Lecture 2

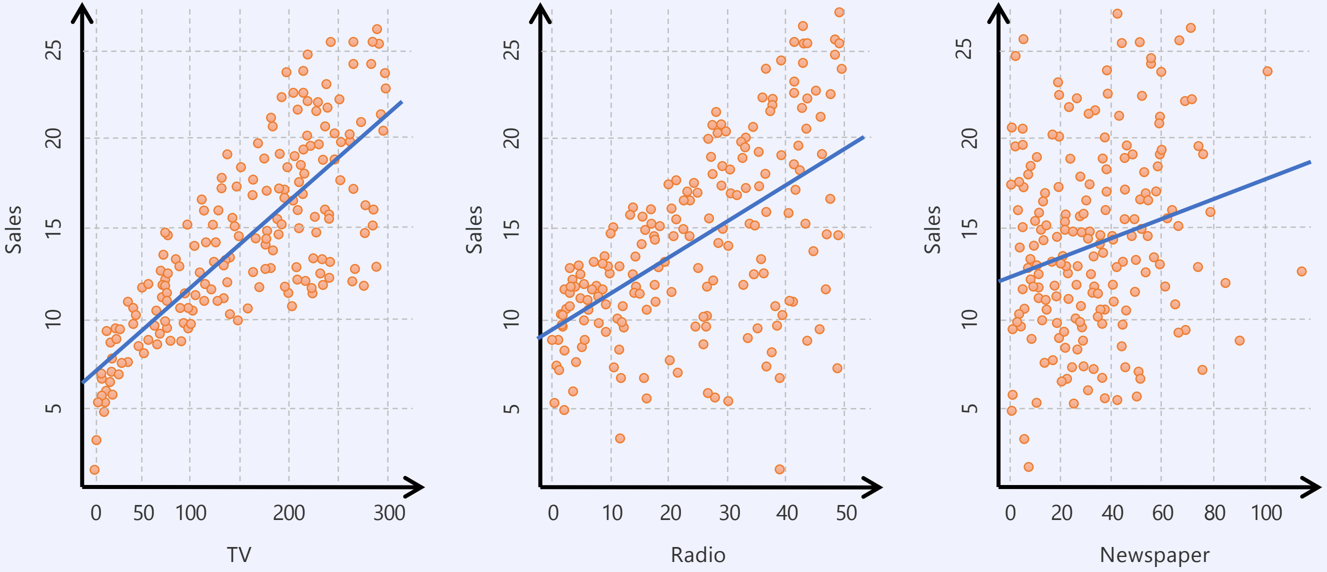
Linear Regression

ISLR 3, ESL 3

Krikamol Muandet
Jilles Vreeken

UNIVERSITÄT DES SAARLANDES
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Looking for Linear Relationships



Numbers are in thousands of dollars
In general, sales increase as advertising is stepped up.
The blue lines result from least-squares linear regression
to the variable along the x-axis

Questions

1. Is there a relationship between advertising budget and sales?
   - if the evidence is weak, advertising may not be effective
2. How strong is the relationship between advertising and sales?
   - can sales be predicted accurately based on the advertising budget?
3. Which media contribute to sales?
   - are all three media effective?
4. How accurately can we estimate the effect of a medium on sales?
   - what is the expected range of sales increase per dollar spent on a medium?
5. How accurately can we predict future sales?
6. Is the relationship in fact linear?
7. Is there synergy among advertising media?

Simple Linear Regression

ISLR 3.1, ESL 3.2

Simple Linear Regression

We assume that X and Y are related as $Y \approx \beta_0 + \beta_1 X$

- for example, $sales \approx \beta_0 + \beta_1 \times TV$
- the estimated value of Y for input $X = x_i$ is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- the intercept, β_0, and slope, β_1, are coefficients or parameters
- this is also known as simple or univariate linear regression

Given training data set of n observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Goal estimate the unknown coefficients β_0 and β_1 such that

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

for all $i = 1, \ldots, n$ and for future values of x

Estimating the Coefficients

We measure the deviation of the estimate to the true value by a loss function

- let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, then $e_i = y_i - \hat{y}_i$ is the residual

In regression, we mostly use the residual sum of squares (RSS)

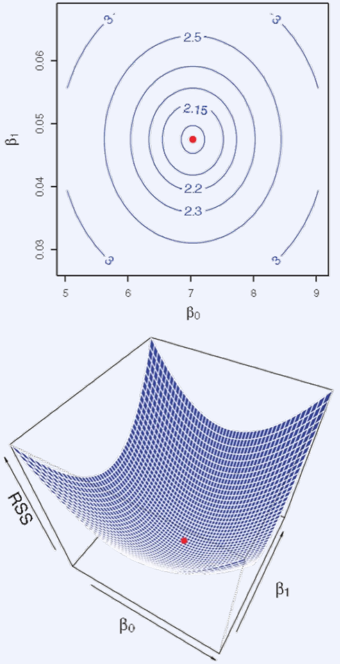
$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$= \left(y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right)^2 + \left(y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_2)\right)^2 + \cdots + \left(y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)\right)^2$$

- this function is quadratic in β_0 and β_1
- setting its derivative to zero yields the least-square coefficient estimates

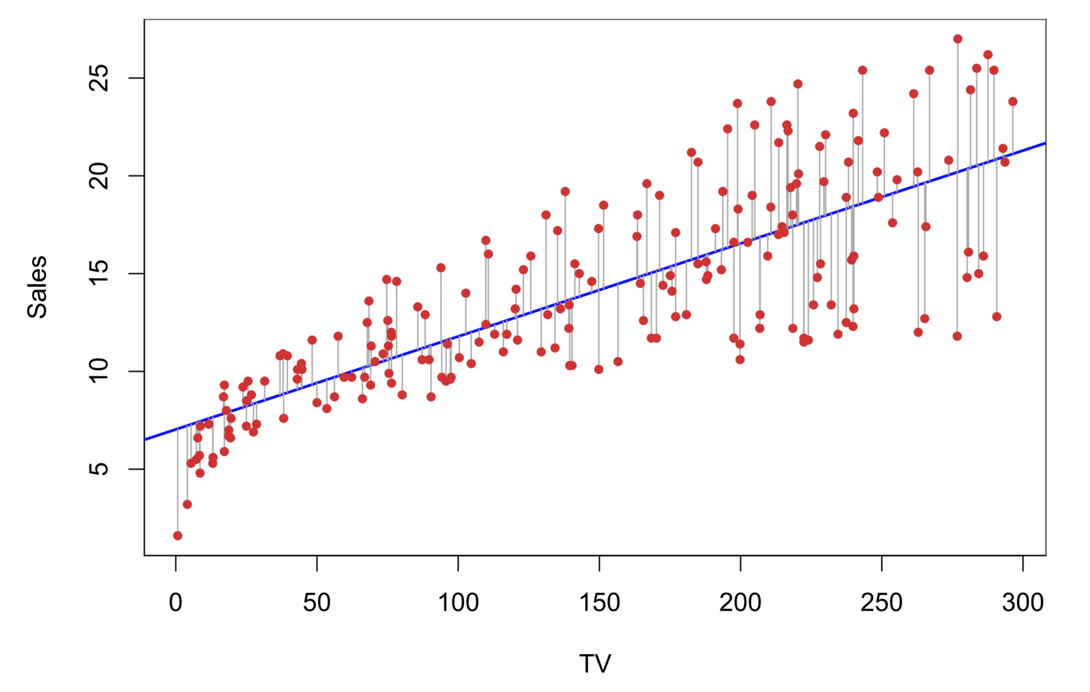
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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



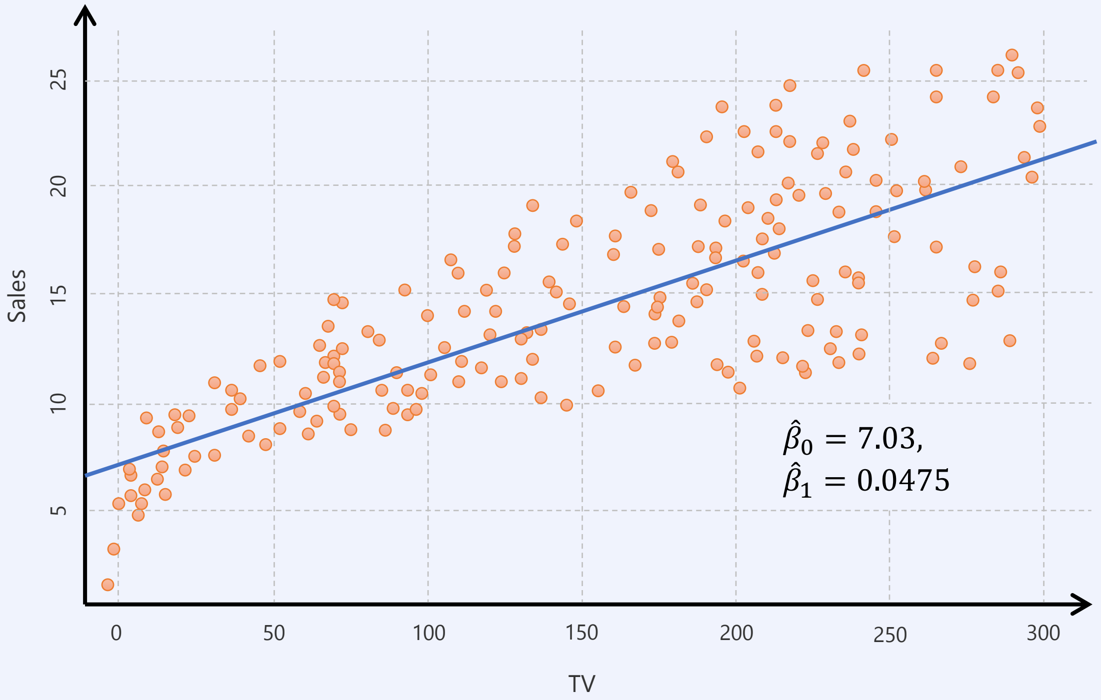
Contour and 3D plots of the RSS

Estimating the Coefficients



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

Estimating the Coefficients



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

Accuracy of Coefficient Estimates

We assume the true relationship includes **noise** that is **independent** from the observations

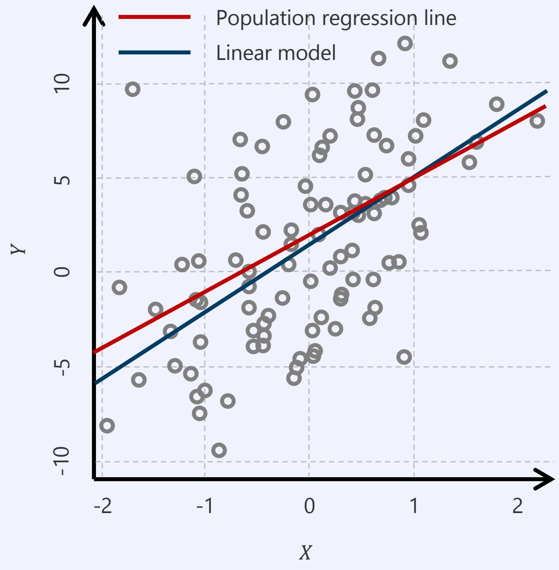
$$Y = \beta_0 + \beta_1 X + \epsilon \qquad (*)$$

- if this is true, the population regression line is the best linear approximation to the relationship between X and Y
- the population regression line is usually unobserved

The least-squares fit on the training data is given by

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

- the fit depends on the (finite!) training data



Least-squares fit (blue) and population regression line (red) on simulated data $Y := 2 + 3X + \epsilon$ with Gaussian error ϵ with 0–mean

Accuracy of Coefficient Estimates

We assume the true relationship includes **noise** that is **independent** from the observations

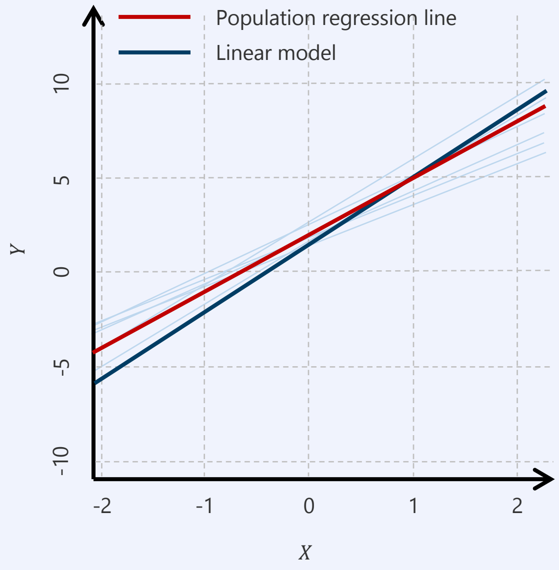
$$Y = \beta_0 + \beta_1 X + \epsilon \qquad (*)$$

- if this is true, the population regression line is the best linear approximation to the relationship between X and Y
- the population regression line is usually unobserved

The least-squares fit on the training data is given by

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

- the fit depends on the (finite!) training data



Least-squares fit on ten different randomly chosen training data sets

Unbiased Estimates

How do we estimate the mean μ of a random variable Y?

- the sample estimate over a finite set of observations is the average

$$\operatorname{avg}(y_1, y_2, \dots, y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

- on average, we have $\bar{y} = \mu$
- $\bar{y}$ is an **unbiased estimate** for μ

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 the **smallest variance**
- Gauss-Markov Theorem; if you learn one thing from EML, this should be it.

Assessing the Accuracy of Estimates

How accurately does $\hat{\mu}$ estimate μ?

- assuming every sample is independent, we have the standard error of $\hat{\mu}$

$$SE(\hat{\mu}) = \sqrt{Var(\hat{\mu})} = \sqrt{\sigma^2/n}$$

- where n is the number of samples, and σ is the population standard deviation
- the more samples, the smaller the standard error

The standard errors of the least-square coefficients β_0 and β_1 are

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \qquad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- we again assume that errors are independent, uncorrelated, and have a common variance $\sigma^2 = Var(\epsilon)$

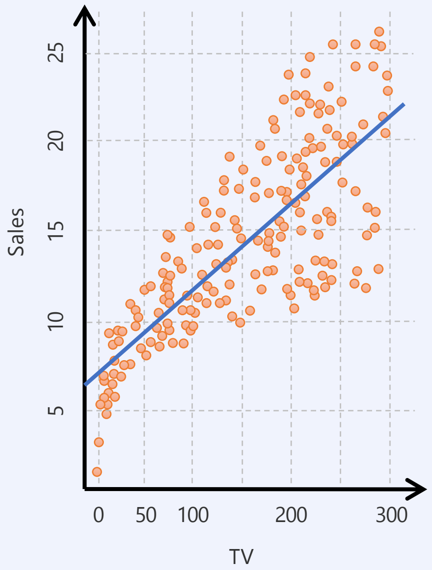
(assuming i.i.d., that samples are independently and identically drawn from the same distribution is one of Machine Learning's big lies)

Assessing the Accuracy of Estimates

Observations

1. $SE(\hat{\beta}_1)$ decreases as the $\boldsymbol{x_i}$ are more spread out, making the slope is the easier to determine
2. $SE\left(\hat{\beta}_0\right) = SE(\hat{\mu})$ if $\bar{x} = 0$ in which case $\hat{\beta}_0 = \bar{y}$
3. $\boldsymbol{\sigma}$ is generally not known, but, we can provide a sample estimate for it: the residual standard error

$$RSE = \sqrt{RSS/(n-2)}$$



Computing Confidence Intervals

The famous 95% confidence interval

- interval that with 95% probability contains the true value
- we compute the limits from the sample (training) data
- for linear regression coefficient $\hat{\beta}_0$ we have

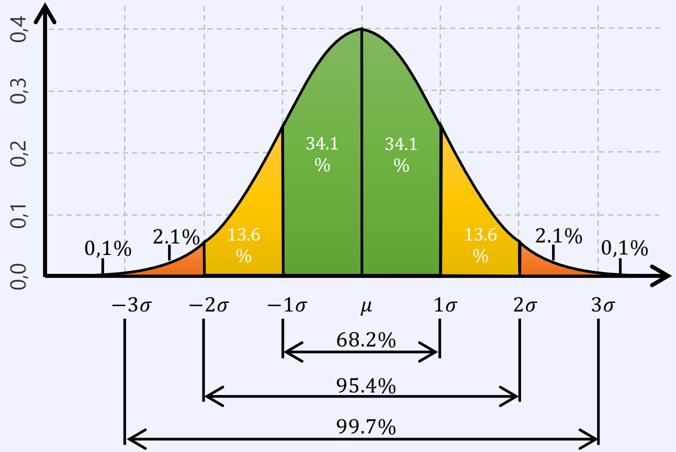
$$[\hat{\beta}_0 - 2 \cdot SE(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot SE(\hat{\beta}_0)]$$

- while for $\hat{\beta}_1$ we analogously have

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

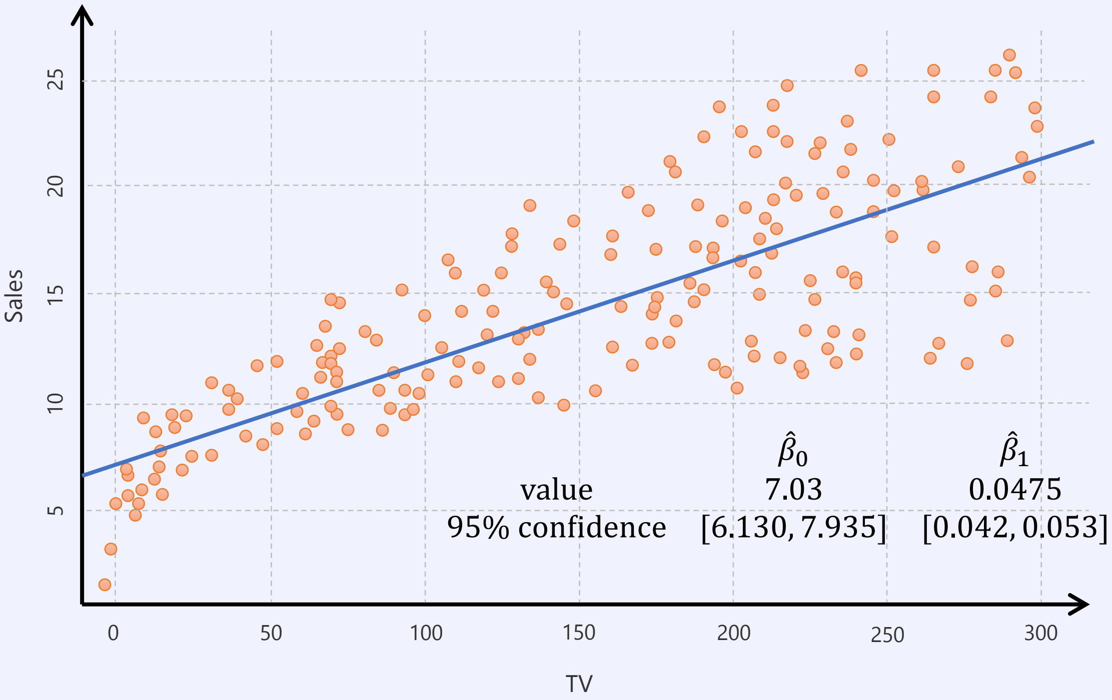
Why is this the case?

- we assume that the error in the output is Gaussian distributed
- the coefficient estimates are then also Gaussian distributed (!)



Probability mass in a Gaussian

Example Advertising Data



	$\hat{\beta}_0$	$\hat{\beta}_1$
value	7.03	0.0475
95% confidence	$[6.130, 7.935]$	$[0.042, 0.053]$

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

Hypothesis Testing

When can we determine if there is a significant relationship between X and Y?

- we can statistically test the null hypothesis $\boldsymbol{H_0}$ against the alternative hypothesis $\boldsymbol{H_a}$
- in our setting, this means testing $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

How do we determine if β_1 is far enough from zero?

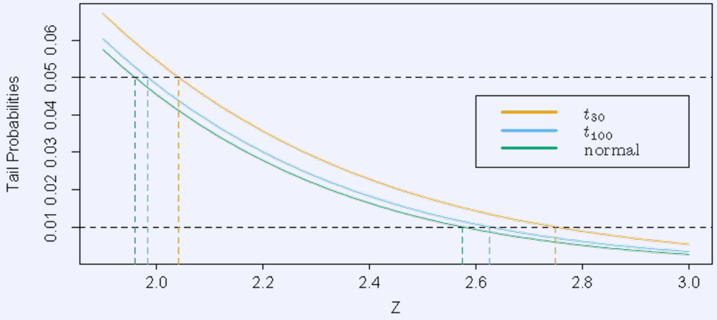
- depends on the accuracy of $\hat{\beta}_1$, i.e. depends on $SE(\hat{\beta}_1)$

The $\boldsymbol{t}$-statistic is the normalized deviation of $\hat{\beta}_1$ from zero

Null-hypothesis

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

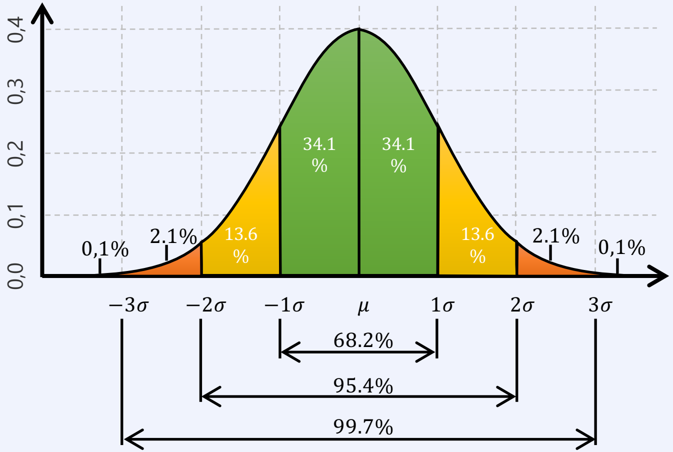
- this also known as the z-score, and it has a bell shape
- for $n > 30$, it is quite similar to the normal distribution



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 $n = 30$

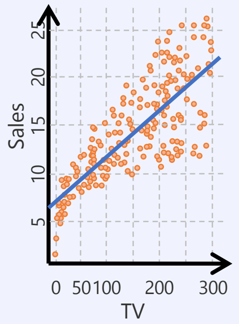
(for $|t| > 2$ the p-value is 4.6%, as 100 – 2 (34.1+13.6) = 4.6%)

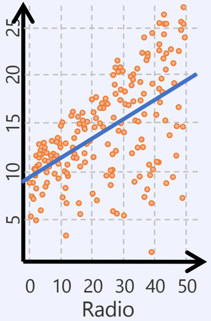
Example Significance of Coefficients

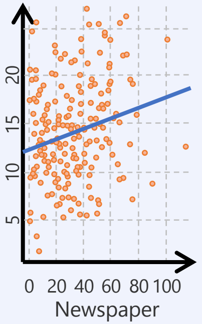
	Coefficient	Std. error	t-statistic	p-value
intercept	7.0325	0.4578	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

	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	<0.0001







Other Scores RSE and R^2

Residual Standard Error (RSE)

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- absolute measure of error measured in units of Y
- RSE estimates the standard error (roughly the average deviation) made by the regression line
- for the advertising data, $RSE = 3.26$, the mean sales is about 14, so the percentage error is 23%

R^2-statistic

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- proportion of variance of Y explained by X
- $R^2 \in [0,1]$ and independent of the scale of Y
- RSS measures variance unaccounted for after regression
- the total sum of squares, or $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, measures the total variance in Y
- $TSS - RSS$ measures variance removed by regressing
- high R^2 means an accurate model

Other Scores Correlation

Correlation

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- the sample estimate of correlation measures how linear the relationship between X and Y is
- in the univariate case, we can show that for the least-squares linear model, $Cor(X,Y)^2 = R^2$
- this does not extend to the multivariate case, nor to models other than least-squares!

Multiple Linear Regression

ISLR 3.2, ESL 3.2.3

Multiple Linear Regression

For linear regression with multiple predictors we assume a model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$
$$= \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon = \boldsymbol{X\beta} + \epsilon$$

- where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ and $\boldsymbol{X} = (1, X_1, \dots, X_p)$ are vectors
- for the advertising example we have $\mathbf{sales} = \beta_0 + \beta_1 \times \mathbf{TV} + \beta_2 \times \mathbf{radio} + \beta_3 \times \mathbf{newspaper} + \epsilon$

For the multivariate case, the residual sum of squares becomes

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \boldsymbol{x}_{ij}^T \beta_j \right)^2 = (Y - \mathbf{X}\beta)^T (Y - \mathbf{X}\beta) \quad *)$$

- which we can again solve by setting the (multidimensional) derivative to zero

*) we slightly misuse notation here, because $\hat{y}$ is actually a function of the β_i. We omit the hats on the β_i since we treat them as variables.

Estimating β for Multiple Linear Regression

To minimize the RSS, we can differentiate w.r.t. β and obtain

$$\frac{\delta RSS}{\delta\beta} = -2\mathbf{X}^T(Y - \mathbf{X}\beta) \qquad \frac{\delta^2 RSS}{\delta\beta\delta\beta^T} = 2\mathbf{X}^T\mathbf{X}$$

- we assume that $\mathbf{X}$ has full column rank, i.e. that $\mathbf{X}^T\mathbf{X}$ is positive definite*
- the RSS then has a unique minimum at which the first derivative vanishes

We set the (multidimensional) derivative to zero

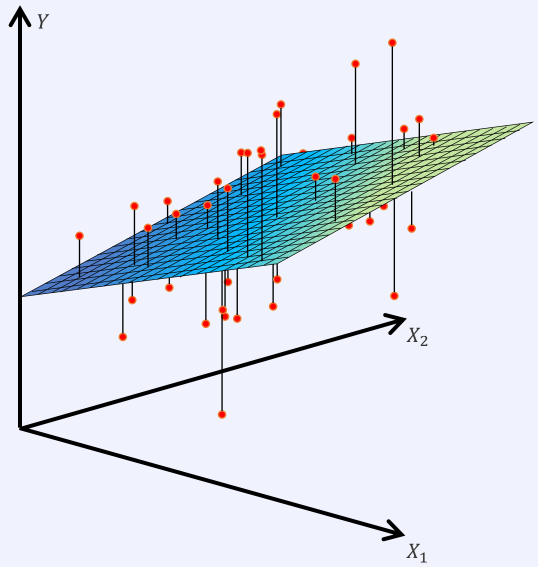
$$2\mathbf{X}^T(Y - \mathbf{X}\beta) = 0$$

- solving for β yields $\hat{\beta}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$
- solving for just one β_i yields $\hat{\beta_i} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- overall, we have $\hat{Y} = \mathbf{X}\widehat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y$

aka the hat matrix, or $\mathbf{H}$

* A matrix $\mathbf{A}$ is positive definite if for all vectors $\boldsymbol{x} \neq \mathbf{0}$ we have $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq \mathbf{0}$

Interpreting Multiple Linear Regression



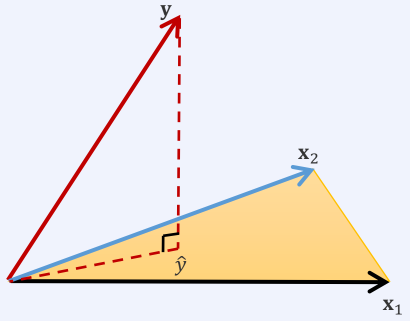
visualization in the space $\mathbb{R}^p$ spanned by the p features

Geometric interpretation 1

- the $\boldsymbol{p}$ features together span a $\boldsymbol{p}$-dimensional space in which $\boldsymbol{n}$ observations live
- the regression plane is the plane that hugs those points best
- best is quantified by minimum

$$RSS(\beta) = ||Y - \mathbf{X}\beta||^2$$

Interpreting Multiple Linear Regression



visualization in the space $\mathbb{R}^n$
spanned by the n observations

Geometric interpretation 2

- $x_0, \dots, x_p$ with $x_0 \equiv 1$ span a p-dimensional subspace of $\mathbb{R}^n$, the column space
- minimizing $RSS(\beta) = ||Y - \mathbf{X}\beta||^2$ implies an orthogonal projection of the $\boldsymbol{y}$-vector onto this subspace
- $\mathbf{H}$ computes this projection, and is hence also called projection matrix

Multiple (Multivariate) Linear Regression

Linear least-squares models are unbiased

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

$$Y = \mathbf{X}\beta + \epsilon$$

- to see this, substitute line 2 into line 1

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon)\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\\ &= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\end{aligned}$$

- and compute expectations

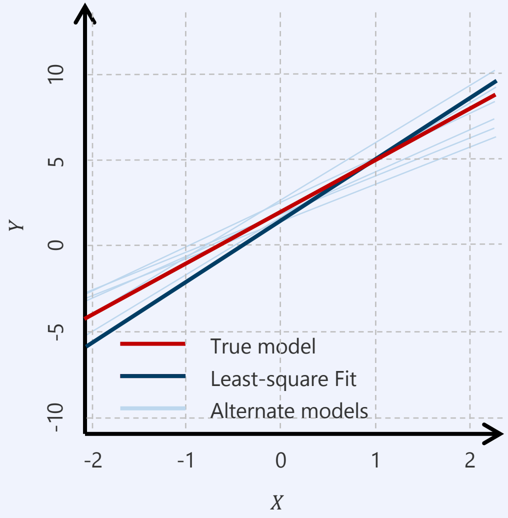
Inputs and errors are independent!

$$\begin{aligned}\mathrm{E}\left[\hat{\beta} \mid \mathbf{X}\right] &= \mathrm{E}[\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon \mid \mathbf{X}]\\ &= \mathrm{E}[\beta \mid \mathbf{X}] + \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon \mid \mathbf{X}]\\ &= \mathrm{E}[\beta \mid \mathbf{X}] + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}[\epsilon] = \beta\end{aligned}$$

$$\mathrm{E}\left[\hat{\beta}\right] = \int \mathrm{E}[\hat{\beta}|\mathbf{X}]\, d\,\Pr(\mathbf{X}) = \int \beta d\,\Pr(\mathbf{X}) = \beta$$

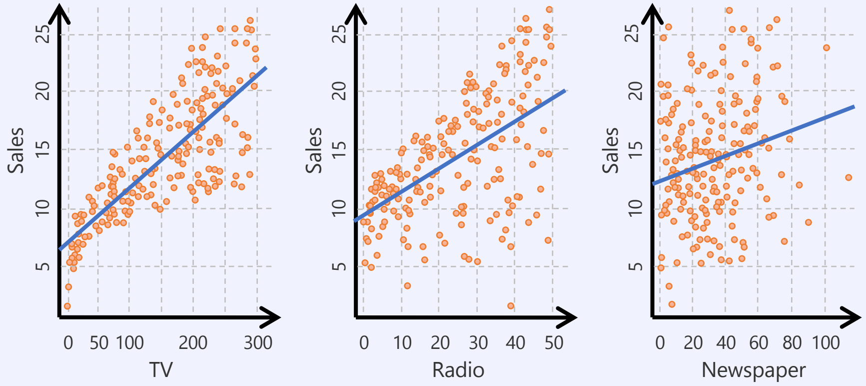
Noise is zero-mean!

Law of total expectation



Among all unbiased linear estimators,
the least-square fit has the smallest variance
(Gauss-Markov Theorem)

Multiple (Multivariate) Linear Regression



Univariate regression

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

	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

Multivariate regression

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

Multiple (Multivariate) Linear Regression

Why is **newspaper** significant in the univariate model, but not in the multivariate one?

- the correlation between **newspaper** and **radio** is 0.35, that is, we spend more on **newspaper** advertising in markets where we also spend more on **radio** advertising
- in the univariate case, we attribute sales to **newspaper** that can also be due to **radio**: **newspaper** is a surrogate for **radio**

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Correlation matrix between inputs

Multiple (Multivariate) Linear Regression

Why is **newspaper** significant in the univariate model, but not in the multivariate one?

- the correlation between **newspaper** and **radio** is 0.35, that is, we spend more on **newspaper** advertising in markets where we also spend more on **radio** advertising
- in the univariate case, we attribute sales to **newspaper** that can also be due to **radio**: **newspaper** is a surrogate for **radio**

Examples of correlations

- number of storks is highly correlated with number of births
- number of gas stations is highly correlated with number of divorces

In these examples, another factor exists that actually causes these features

- if this factor is part of the data we can find it using a multivariate model
- if not, it is a hidden confounder, and we will inferring causally wrong relationships between features