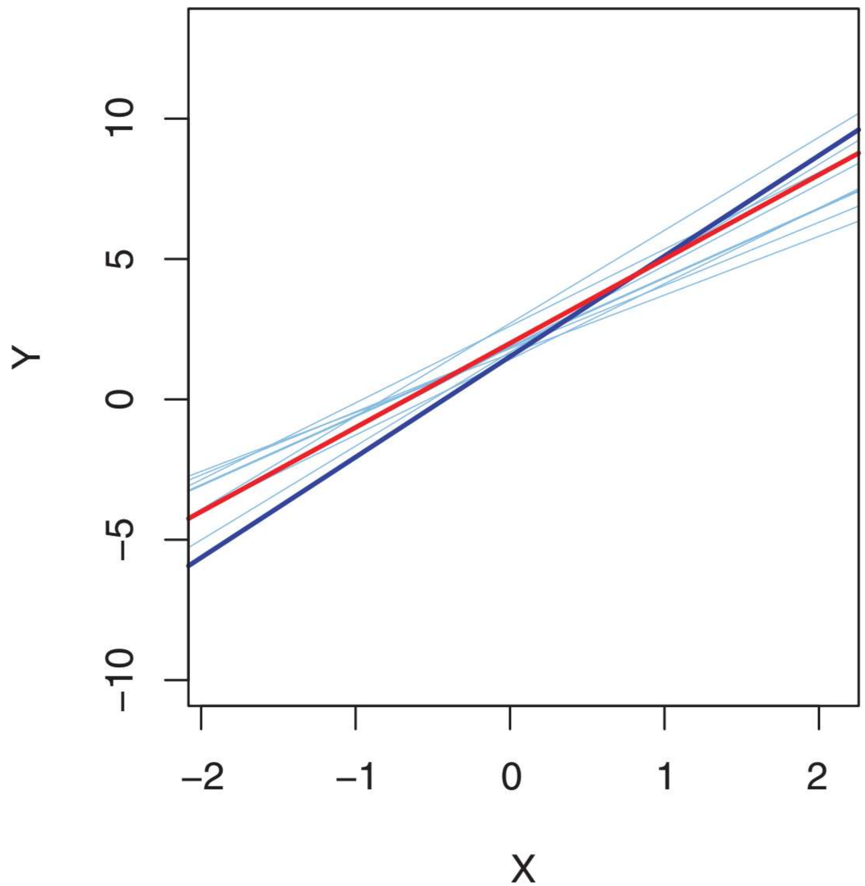
  - The sample mean and the population mean are different, but in general, the **sample mean will provide a good estimate of the population mean**.

- In analogy with this, the unknown coefficients β_0 and β_1 in linear regression define the population regression line.

  - The estimated coefficients (using the formulae) define the **least squares line**.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Assessing the accuracy of the coefficient estimates



population true regression line

10 least squares lines from separate sets of random observations

least squares line (estimate of $f(X)$ based on the observed data)

$$Y = 2 + 3X + \epsilon$$

- We usually have a set of observations for which we can calculate the least square lines.
- The light blue plots correspond to 10 least squares lines for the model generated from 10 different datasets.
- Different datasets generated from the same true model -> slightly different least squares lines.
- Least square lines are different but **on average they are quite close to the regression line**.

Assessing the accuracy of the coefficient estimates

- Based on one set of observations $y_1, y_2, \ldots, y_n$, $\hat{\mu}$ might overestimate μ.

  - And, on the basis on another set, it might underestimate it.

- If we could **average a huge number of estimates**
  - $\hat{\mu}$ would equal μ

- An unbiased estimator does not systematically under- or over-estimate the true parameter.

- The same applies to the least-squares coefficient estimates...

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

How far off will a single estimate of μ be?

- The answer is given by computing the standard error of $\hat{\mu}$:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

σ being the standard deviation of each y_i of Y

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

- In other terms, the standard error tells us the **average amount that $\hat{\mu}$ differs from the actual value μ**.

- Also, the deviation when the number of observations increases.

How far off will a single estimate of μ be?

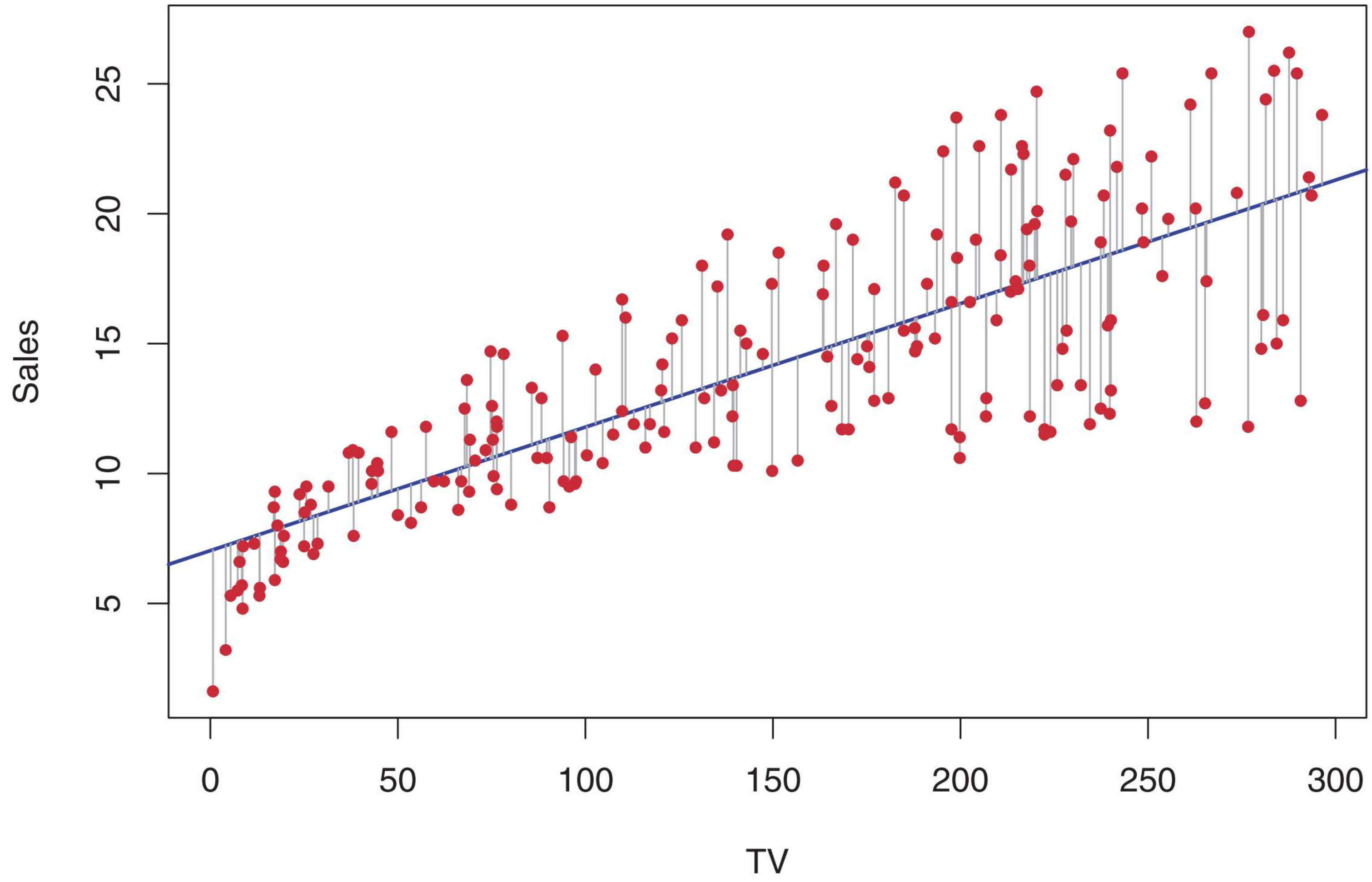
- In analogy, we can see how close are $\hat{\beta}_0$ and $\hat{\beta}_1$ to the true coefficient β_0 and β_1, using:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

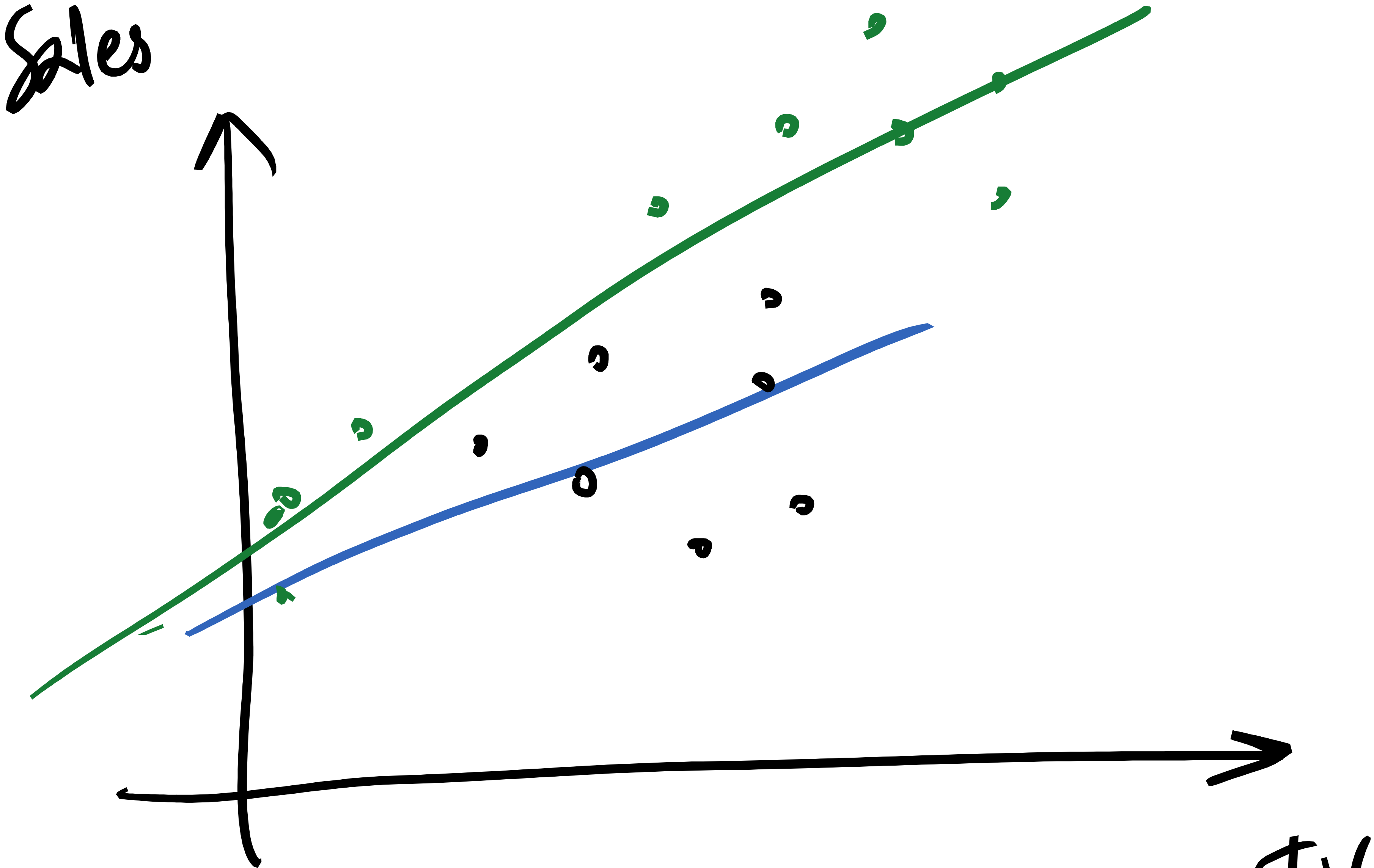
$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Note that $\boldsymbol{SE}(\widehat{\boldsymbol{\beta}}_1)$ **tends to be smaller when $\boldsymbol{x_i}$ are more spread out**.

How far off will a single estimate of μ be?



- Imagine that we have all observations concentrated within a certain range of the x-axis.
- The restriction on the slope would be less -> more variance because many lines could fit the points -> **the more the points are spread out -> lower SE is expected**.
- Collecting data that is spread out across the axis -> eventually lead to **more precise line**.


Sales
TV

Confidence intervals

- **Standard errors can be used to compute confidence intervals.**

- A 95% confidence interval is the range of values such that with **95% probability, it will include the true unknown value of a parameter**.

- It is defined by the lower and upper limits computed from the sample of data.

- In linear regression, the **95% confidence interval for $\boldsymbol{\beta_1}$ is approx**.:

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1) \qquad \hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

Confidence intervals – example

Suppose that for the advertising data, the 95% confidence intervals for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are $[6.130, 7.935]$ and $[0.042, 0.053]$.

- **What would the sales be on average, in the absence of any advertising?**

Confidence intervals – example

Suppose that for the advertising data, the 95% confidence intervals for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are $[\mathbf{6.130, 7.935}]$ and $[\mathbf{0.042, 0.053}]$.

- **What would the sales be on average, in the absence of any advertising?**
  - somewhere between 6130 and 7935 units

- **What would be the average increase in sales for each $\$\mathbf{1000}$?**

$$\text{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 \, \text{TV}$$

Confidence intervals – example

Suppose that for the advertising data, the 95% confidence intervals for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are $[\mathbf{6.130, 7.935}]$ and $[\mathbf{0.042, 0.053}]$.

- **What would be the average increase in sales for each $\$\mathbf{1000}$?**
  - an average increase between 42 and 53 units

$\boldsymbol{\beta_1}$ with confidence interval $[\mathbf{0.042, 0.053}]$ -> TV advertising has a positive effect on sales.

The confidence interval also reflects how large is the effect of TV advertising on sales.

Coefficients estimates

- Standard errors can also be used to perform **hypothesis testing** on coefficients.

- The most common hypothesis test involves testing the null vs alternative hypotheses:

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$
$$H_a : \text{There is some relationship between } X \text{ and } Y$$

- Mathematically:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Coefficients estimates

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

- To test the null hypothesis, we **need to determine whether $\boldsymbol{\hat{\beta}_1}$** (our estimate) **is sufficiently far from zero** so that we can be confident that it is non-zero.

- How far is enough?
  - depends on how accurate is our estimate $\hat{\beta}_1$.
    - If $SE(\hat{\beta}_1)$ is small, even relatively small values of $\hat{\beta}_1$ can provide strong evidence that it is non-zero.
    - If it is large, then $\hat{\beta}_1$ must be very large in absolute value so we can reject the null hypothesis.

Coefficients estimates

- In practice, we compute the t-statistic, given by:

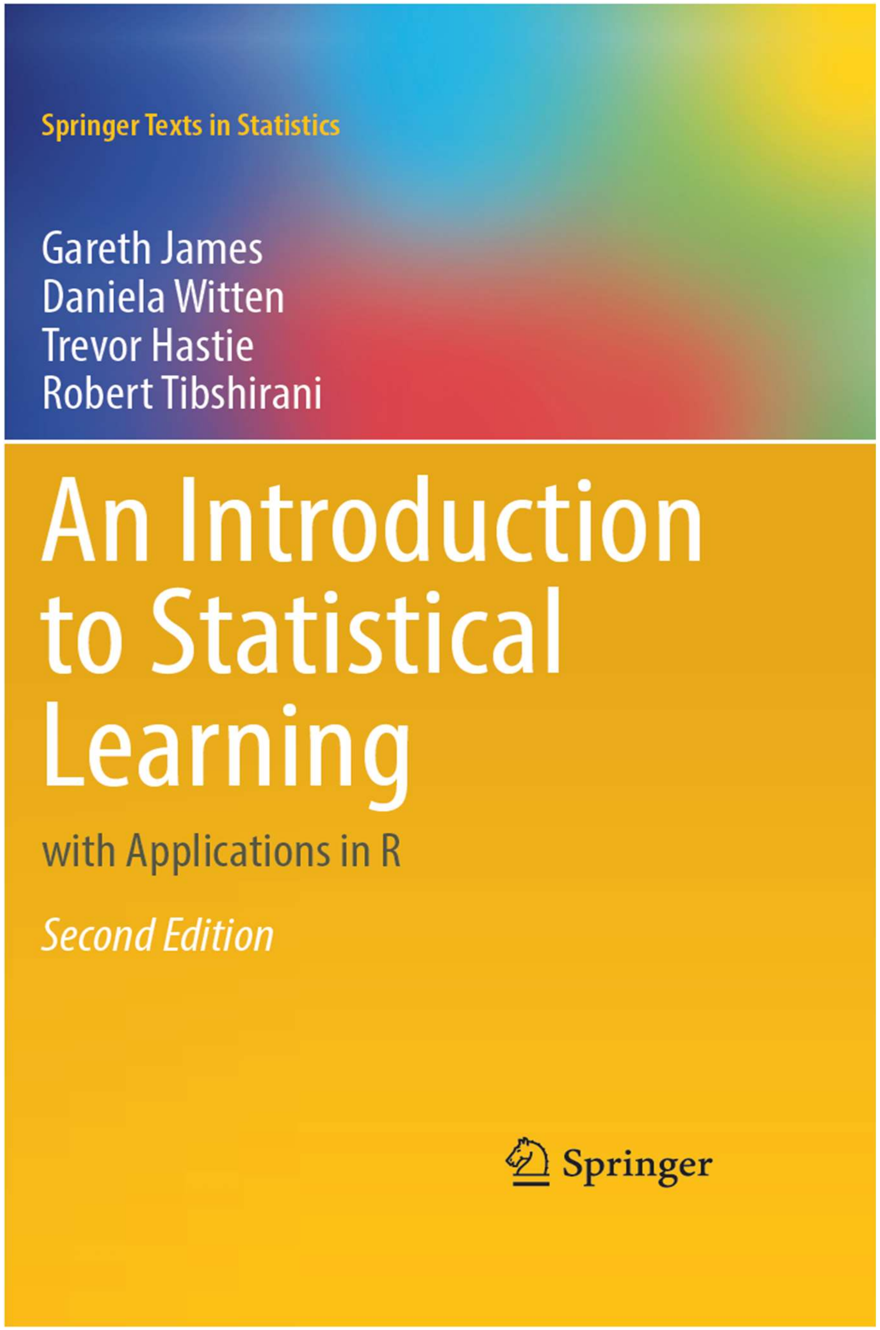
$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- We can see that it measures how far $\hat{\beta}_1$ deviates from zero.

  - If no relationship between X and Y -> t-distribution with $n-2$ degrees of freedom.
  - The distribution has a bell shape and for values approx. $\geq$ 30 -> similar shape to the normal distribution.

H_0 : There is no relationship between X and Y

Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer