Recap

Lecture 2

Linear Regression

ISLR 3, ESL 3

Krikamol Muandet
Jilles Vreeken

UNIVERSITÄT DES SAARLANDES

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

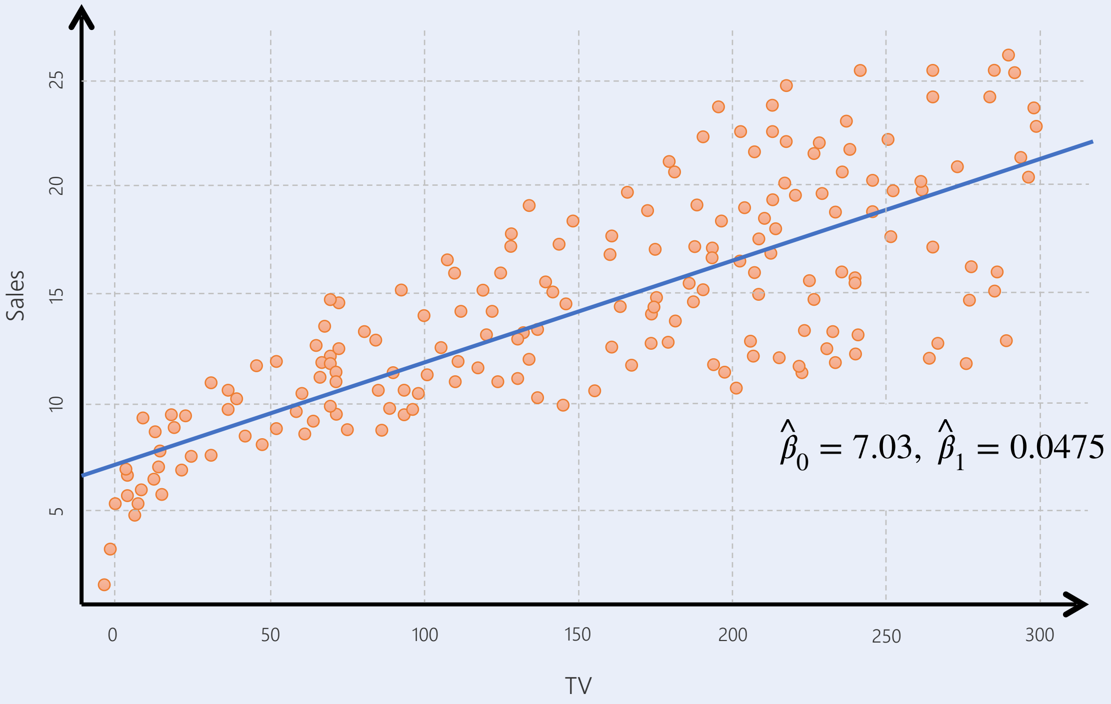
Recap (Lecture 2)

- Simple Linear Regression
- Assessing the goodness of fit
- Multiple Linear Regression

Recap (Lecture 2)

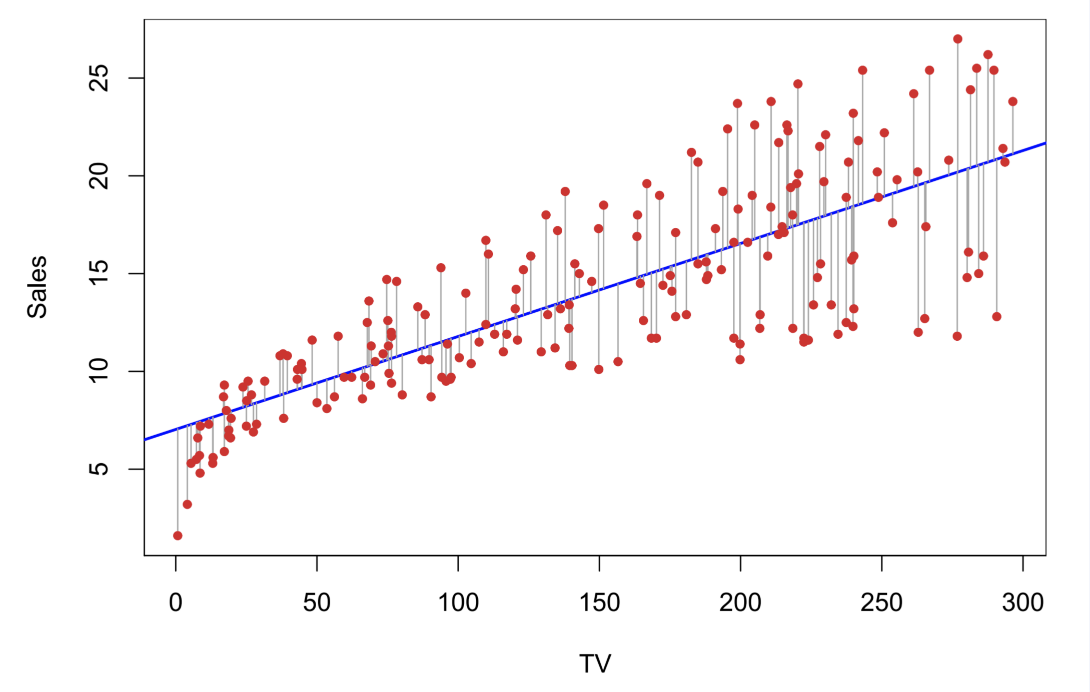
- Simple Linear Regression
  - Looking for linear relationships between a predictor (e.g. TV ads) and outcome (e.g. sales)
  - Estimating the linear coefficients: least-square fit
  - Gauss-Markov theorem: among all unbiased linear estimators, the least-square fit is the one with the smallest variance
- Assessing the goodness of fit
- Multiple Linear Regression

Looking for Linear Relationships: Coefficients



Linear fit of the advertising data appears appropriate for all but the smallest advertising budgets

Looking for Linear Relationships: Errors



$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2 = \left(y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right)^2 + \left(y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_2)\right)^2 + \ldots + \left(y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)\right)^2$$

Estimating the linear coefficients

Minimizing the RSS for a single predictor X, we obtain

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

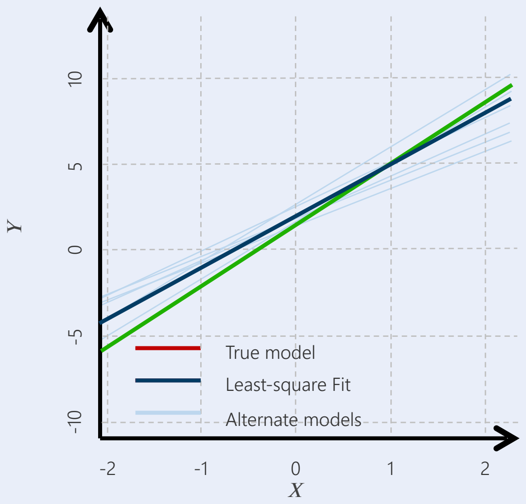
$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

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

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

Gauss-Markov Theorem

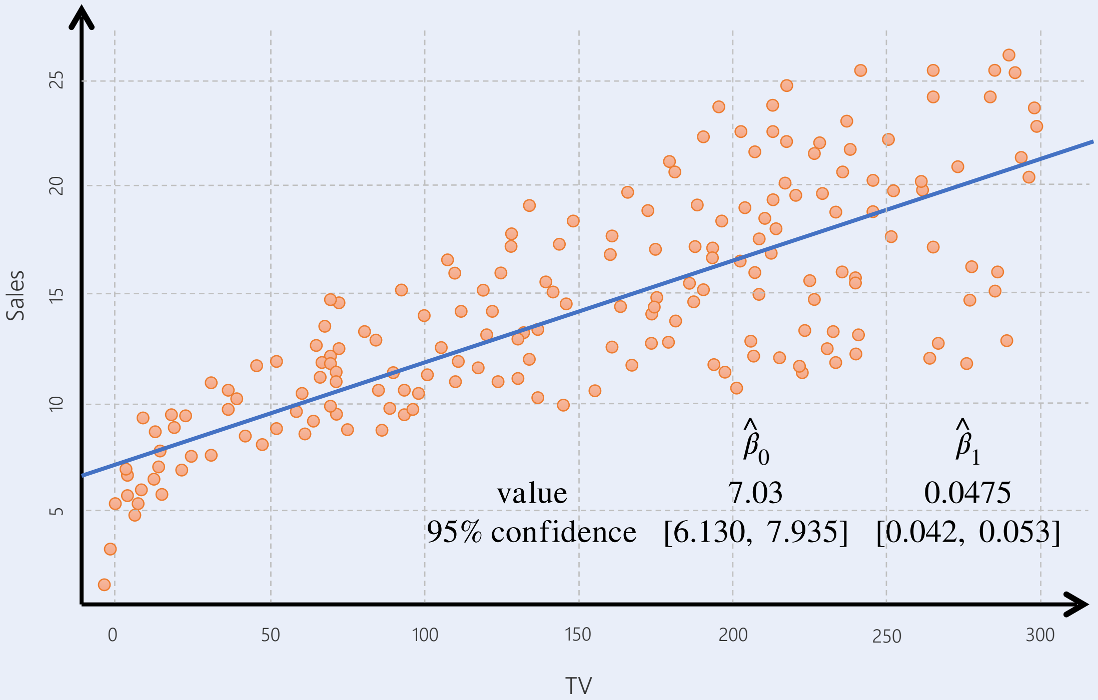
- The least-square fit is an unbiased estimate for the population regression line
- Among all unbiased linear estimators, the least-square fit is the one with smallest variance



Recap (Lecture 2)

- Simple Linear Regression
- Assessing the goodness of fit
  - Standard Error: How accurately does least-squares estimate the population coefficients?
  - Confidence Intervals & Hypothesis Testing: How confident are we about the estimates?
  - Other Scores: RSS, R2 and Correlation
- Multiple Linear Regression

Confidence Intervals

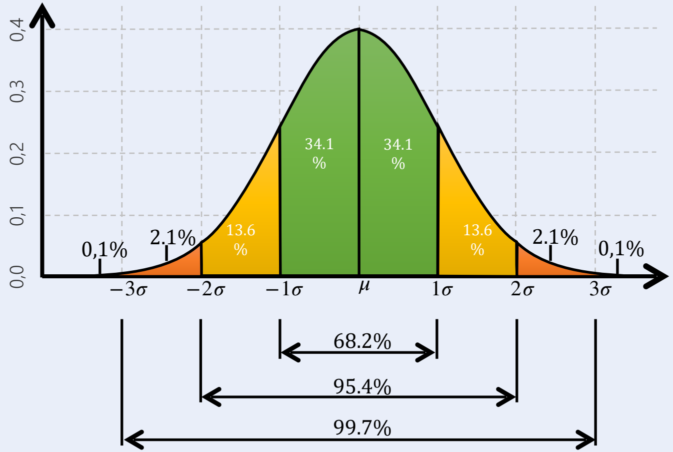


Linear fit of the advertising data appears appropriate for all but the smallest advertising budgets

Confidence Intervals & Hypothesis Testing

We can determine the probability that $|t|$ exceeds a certain value from the figure on the right

- for $|t| > 2$ it is roughly 5%
- this probability is called the p-value



If a p-value is small, it is unlikely that the observed association of input and output is due to chance

- a p-value of 5% means that, if the null-hypothesis holds, an equal or better result will happen in at most 5% of all datasets
- we reject the null hypothesis at a significance level α if the p-value $\leq \alpha$

Typical significance levels α for rejecting the null hypothesis are 5% and 1%

- the figure shows the values for $n = 30$

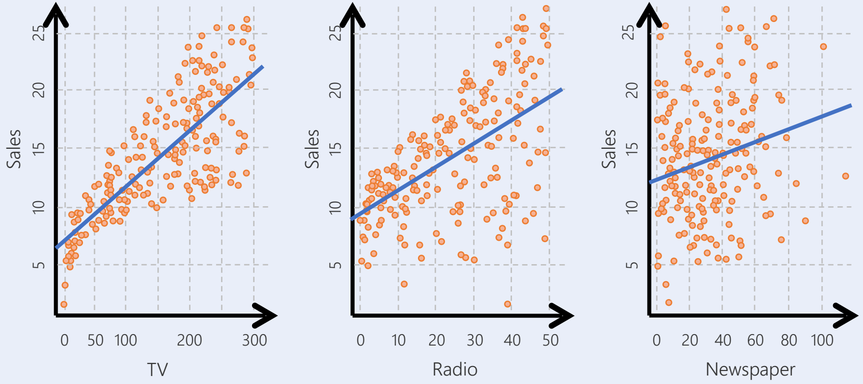
II

(for $|t| > 2$ the p-value is 4.6%, as 100 – 2 (34.1+13.6) = 4.6%)

Recap (Lecture 2)

- Simple Linear Regression
- Assessing the goodness of fit
- Multiple Linear Regression
  - Looking for linear relationships between multiple predictors and an outcome
  - Estimating the linear coefficients: least-square fit in the multivariate case

Looking for Linear Relationships: univariate vs. multivariate



Univariate regression

For each value of the considered input, ignore the values of all other features

	Coefficient	Std. error	-statistic	-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

Multivariate regression

For each value of the considered input, keep the values of all other features fixed

Estimating the linear coefficients: univariate vs. multivariate

- Minimizing the RSS for a single predictor X, we obtain

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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

- For multivariate $\mathbf{X}$, we obtain (assuming that $\mathbf{X}^T\mathbf{X}$ is positive definite, i.e. for all $a \in \mathbb{R}_n$, $a^T\mathbf{X}^T\mathbf{X}a > 0$):

$$\hat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T Y$$

Univariate regression

For each value of the considered input, ignore the values of all other features

Multivariate regression

For each value of the considered input, keep the values of all other features fixed