Predictive Modeling

Mirko Birbaumer

HSLU T&A

1. Introduction Predictive Modeling

2. Introduction to Regression Analysis
   - Introduction
   - Simple Linear Regression Model
   - Estimating the Coefficients
   - Hypothesis Test and Confidence Interval for Regression Coefficients

3. Confidence and Prediction Interval for the Response

Predictive Modeling

We have tried to predict the future since ancient times when shamans looked for patterns in smoking entrails, or when priests asked the oracle of Delphi for advice.

Predictive Modeling

Recently, **predictive models** have been introduced

i. in *automated guided vehicles* to predict the motion of surrounding vehicles

ii. in *medicine* for detecting and predicting skin cancer

iii. in *hospitals* to prioritize patients for medical interventions based on their predicted risk of complications

iv. in *personalized medicine* to predict the reaction of patients to drugs

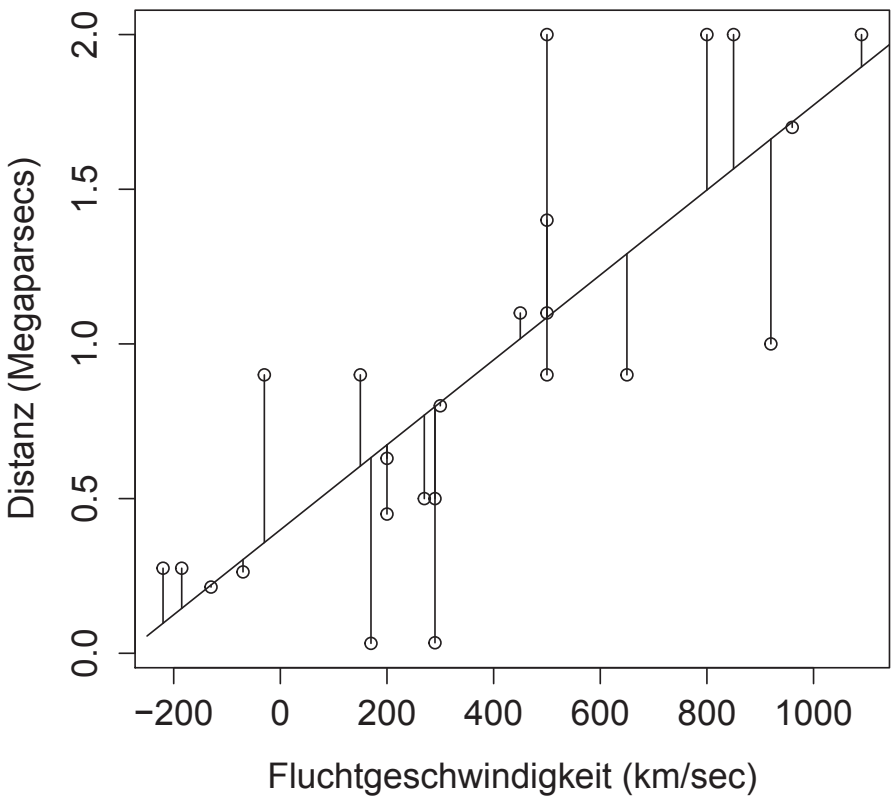
v. in *economy* for predicting economic well-being at a granular level using mobile data, satellite imagery, or Google Street View

vi. in *criminal justice* to predict whether an individual will commit a crime

What is Predictive Modeling?

- **Supervised machine learning** : software programs take as input training data sets and estimate or *learn* parameters that can be used to make predictions on new data
- Major challenge for using data to make predictions is distinguishing what is **meaningful** from **noise**
- With the Internet of things we can expect an explosion of diverse, heterogeneous data: ability to generate accurate predictions and high-quality analyses that include support for and evidence against predictions will be critical
- This course aims at providing the **fundamentals** to understand and apply predictive models

Example Linear Regression: Cosmology



- Edwin Hubble's 1929 article *A relation between distance and radial velocity among extra-galactic nebulae* marked a turning point in understanding the universe (dataset on the left)
- Big Bang Theory : there is a *linear relationship* between the distance y and the recession velocity x of a **galaxy**
- *Linear Model:*

$$y = \beta_0 + \beta_1 x$$

Example Linear Regression: Cosmology

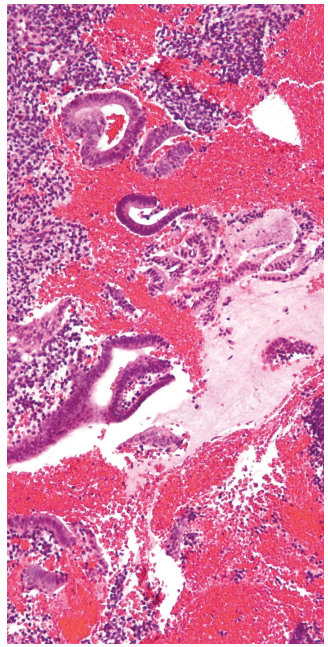
- *Linear Model:*

$$y = \beta_0 + \beta_1 x$$

- Hubble-Dataset: $\hat{\beta}_1 = 0.00137$ and $\hat{\beta}_0 = 0.39910$
- Age of the universe T corresponds to the parameter β_1 (Unit : megaparsec-second per kilometer) $\longrightarrow$ 1.34 billion years (in contradiction with radiometric results)
- Age of Universe according to *Planck 2015* data : 13.813 ± 0.038 billion years (with **uncertainty**!)

Example Classification: Tumor Type Prediction

- The small, round blue cell tumors (SRBCTs) contain various kinds of tumors that have a similar appearance on routine histology (cf. image on the right)
- Treatment varies widely between the tumor kinds. Thus exact diagnosis is important
- Traditional diagnosis is error prone



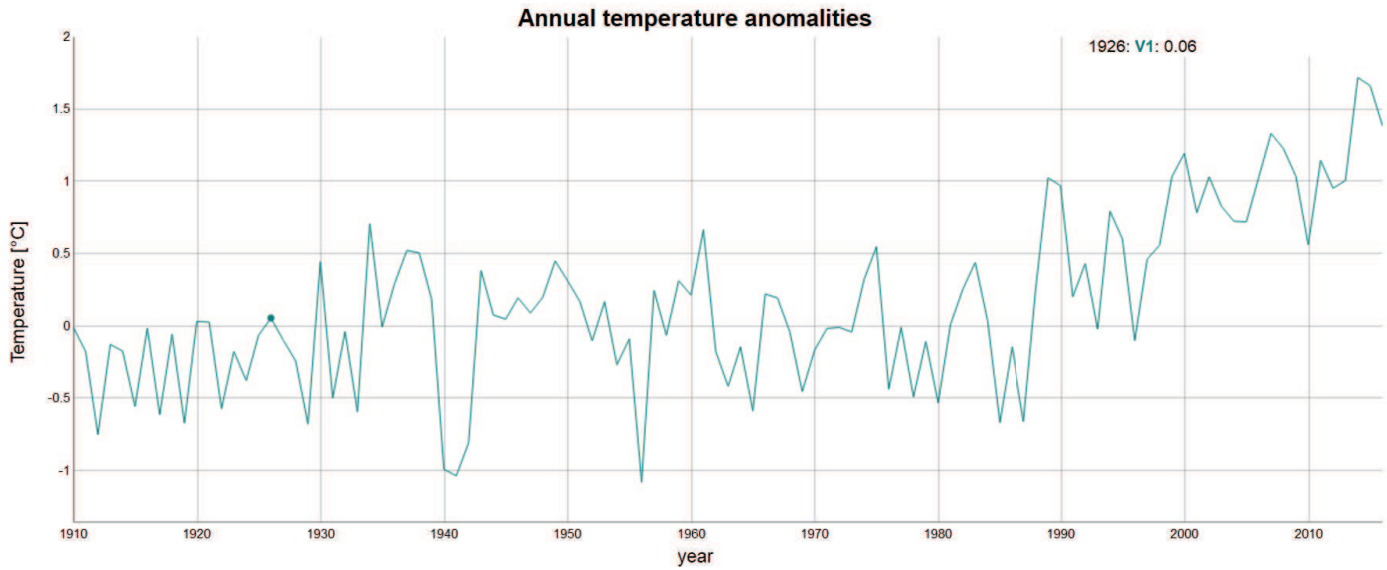
Example Classification: Tumor Type Prediction

Predictive Modeling: Use labeled gene micro array data of 6567 genes and learn a classification model on the data to predict the tumor type.

Tumor type	Var. 1	Var. 2	...	Var. 1480	...	Var. 6567
RMS	0.3	1.180	...	−0.08	...	−1.93
NB	0.679	1.289	...	−0.17	...	−2.00
⋮						
NB	0.31	0.48	...	0.48	...	1.48
⋮						
NHL	0.35	−0.27	...	−0.37	...	1.21

Example Time Series: Temperature Anomalities

- Figure below shows the *annual average temperature anomalities* with respect to the average between 1910 and 2000 in Europe.



- **Predictive Modelling:** Based on this time series and potentially further data, make a forecast of the temperature anomalities in the near future

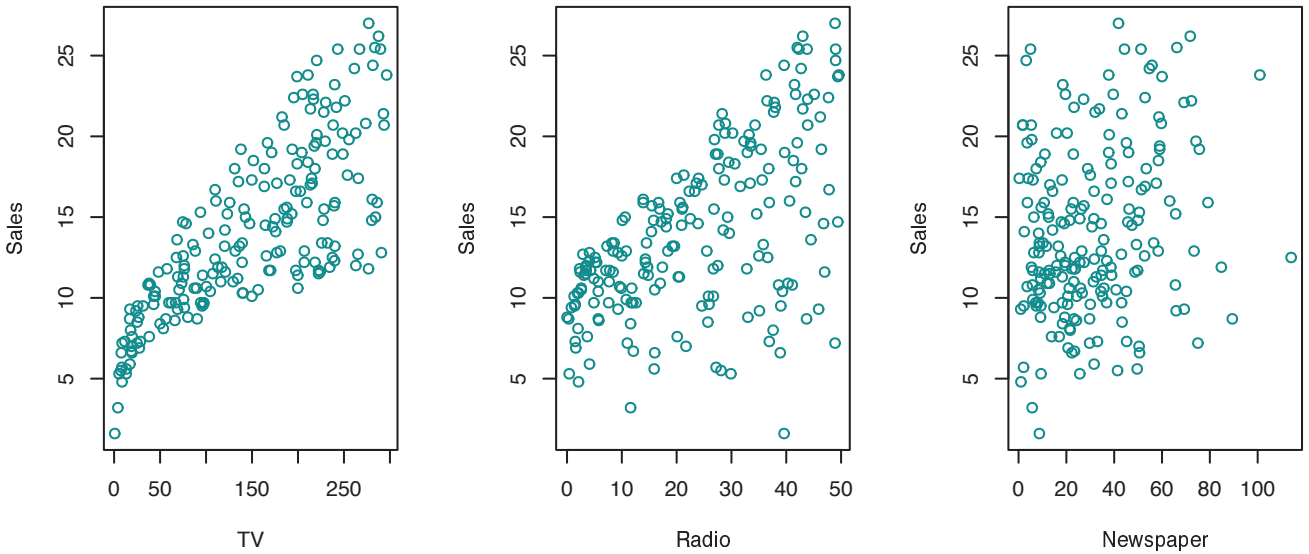
Organization of Predictive Modeling Course

- This course consists of **three** parts:
  - Linear Regression (SW 1-4)
  - Classification (SW 5-10)
  - Time Series Analysis (SW 11-14)
- **Lecture notes** are available on Moodle
- Weekly **exercises** including solutions are available on Moodle
- Relevant `Python` or `R` code available on `renkulab.io`

Introductory Example : `Advertising`

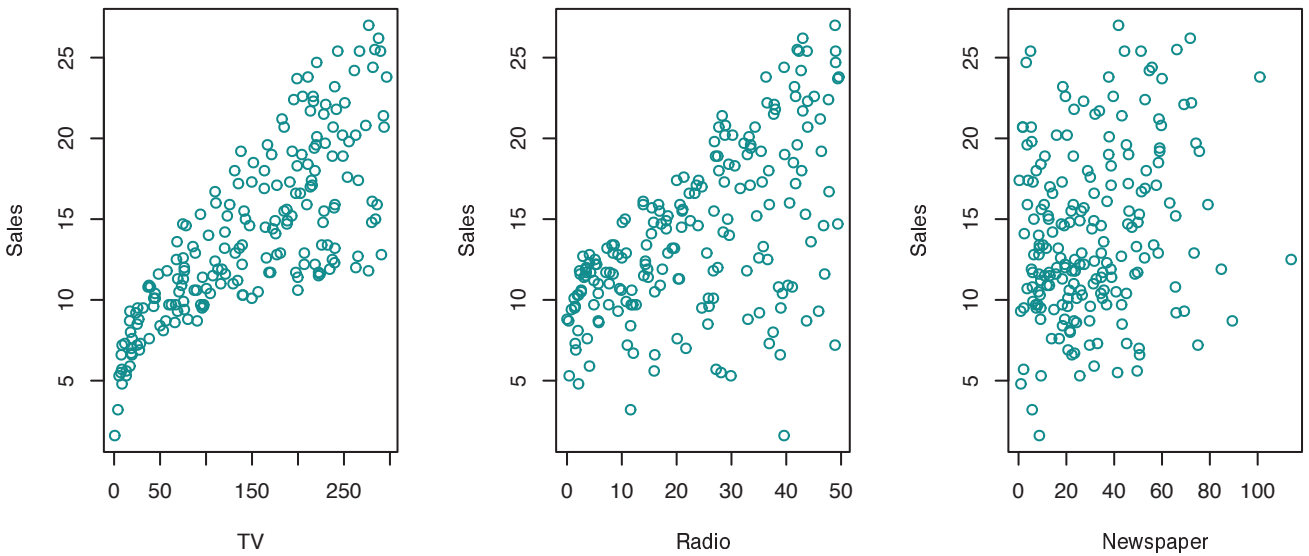
`Advertising` data set consists of

- `sales` (in thousand of units) of a product in 200 different markets
- advertising budgets for that product in each of those markets for three different media : `TV`, `radio` and `newspaper` (in thousands of CHF)



Introductory Example : `Advertising`

Goal: develop an accurate **model** to **predict** sales on the basis of the three media budgets `TV`, `radio` and `newspaper` to adjust advertising budgets and to increase `sales`



See Example `1.1` in the `Introduction to Regression Analysis` chapter.

Introductory Example : Mathematical Formulation

Required: Function f, that predicts the `sales` Y based on the three advertising budgets X_1 (`TV`), X_2 (`radio`) and X_3 (`newspaper`) :

$$Y \approx f(X_1, X_2, X_3)$$

- Y : **response variable** or **output variable**
- X_1, X_2 and X_3 are called **predictors** or **input variables**

Remark: In relation $Y \approx f(X_1, X_2, X_3)$ there is no equal sign : since the plots do not represent graphs of a function. f represents the relation between X_1, X_2, X_3 and Y only *approximately*.

Introduction

Statistical Regression Analysis

Regression analysis represents a statistical method to study and model the relationship between a **response variable** and **predictor variables**. Principal goal of regression analysis is to:

- *predict* data points based for some new values of the predictor variables (**prediction**)
- *understand* how the response variable is affected by a change of the predictor variables (**inference**)

Applications : Regression analysis is one of the major methods in data analysis and applied in many different fields

Statistical Regression Analysis

Assumption: There is a relationship between Y and X_1, X_2, X_3 and deviations are random

Mathematical Formalization:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- f is some fixed but *unknown* function of $X_1, X_2, \ldots, X_p$
- ε is a **random error term** which is:
  - independent of $X_1, X_2, \ldots, X_p$
  - has mean zero

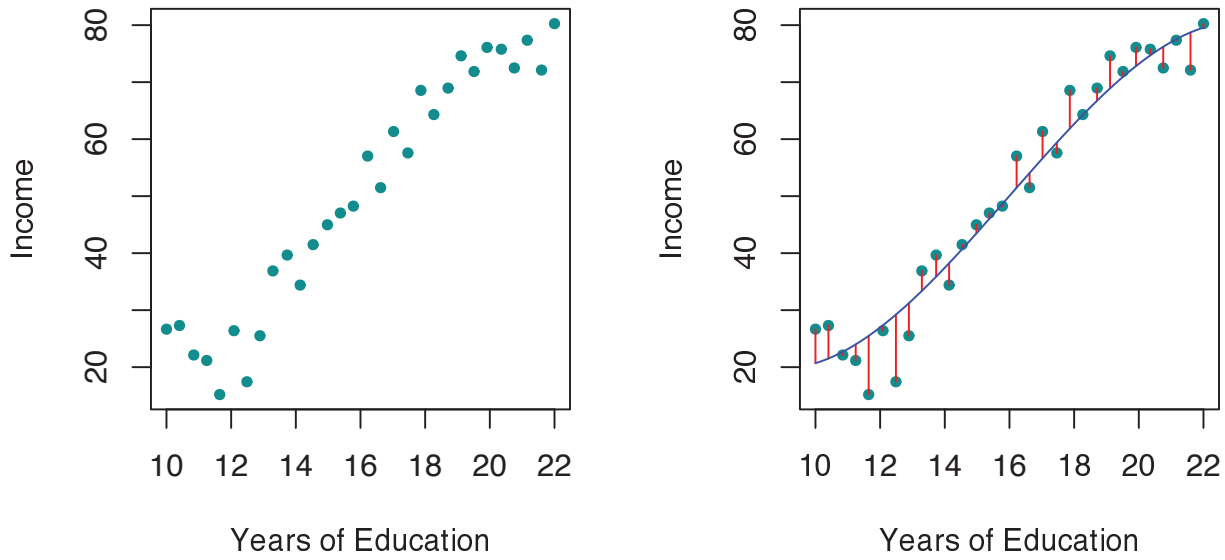
Remarks:

- Error ε is described in terms of a probability distribution
- Quantity ε may contain **unmeasured** or **unmeasurable** variables
- Simple assumption : f is linear. Linear regression allows as well to fit non-linear curves.

Example : Income as function of years of education

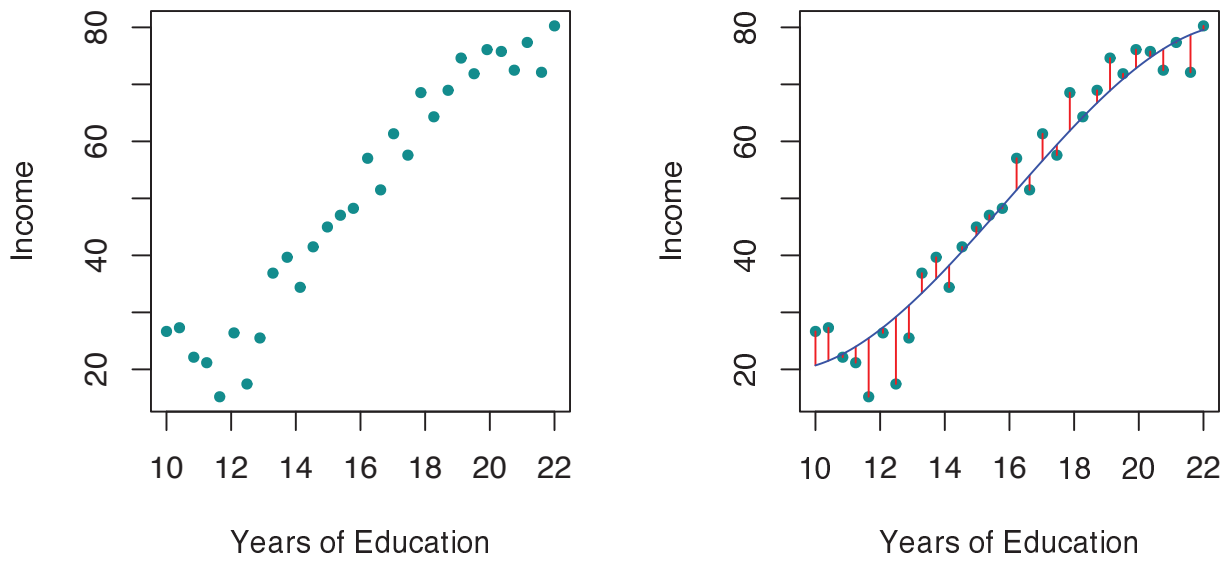
Simulated data set `Income` consists of

- `income` of 30 individuals
- `years of education` (in years)



Example : Income as function of years of education

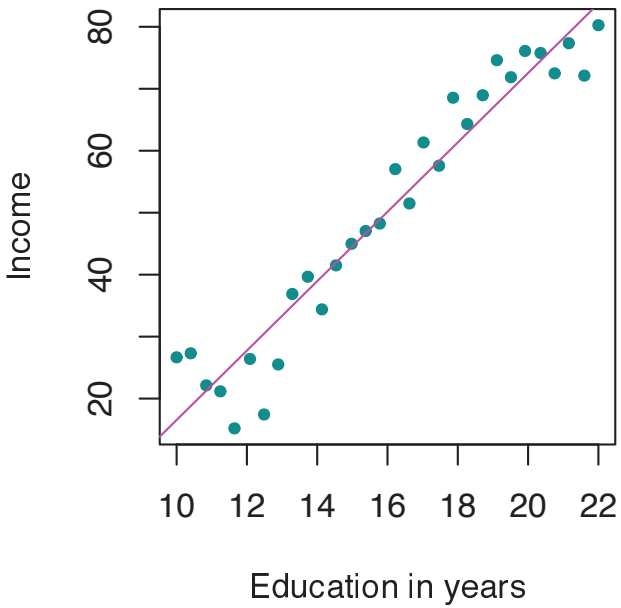
Since the data set `Income` is simulated, the function f is known and is shown by the blue curve in the right-hand panel of the figure below. Vertical (red) lines represent the error terms ε (mean value is approximately 0).

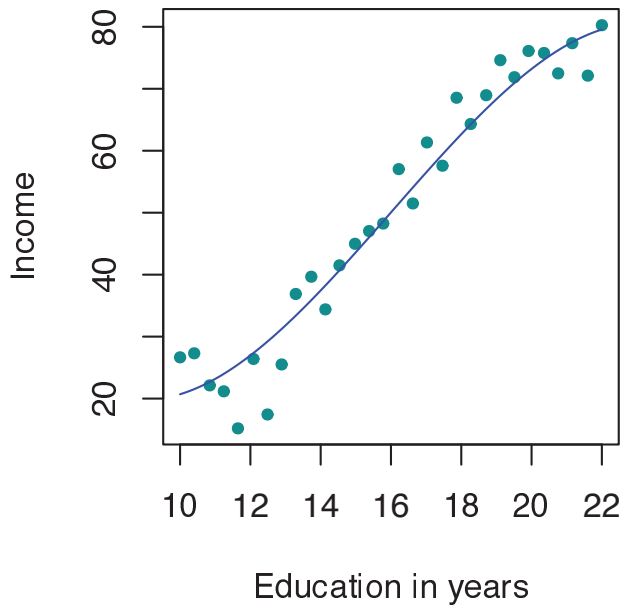


Example : Income as function of years of education

Question: Which *model* do we select or which is the form f should have?

- Linear model : $f(X) = \beta_0 + \beta_1 X$
- Cubic model : $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$





- How do you interpret β_0 and β_1 ?
- How do you estimate f resp., β_0 and β_1?

Questions Surrounding Regression Analysis

`Advertising` data set : `sales` for a specific product as a function of the advertising budgets for the media `TV`, `radio` and `newspaper`

- *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 Model

We assume that there is approximately a **linear** relationship between X and Y:

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

- β_0 represents the **intercept** of the regression line
- β_1 measures the **slope** of the regression line
- $\approx$ reads "is approximately modeled as"

Simple Linear Regression Model

We assume that there is approximately a **linear** relationship between X and Y:

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

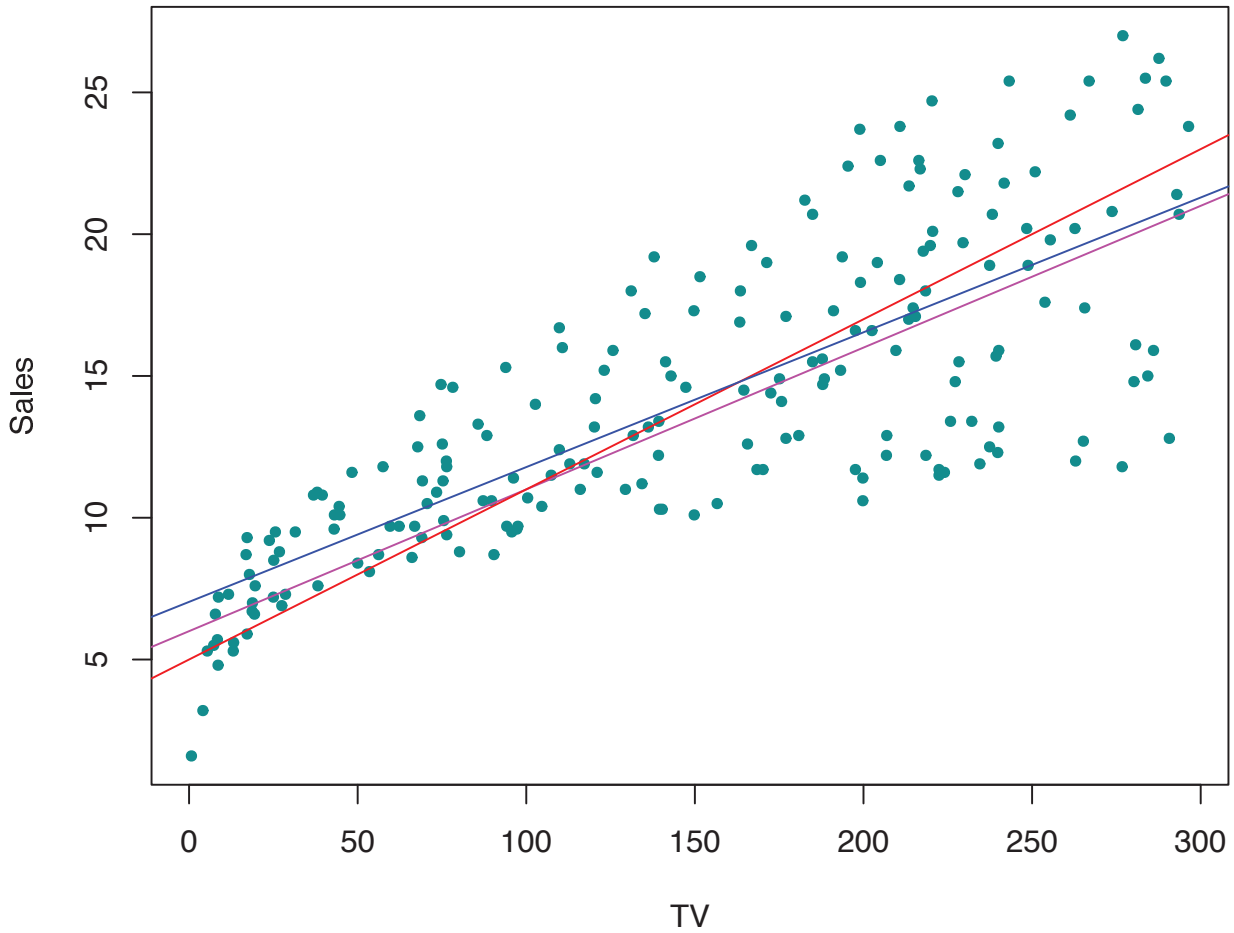
Example: In the example with the `Advertising` data set, X may represent `TV` advertising and Y may represent `sales`. Then, we can regress `sales` onto `TV` by fitting the model

$$\texttt{sales} \approx \beta_0 + \beta_1 \cdot \texttt{TV}$$

Regression coefficients or **parameters** β_0 and β_1 are estimated on the basis of the data: coefficient estimates are denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$

Estimating the Coefficients

Goal: to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ fits the available data as well as possible



Least Squares Method

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X
- ith **residual** : $r_i = y_i - \hat{y}_i$
  difference between the ith observed response value and the ith response value that is predicted by our linear model
- **Least squares** approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the **residual sum of squares** (RSS)

$$\text{RSS} = r_1^2 + r_2^2 + \ldots + r_n^2$$

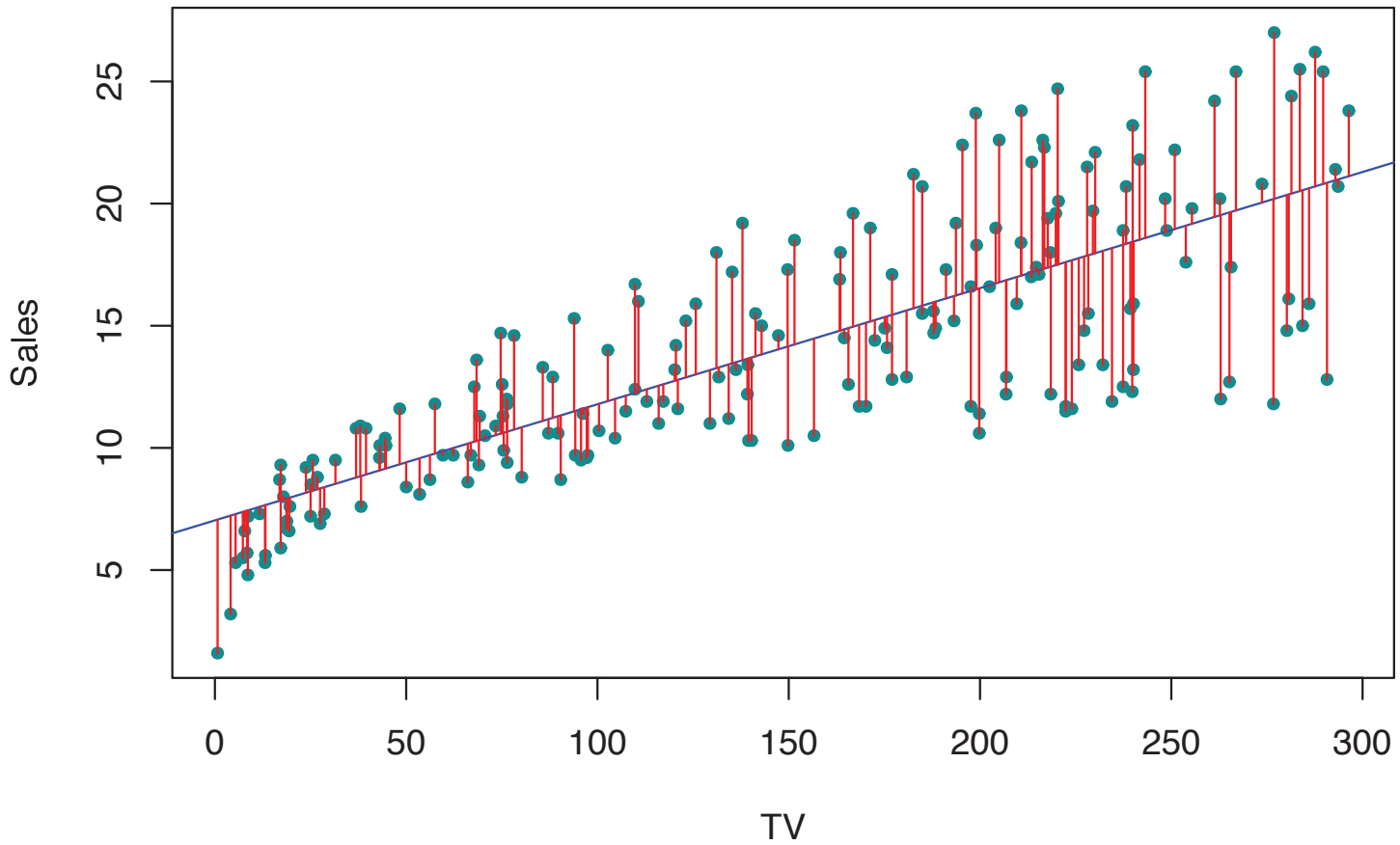
or equivalently

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

$\longrightarrow$ derivative with respect to $\hat{\beta}_0$, resp. to $\hat{\beta}_1$ and set expressions equal to zero!

Example : Advertising

See Example 2.4 in the Simple Linear Regression chapter for the Advertising data set



Estimating the Coefficients

Least Squares Coefficient Estimates

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

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

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

How much do coefficient estimates scatter?

Example:

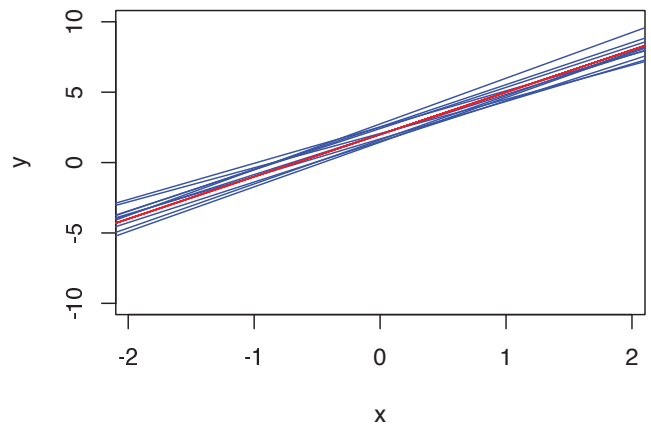
- In this example, we assume that we know the **true** relationship between X and Y : $f(X) = 2 + 3X$
- We **simulate** X and Y from the model

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

where ε is normally distributed with mean 0, thus $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We create 100 random values of X and generate the corresponding values of Y

- Every simulation can be considered as a **random experiment**
- After every experiment, we estimate the coefficients β_0 and β_1 and plot the corresponding regression line
- See Example `3.1` in the `Simple Linear Regression` chapter

How much do coefficient estimates scatter?



$\hat{\beta}_0$ and $\hat{\beta}_1$ are **scattered** around the true values β_0 and β_1 with

Standard Error

$$\text{se}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right) \quad \text{and} \quad \text{se}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where $\sigma^2 = \text{Var}(\varepsilon)$

Residuals and Estimation of Variance

In general, σ^2 (variance of the error term ε) is **not** known: but can be estimated on the basis of the data. The **error term** ε

- cannot be observed
- cannot be derived from $\varepsilon = Y - (\beta_0 + \beta_1 X)$ since β_0 and β_1 are unknown.
- *Approximation* for ε : residuals $r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Residuals and Estimation of Variance

- *Approximation* for ε : residuals $r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
- Estimation of σ

$$\hat{\sigma} = \text{RSE} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{r_1^2 + r_2^2 + \ldots + r_n^2}{n-2}}$$

- Factor $1/(n-2)$ is chosen so that the estimate of σ turns out to be **unbiased**
- This estimate is known as the **residual standard error** (RSE)
- See Example `3.3` in the `Simple Linear Regression` chapter

Hypothesis Test

Most common hypothesis test:

- H_0 : There is **no** relationship between X and Y
- H_A : There is **some** relationship between X and Y

Mathematically, this corresponds to testing:

- H_0 : $\beta_1 = 0$ versus
- H_A : $\beta_1 \neq 0$

If $\beta_1 = 0$, then model reduces to $Y = \beta_0 + \varepsilon$ and X is **not** associated with Y

Hypothesis Test

- To test the null hypothesis, we need to determine whether $\hat{\beta}_1$ is sufficiently far from zero that we can be confident that β_1 is non-zero.
- **Question** : *How* far from 0 is far enough?
- **Answer** : It depends on se($\hat{\beta}_1$)
  - if se($\hat{\beta}_1$) is **large**: then $\hat{\beta}_1$ must be large in absolute value, to reject H_0
  - if se($\hat{\beta}_1$) is **small**: then even relatively small values of $\hat{\beta}_1$ may provide evidence that $\beta_1 \neq 0$, that is to reject H_0

Hypothesis Test, Test statistic and P-Value

- **Test statistic** : $T = \dfrac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)}$: measures the number of standard deviations that $\hat{\beta}_1$ is away from 0
- If there is really **no** relationship between X and Y, that is H_0 is **true**, then we expect that T follows a t-distribution with $n-2$ degrees of freedom
- We perform an experiment and measure the realization t of the test statistic T
- **p-value** : probability of observing any value of T larger than $|t|$
  - If p-value is **smaller** than α (typically $\alpha = 0.05$), then we **reject** H_0 and conclude: there is a relationship between X and Y
  - If p-value is **larger** than α, then we **retain** H_0 (there is no relationship between X and Y)

t-Test in Linear Regression

1. **Model**:
$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

2. **Null Hypothesis**: $H_0 : \beta_1 = 0$
Alternative Hypothesis: $H_A : \beta_1 \neq 0$ (two-sided Test)
3. **Test statistic**:
$$T = \frac{\text{observed} - \text{expected}}{\text{estimated standard error}} = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)}$$
Null Distribution assuming H_0 is true: $T \sim t_{n-2}$
4. **Significance Level**: α
5. **Rejection Region for Test Statistic**:
$$K = \left(-\infty, t_{n-2;\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}}, \infty\right)$$

6. **Test Decision**: Verify whether observed t falls into rejection region

Example `Advertising`

Realization of test statistic T

$$t = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{0.047537 - 0}{0.002691} = 17.66518$$

This value can be found as well under `t value` in the output of Example `3.3` in the `Simple Linear Regression` chapter

Example `Advertising`

- $\hat{\beta}_1$ is approximately 18 standard errors $\mathrm{se}(\hat{\beta}_1)$ away from 0
- **p-value**: probability of observing a value of the t-statistic larger than $|t| = 17.66518$
- Assuming $\beta_1 = 0$, T will follow a t-distribution with $n - 2 = 198$ degrees of freedom, hence the **two-sided p-value** is approximately 0 (see the output in Example `3.4` in the `Simple Linear Regression` chapter)
- See example `3.3` in the `Simple Linear Regression` chapter: In table `Coefficients` under `Pr(>|t|)` the p-value $< 2 \cdot 10^{-16}$ is listed

Confidence Intervals

Confidence Interval

A 95 % confidence interval is defined as a range of values such that with a 95 % probability, the range will contain the true unknown parameter. The range is defined in terms of lower and upper limits computed from the sample of data.

Confidence Intervals for Linear Regression

- For linear regression, the 95 % confidence interval for β_1 takes approximately the form

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

- **Remark**: Exact formula for the 95 % confidence interval is obtained by replacing the factor 2 by $t_{0.975;n-2}$; $t_{0.975;n-2}$ is the 0.975 quantile of a t-distribution with $n-2$ degrees of freedom

Example : `Advertising`

- See example `3.5` in the `Simple Linear Regression` chapter
- 95 % confidence interval for β_0 : $[6.130, 7.935]$
- 95 % confidence interval for β_1 : $[0.042, 0.053]$
- In the absence of any advertising: sales will fall somewhere in between of 6130 and 7935 units
- For each CHF 1000 increase in `TV` advertising, there will be an average increase in `sales` of between 42 and 53 units

Confidence Interval for Response - `Advertising` Data Set

- What is the value of `sales`, if 100 000 CHF is spent for `TV` advertising?
  - **Answer**: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_0$ where $\hat{y}_0$ is the predicted value of `sales`, $\hat{\beta}_0 = 7.032594$ and $\hat{\beta}_1 = 0.047537$; we thus find $\hat{y}_0 = 11786$ units

Confidence Interval for Response - `Advertising` Data Set

- We now want to compute the expected value of the predicted response `sales` if 100 000 CHF is spent for `TV` advertising

- **Expected value** of $\hat{y}$ for a given value x_0 of the predictor then is $\mathrm{E}[\hat{y}|x_0] = \mathrm{E}[\hat{y}_0] = \mathrm{E}[\hat{\beta}_0 + \hat{\beta}_1 \cdot x_0] = \beta_0 + \beta_1 x_0$; this corresponds to the **true** value $y_0 = f(x_0)$ which is unknown
  - **Answer**: For the true value y_0, we only can determine a **confidence interval**, in this case an approximate 95% confidence interval

$$[\hat{y}_0 - 2 \cdot \mathrm{se}(\hat{y}_0), \hat{y}_0 + 2 \cdot \mathrm{se}(\hat{y}_0)]$$

where

$$\mathrm{se}(\hat{y}_0)^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$$

Confidence Interval for the True Predicted Value of $\hat{y}_0$

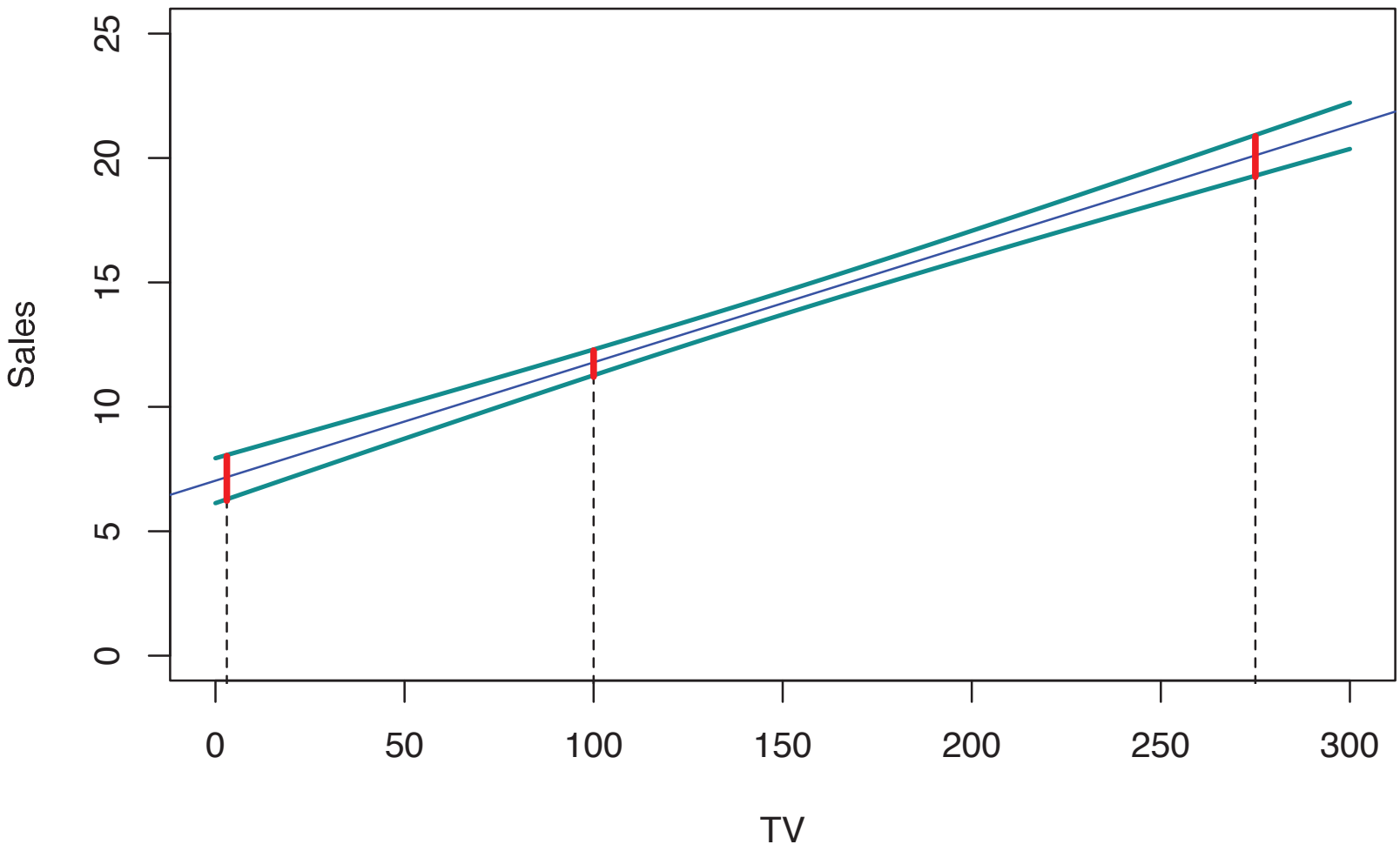
- `Advertising` data set : we determine the 95 % confidence intervals for the following values of `TV`: 3, 100 and 225
  - For $x_0 = 100$ the 95 % confidence interval is given by

$$[11.268, 12.305]$$

  - The expected value of $\hat{y}$ given the predictor $x_0 = 100$, that is $\mathrm{E}[\hat{y}|x_0]$, is contained in this interval with a probability of 95 %

- See example `4.1` in the `Simple Linear Regression` chapter

Confidence Band for the `Advertising` Data Set



Prediction Intervals

Question: Given that 100 000 CHF is spent on `TV` advertising, which interval contains with a probability of 95 % the true value of `sales` for a particular city?

- To indicate this interval we now have to account as well for the scatter of the data points around the regression line which is expressed by the error term ε in our regression model

- Standard error of future observation y_0 :

$$\text{se}(y_0)^2 = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$$

- 95% **prediction interval** for future observation

$$[\hat{y}_0 - 2 \cdot \text{se}(y_0), \hat{y}_0 + 2 \cdot \text{se}(y_0)]$$

Prediction Interval : Example `Advertising`

In the case of the `Advertising` data we determine the 95 % prediction interval for the x-values 3, 100 and 225.

- For the predictor value $x_0 = 100$ the 95 % **prediction interval** is given by

$$[5.339, 18.233]$$

- A future observation y_0 for given $x_0 = 100$ will fall with a probability of 95 % into this interval
- Prediction interval thus is clearly **larger** than the confidence interval for the expected value of $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x$
- See example `4.4` in the `Simple Linear Regression` chapter

Prediction Band : Example `Advertising`