# Machine Learning for Computational Linguistics

## Regression

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

April 26/28, 2016

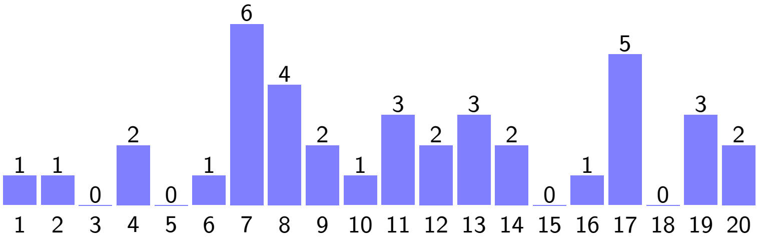
- Course credits:

**9 ECTS** with term paper
**6 ECTS** without term paper

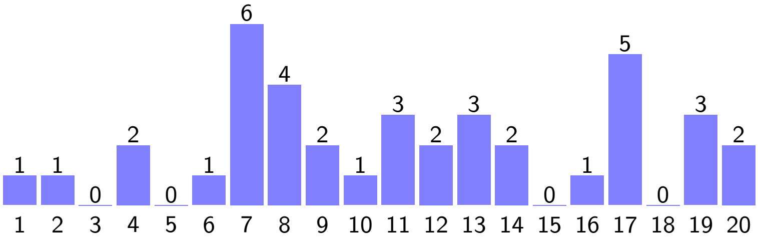
- Homeworks & evaluation:
  For each homework, you either get
  - $0$ not satisfactory or not submitted
  - $[6, 10]$ satisfactory and on time
    - Late homeworks are not accepted

  Please follow the instructions precisely!

# Entropy of your random numbers

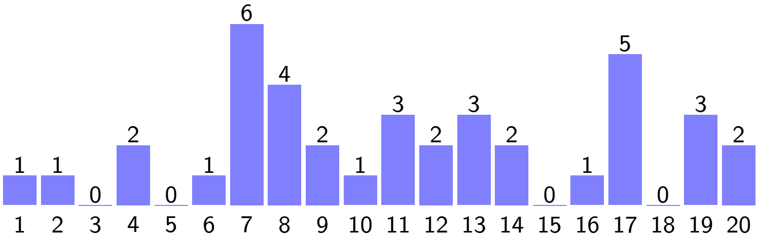


# Entropy of your random numbers




$$H(X) = -\sum_{x} P(x) \log_2 P(x)$$
$$= 2.61$$

# Entropy of your random numbers




$$
\begin{aligned}
H(X) &= -\sum_{x} P(x) \log_2 P(x) \\
&= 2.61
\end{aligned}
$$

If the data was really uniformly distributed: $H(X) = 4.32$.

# Coding a four-letter alphabet

| letter | prob | code 1 | code 2 |
|---|---|---|---|
| a | 1/2 | 0 0 | 0 |
| b | 1/4 | 0 1 | 10 |
| c | 1/8 | 1 0 | 110 |
| d | 1/8 | 1 1 | 111 |

Average code length of a string under code 1:

$$\frac{1}{2}2 + \frac{1}{4}2 + \frac{1}{8}2 + \frac{1}{8}2 = 2.0\text{bits}$$

Average code length of a string under code 2:

$$\frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = 1.75\text{bits} = \mathrm{H}$$

# Statistical inference and estimation

- Statistical inference is about making generalizations that go beyond the data at hand (training set, or experimental sample)
- In a typical scenario, we (implicitly) assume that a particular class of models describe the real-world process, and try to find the best model within the class of models
- In most cases, our models are parametrized: the model is defined by a set of parameters
- The task, then, becomes estimating the parameters from the training set such that the resulting model is useful for unseen instances

# Estimation of model parameters

A typical statistical model can be formulated as

$$y = f(x; w) + \epsilon$$

- $x$ is the input to the model
- $y$ is the quantity or label assigned to for a given input
- $w$ is the parameter(s) of the model
- $f(x; w)$ is the model's estimate of output $y$ given the input $x$, sometimes denoted as $\hat{y}$
- $\epsilon$ represents the uncertainty or noise that we cannot explain or account for

- In machine learning, focus is correct prediction of $y$
- In statistics, the focus is on inference (testing hypotheses or explaining the observed phenomena)

# Estimating parameters: Bayesian approach

Given the training data $\mathbf{X}$, we find the posterior distribution

$$p(\boldsymbol{w}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathbf{X})}$$

- The result, posterior, is a probability distribution of the parameter(s)
- One can get a point estimate of $\boldsymbol{w}$, for example, by calculating the expected value from the distribution
- The posterior distribution also contains the information on the uncertainty of the estimate
- Prior information can be specified by the *prior* distribution

# Estimating parameters: frequentist approach

Given the training data $\mathbf{X}$, we find the value of $\boldsymbol{w}$ that maximizes the likelihood

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \mathrm{p}(\mathbf{X}|\boldsymbol{w})$$

- The likelihood function $\mathrm{p}(\mathbf{X}|\boldsymbol{w})$, often denoted $\mathcal{L}(\boldsymbol{w}|\mathbf{X})$, is the probability of data given $\boldsymbol{w}$ for discrete variables, and the value of probability mass function for the continuous variables
- The problem becomes searching for the maximum value of a function
- Note that we cannot make probabilistic statements about $\boldsymbol{w}$
- Uncertainty of the estimate is less straightforward

# A simple example: estimation of the population mean

We assume that data observed comes from the model:

$$y = \mu + \epsilon$$

where, $\epsilon \sim N(0, \sigma^2)$
An example:

- Let's assume that we are estimating the average number of characters in twitter messages. We will use two data sets:
- $87, 101, 88, 45, 138$
  - The mean of the sample ($\bar{x}$) is $91.8$
  - Variance of the sample ($sd^2$) is $1111.7$ ($sd = 33.34$)
- $87, 101, 88, 45, 138, 66, 79, 78, 140, 102$
  - $\bar{x} = 92.4$
  - $sd^2 = 876.71$ ($sd = 29.61$)

# Estimating mean: Bayesian way

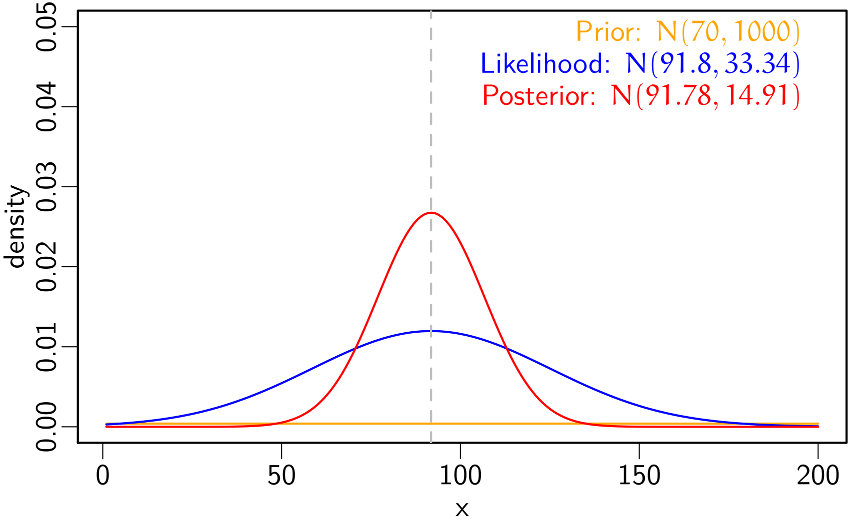
We simply use Bayes' formula:

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{p(D)}$$

- With a vague prior (high variance/entropy), the posterior mean is (almost) the same as the mean of the data
- With a prior with lower variance, posterior is between the prior and the data mean
- Posterior variance indicates the uncertainty of our estimate. With more data, we get a more certain estimate

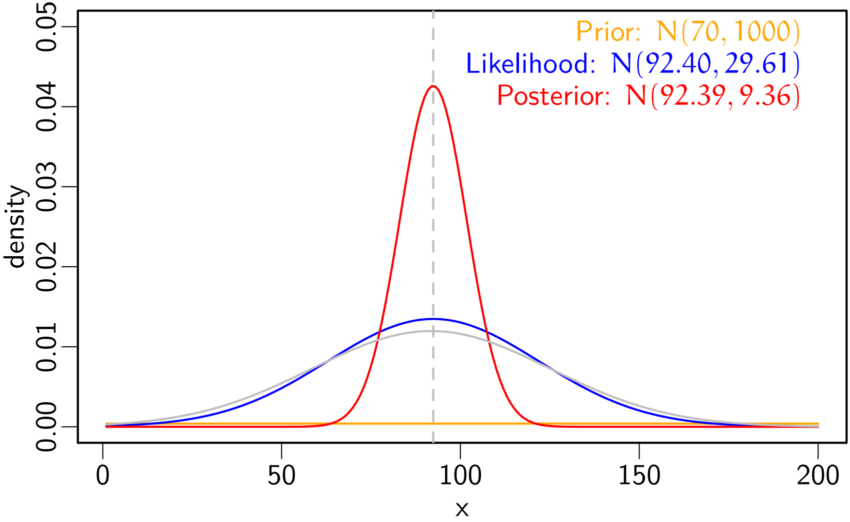
# Estimating mean: Bayesian way

## vague prior, small sample



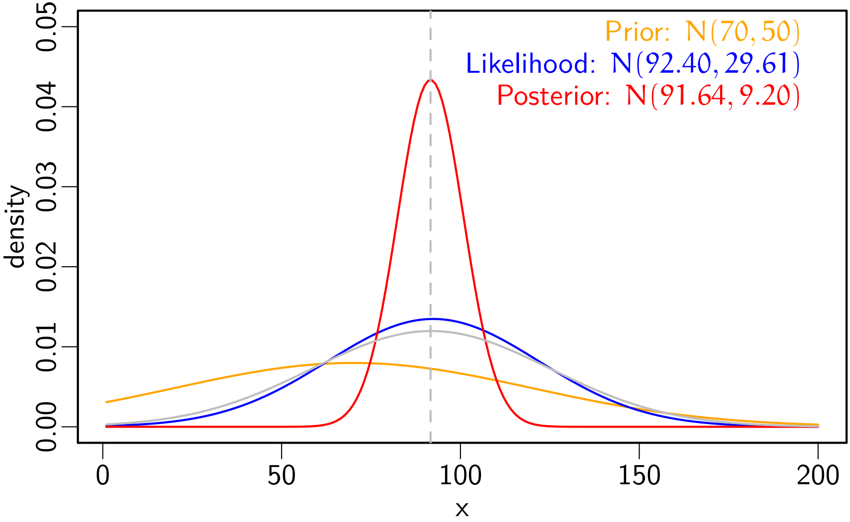
# Estimating mean: Bayesian way

vague prior, larger sample



# Estimating mean: Bayesian way

## visualization



# Estimating mean: frequentist way

- The MLE of the mean of the population is the mean of the sample
  - For 5-tweet sample: $\hat{\mu} = \bar{x} = 91.8$
  - For 10-tweet sample: $\hat{\mu} = \bar{x} = 92.4$
- We express the uncertainty in terms of standard error of the mean ($SE$)

$$SE_{\bar{x}} = \frac{sd_x}{\sqrt{n}}$$

  which corresponds to the means of the (hypothetical) samples of the same size drawn from the same population.
  - For 5-tweet sample: $SE_{\bar{x}} = 33.34/\sqrt{5} = 14.91$
  - For 10-tweet sample: $SE_{\bar{x}} = 29.61/\sqrt{10} = 9.36$
- A rough estimate for a 95% confidence interval is $\bar{x} \pm 2SE_{\bar{x}}$
  - For 5-tweet sample: $91.8 \pm 2 \times 14.91 = [61.98, 121.62]$
  - For 10-tweet sample: $92.4 \pm 2 \times 9.36 = [83.04, 101.76]$

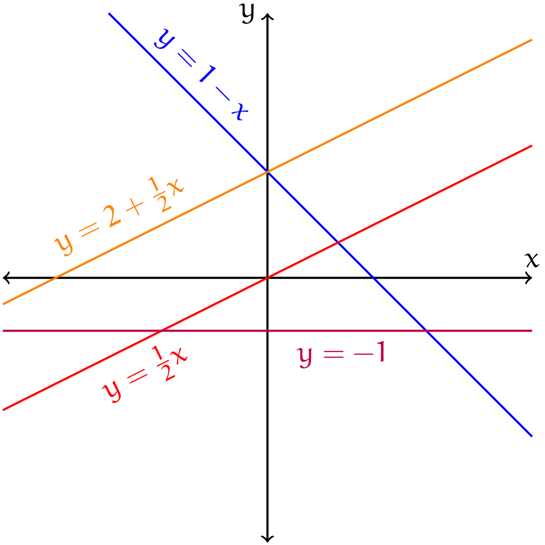
# Regression

- Regression is a supervised method for predicting value of a continuous response variables based on a number of predictors
- We estimate the conditional expectation of the outcome variable given the predictor(s)
- If the outcome is a label, the problem is called classification. But the border between the two often is not that clear

# The linear equation: a reminder

$$y = a + bx$$

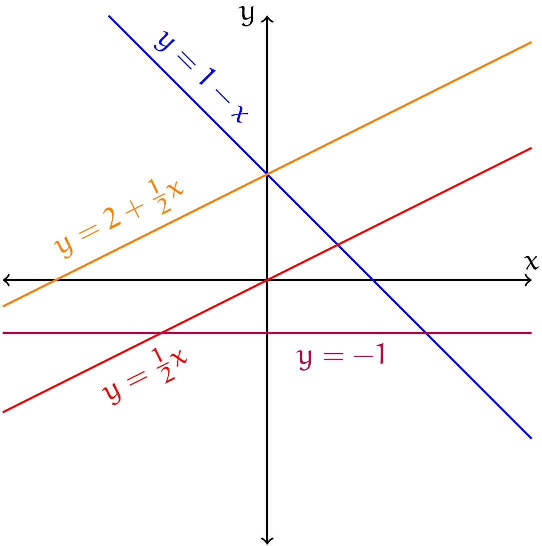
- a (intercept) is where the line crosses the $y$ axis.
- b (slope) is the change in $y$ as $x$ is increased one unit.



# The linear equation: a reminder

$$y = a + bx$$

- a (intercept) is where the line crosses the $y$ axis.
- b (slope) is the change in $y$ as $x$ is increased one unit.




What is the correlation between $x$ and $y$ for each line (relation)?

# The simple linear model

$$y_i = a + bx_i + \epsilon_i$$

- $y$ is the *outcome* (or response, or dependent) variable. The index $i$ represents each unit observation/measurement (sometimes called a 'case').
- $x$ is the *predictor* (or explanatory, or independent) variable.
- $a$ is the *intercept*.
- $b$ is the *slope* of the regression line.
- $a$ and $b$ are called *coefficients* or *parameters*.
- $a + bx$ is the *deterministic* part of the model. It is the model's prediction of $y$ ($\hat{y}$), given $x$.
- $\epsilon$ is the residual, error, or the variation that is not accounted for by the model. Assumed to be normally distributed with 0 mean

# Notation differences for the regression equation

$$y_i = a + bx_i + \epsilon_i$$

# Notation differences for the regression equation

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- Sometimes, Greek letters $\alpha$ and $\beta$ are used for intercept and the slope, respectively.

# Notation differences for the regression equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Sometimes, Greek letters $\alpha$ and $\beta$ are used for intercept and the slope, respectively.
- Another common notation to use only $b$, $\beta$ $\theta$, but use subscripts, $0$ indicating the intercept and $1$ indicating the slope.

# Notation differences for the regression equation

$$y_i = w_0 + w_1 x_i + \epsilon_i$$

- Sometimes, Greek letters $\alpha$ and $\beta$ are used for intercept and the slope, respectively.
- Another common notation to use only $b$, $\beta$ $\theta$, but use subscripts, $0$ indicating the intercept and $1$ indicating the slope.
- In machine learning it is common to use $w$ for all coefficients (sometimes you may see $b$ used instead of $w_0$)

# Notation differences for the regression equation

$$y_i = \hat{w}_0 + \hat{w}_1 x_i + \epsilon_i$$

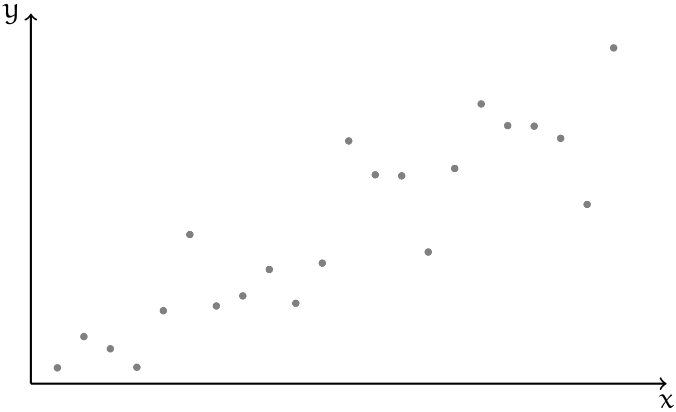
- Sometimes, Greek letters $\alpha$ and $\beta$ are used for intercept and the slope, respectively.
- Another common notation to use only $b$, $\beta$ $\theta$, but use subscripts, $0$ indicating the intercept and $1$ indicating the slope.
- In machine learning it is common to use $w$ for all coefficients (sometimes you may see $b$ used instead of $w_0$)
- Sometimes coefficients wear hats, to emphasize that they are estimates.

# Notation differences for the regression equation

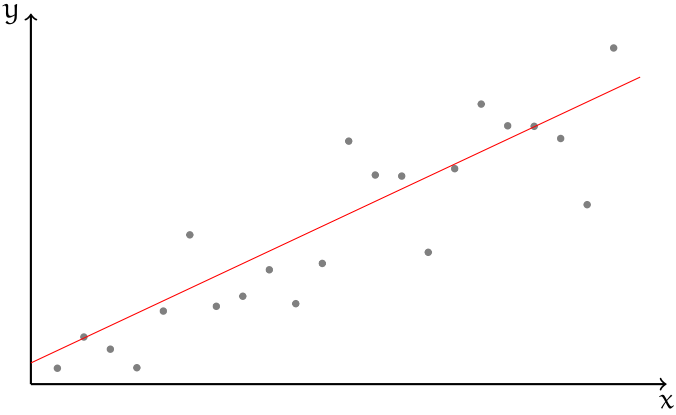
$$y_i = \boldsymbol{w}\boldsymbol{x_i} + \epsilon_i$$

- Sometimes, Greek letters $\alpha$ and $\beta$ are used for intercept and the slope, respectively.
- Another common notation to use only $b$, $\beta$ $\theta$, but use subscripts, $0$ indicating the intercept and $1$ indicating the slope.
- In machine learning it is common to use $w$ for all coefficients (sometimes you may see $b$ used instead of $w_0$)
- Sometimes coefficients wear hats, to emphasize that they are estimates.
- Often, we use the vector notation for both input(s) and coefficients: $\boldsymbol{w} = (w_0, w_1)$ and $\boldsymbol{x_i} = (1, x_i)$

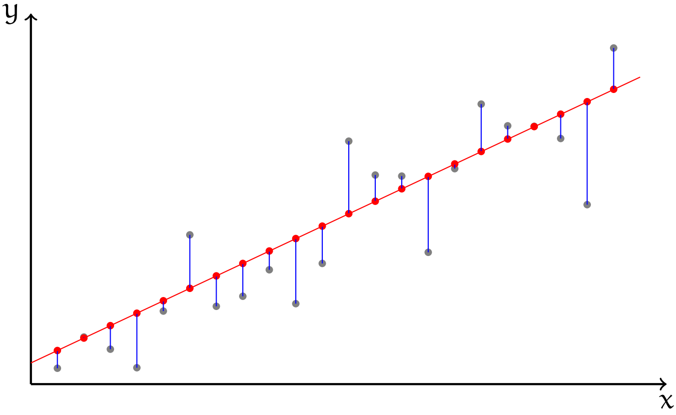
# Visualization of regression procedure



# Visualization of regression procedure



# Visualization of regression procedure



# Least-squares regression

Least-squares regression is the method of determining regression coefficients that minimizes the sum of squared residuals ($SS_R$).

$$y_i = \underbrace{w_0 + w_1 x_i}_{\hat{y}_i} + \epsilon_i$$

# Least-squares regression

Least-squares regression is the method of determining regression coefficients that minimizes the sum of squared residuals ($SS_R$).

$$y_i = \underbrace{w_0 + w_1 x_i}_{\hat{y}_i} + \epsilon_i$$

- We try to find $w_0$ and $w_2$, that minimize the prediction error:

$$\sum_i \epsilon_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - (w_0 + w_1 x_i)\right)^2$$

- This minimization problem can be solved analytically, yielding:

$$\begin{aligned} w_1 &= r\frac{sd_y}{sd_x} \\ w_0 &= \bar{y} - w_1\bar{x} \end{aligned}$$

* See appendix for the derivation.

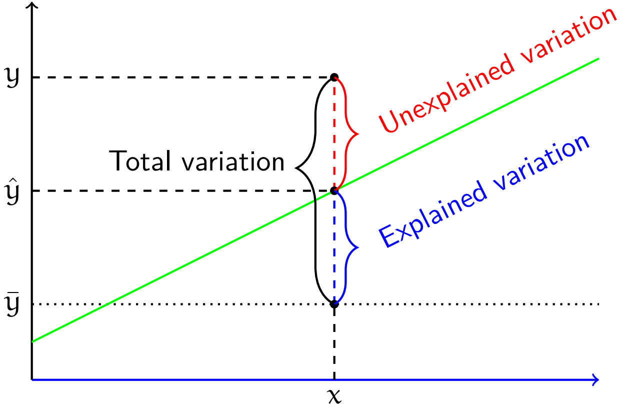
# Short digression: minimizing functions

In least squares regression, we want to find $w_0$ and $w_1$ values that minimize the quantity

$$\sum_i (y_i - (w_0 + w_1 x_i))^2$$

- Note that the above is a quadratic function of $w_0$ and $w_1$
- This is important, since quadratic functions are convex and have a single extreme value: we have a unique solution for our minimization problem
- In case of least squares regression, we are even luckier: we can find an analytic solution
- Even if we do not have an analytic solution, if our error function is convex, a search procedure like gradient descent can find the global minimum

# Explained variation




| | | | | |
|---|---|---|---|---|
| Total variation | = | Unexplained variation | + | Explained variation |
| $y - \bar{y}$ | = | $y - \hat{y}$ | + | $\hat{y} - \bar{y}$ |

# Assessing the model fit: $r^2$

We can express the variation explained by a regression model as:

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2}$$

It can be shown that this value is the square of the correlation coefficient, $r^2$, also called the coefficient of determination.

- $100 \times r^2$ can be interpreted as 'the percentage of variance explained by the model'.
- $r^2$ shows how well the model fits to the data: closer the data points to the regression line, higher the value of $r^2$.

# Regression and inference: an example

## (1) The data

We want to see the effect of mother's IQ to four-year-old children's cognitive test scores (Fake data, based on analysis presented in Gelman&Hill 2007).

| Case | Kid's Score | Mother's IQ |
|---:|---:|---:|
| 1 | 109 | 91 |
| 2 | 99 | 102 |
| 3 | 96 | 88 |
| ... | | |
| 43 | 108 | 101 |
| 44 | 110 | 78 |
| 45 | 97 | 67 |

# Regression and inference: an example

## (2) Analysis (R output)

```
lm(formula = kid.score ~ mother.iq)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5174   24.2375  0.145   0.885
mother.iq   0.6023    0.2471  2.437    0.019 *
---
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214, Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$w_1 = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

# Regression and inference: an example

## (2) Analysis (R output)

```
lm(formula = kid.score ~ mother.iq)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5174   24.2375  0.145   0.885
mother.iq   0.6023    0.2471  2.437    0.019 *
---
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214, Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$w_1 = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

$r^2 = 0.12$ Mothers' IQ explains 12% of the variation in test scores.

# Regression and inference: an example

## (2) Analysis (R output)

```
lm(formula = kid.score ~ mother.iq)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.5174   24.2375  0.145   0.885
mother.iq   0.6023    0.2471  2.437    0.019 *
---
Residual standard error: 22.59 on 43 degrees of freedom
Multiple R-squared: 0.1214, Adjusted R-squared: 0.101
F-statistic: 5.941 on 1 and 43 DF, p-value: 0.019
```

$w_1 = 0.6$ Expected score difference between two children whose mother's IQ differs one unit.

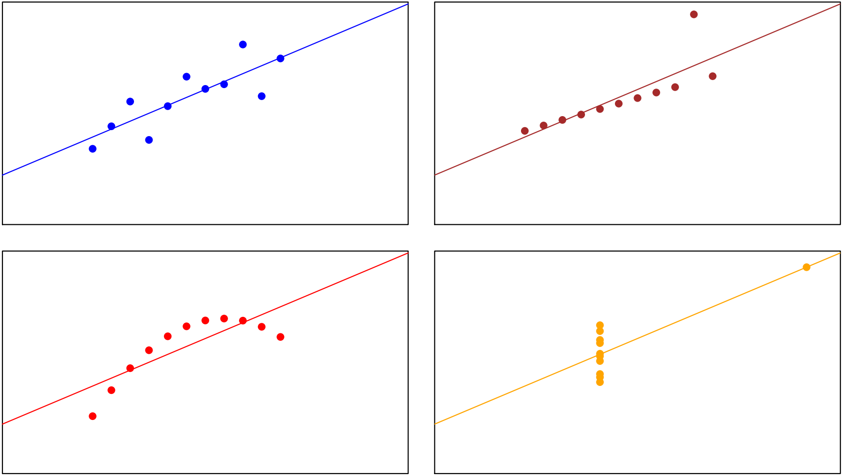
$r^2 = 0.12$ Mothers' IQ explains 12% of the variation in test scores.

$p = 0.02$ Given the sample size, probability of finding a $w_1$ value that far from 0 (two-tailed t-test with null hypothesis $w_1 = 0$).

# Notes/issues on ordinary least squares regression

- Response variable should be linearly related to predictor(s)
- Least squares estimation is sensitive to outliers
- The residuals should be normally distributed

# You should always check your data



* This data set is known as Anscombe's quartet (Anscombe, 1973).

All four sets have the same mean, variance and fitted regression line.

# Regression with multiple predictors

$$y_i = \underbrace{w_0 + w_1 x_{i,1} + w_2 x_{2,i} + \ldots + w_k x_{k,i}}_{\hat{y}} + \epsilon_i = wx_i + \epsilon_i$$

$w_0$ is the intercept (as before).

$w_{1..k}$ are the coefficients of the respective predictors.

$\epsilon$ is the error term (residual).

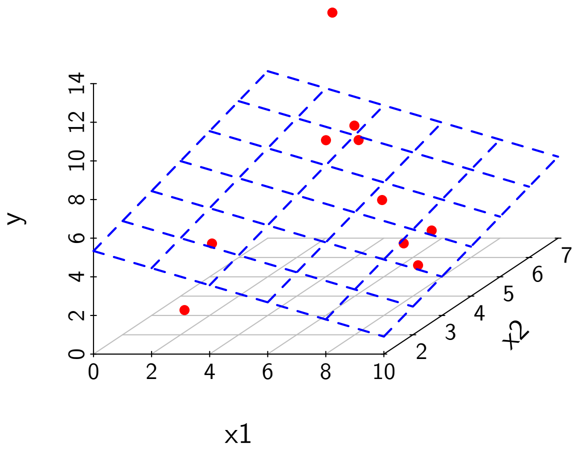
- using vector notation the equation becomes:

$$y_i = \boldsymbol{w}\boldsymbol{x}_i + \epsilon_i$$

where $\boldsymbol{w} = (w_0, w_1, \ldots, w_k)$ and $\boldsymbol{x}_i = \big(1, x_{i,1}, \ldots, x_{i,k}\big)$

It is a generalization of simple regression with some additional power and complexity.

# Visualizing regression with two predictors



# Input/output of liner regression: some notation

A regression with $k$ input variables and $n$ instances can be described as:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{bmatrix}}_{\mathbf{X}} \times \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix}}_{\boldsymbol{w}} + \underbrace{\begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\epsilon}}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

# Estimation in multiple regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

We want to minimize the error (as a function of $\boldsymbol{w}$):

$$\begin{aligned}\boldsymbol{\epsilon}^2 = \mathrm{J}(\boldsymbol{w}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{w})^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{w}\|^2\end{aligned}$$

Our least-squres estimate is:

$$\begin{aligned}\hat{\boldsymbol{w}} &= \operatorname*{arg\,min}_{\boldsymbol{w}} \mathrm{J}(\boldsymbol{w}) \\ &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\end{aligned}$$

Note: the least squares estimate is also the maximum likelihood estimate under the assumption of normal distribution of errors.

# Issues in multiple regression estimation

- Overfitting: many variables cause model to learn noise in the data (we will return to this issue)
- Collinearity: high correlation between predictors increase uncertainty of coefficient estimates
- Model/feature selection is typically needed for both prediction and inference

# Categorical predictors

- Categorical predictors are represented as multiple binary coded input variables
- For a binary predictor, we use a single binary input. For example, (1 for one of the values, and 0 for the other)

$$x = \begin{cases} 0 & \text{for male} \\ 1 & \text{for female} \end{cases}$$

- For a categorical predictor with $k$ values, we use $k - 1$ predictors (various coding schemes are possible). For example, for 3-values

$$\mathbf{x} = \begin{cases} (0, 0) & \text{for neutral} \\ (0, 1) & \text{for negative} \\ (1, 0) & \text{for positive} \end{cases}$$

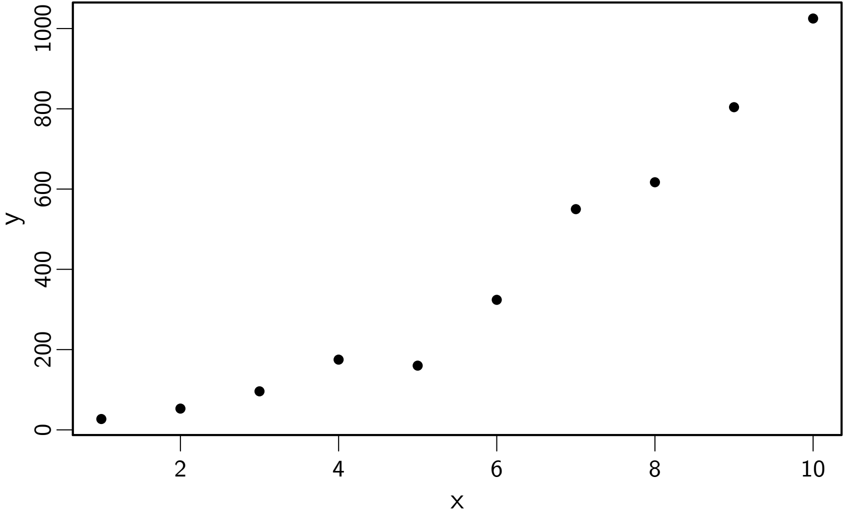
# Dealing with non-linearity (to some extent)

- Least squares works, because the loss function is linear with respect to parameter $\boldsymbol{w}$
- Introducing non-linear combinations of inputs does not affect the estimation procedure. The following are still linear models

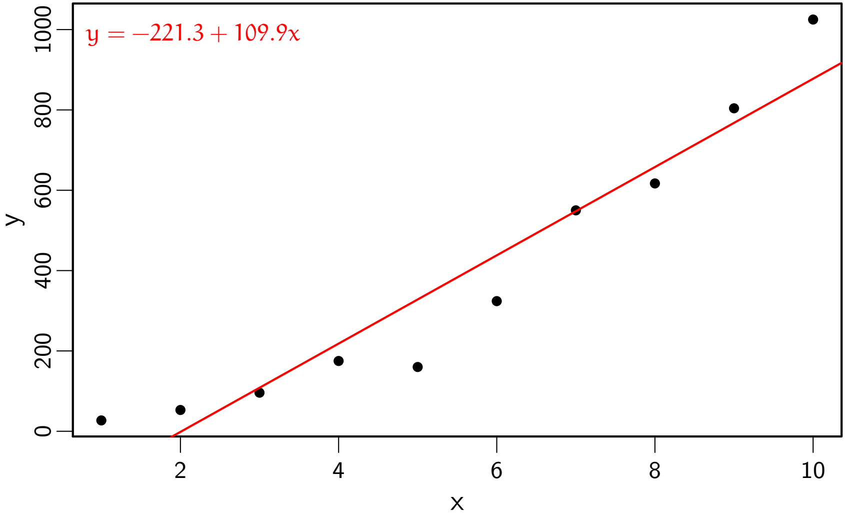
$$
\begin{aligned}
y_i &= w_0 + w_1 x_i^2 + \epsilon_i \\
y_i &= w_0 + w_1 \log(x_i) + \epsilon_i \\
y_i &= w_0 + w_1 x_{i,1} + w_2 x_{i,2} + w_3 x_{i,1} x_{i,2} + \epsilon_i
\end{aligned}
$$

- These *transformations* allow linear models to deal with some non-linearities
- In general, we can replace input $x$ by a function of the input(s) $\Phi(x)$. $\Phi()$ is called a *basis function*

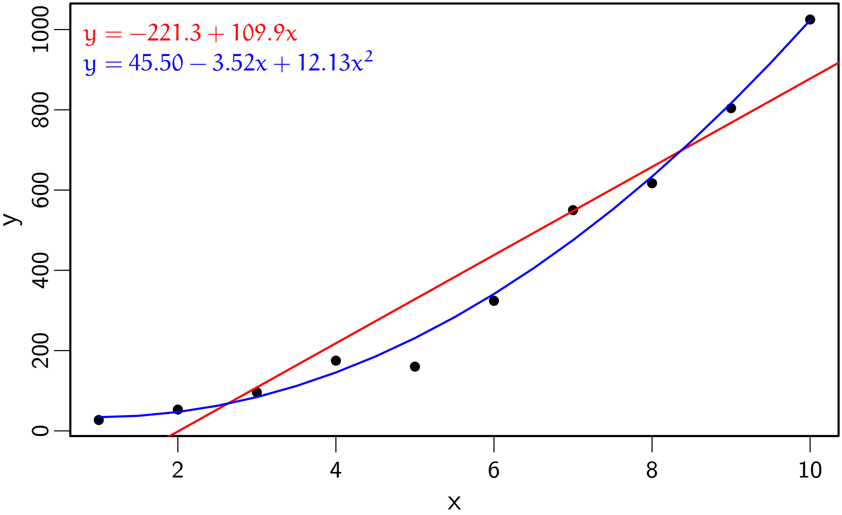
# Example: polynomial basis functions



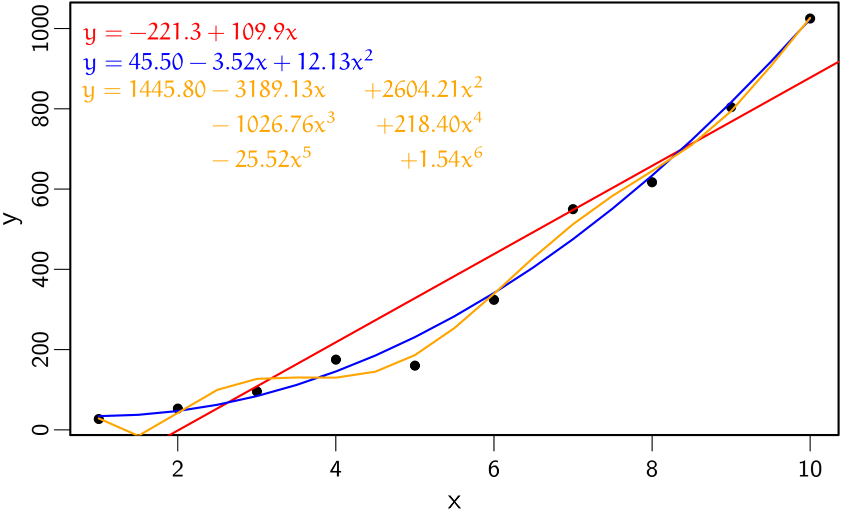
# Example: polynomial basis functions



# Example: polynomial basis functions



# Example: polynomial basis functions



# Next...

| | |
|---|---|
| Tuesday | hands-on exercises with regression |
| Next week | classification |

# Estimating the regression line

We express the sum of squared residuals as a function of the (unknown) regression line:

$$
\begin{aligned}
\sum_{i=1}^{n} \epsilon_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \\
&= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= \sum_{i=1}^{n} (a^2 + 2abx_i - 2ay_i + b^2x_i^2 - 2bx_iy_i + y_i^2)
\end{aligned}
$$

Thus, $\sum_{i=1}^{n} \epsilon_i^2$ is function $f$ in $x$, $y$ with unknown parameters $a$, $b$.

# Estimating the regression line

For a fixed sample $\mathcal{S} = (x, y)$, we want to minimize $f_{ab}(x, y)$ with

$$f_{ab}(x, y) = \sum_{i=1}^{n} (a^2 + 2abx_i - 2ay_i + b^2x_i^2 - 2bx_iy_i + y_i^2)$$

To minimize this function, find $a$ and $b$ such that $f'_{ab}(x, y) = 0$.

Treat $a$ and $b$ as variables and find partial derivatives $\frac{\partial}{\partial a}f$, $\frac{\partial}{\partial b}f$

$$\begin{aligned}
\frac{\partial}{\partial a}f = f'_{xyb}(a) &= \sum_{i=1}^{n} (2a + 2bx_i - 2y_i) \\
\frac{\partial}{\partial b}f = f'_{xya}(b) &= \sum_{i=1}^{n} (2ax_i + 2bx_i^2 - 2x_iy_i)
\end{aligned}$$

# Relationship between correlation and regression

Recall we obtained two partial derivatives (when minimizing sum of squared residuals):

$$f'_{xyb}(a) = \sum_{i=1}^{n}(2a + 2bx_i - 2y_i) \quad (1)$$

$$f'_{xya}(b) = \sum_{i=1}^{n}(2ax_i + 2bx_i^2 - 2x_i y_i) \quad (2)$$

Set (1) to zero:

$$\begin{aligned}
& f'_{xyb}(a) = 0 \\
\Leftrightarrow \quad & n \cdot 2a + \sum_{i=1}^{n}(2bx_i - 2y_i) = 0 \\
\Leftrightarrow \quad & n \cdot 2a + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i = 0 \\
\Leftrightarrow \quad & n \cdot a = n \cdot \overline{y} - n \cdot b\overline{x} \\
\Leftrightarrow \quad & a = \overline{y} - b\overline{x}
\end{aligned}$$

# Relationship between correlation and regression

Plug $a = \overline{y} - b\overline{x}$ into (2) and set to zero:

$$
\begin{aligned}
& f'_{xya}(b) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} (2(\overline{y} - b\overline{x})x_i + 2bx_i^2 - 2x_i y_i) = 0 \\
\Leftrightarrow \quad & (\overline{y} - b\overline{x})(n\overline{x}) + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \\
\Leftrightarrow \quad & n\overline{xy} - b\overline{x}^2 n + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \\
\Leftrightarrow \quad & b(\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n) = \sum_{i=1}^{n} x_i y_i - n\overline{xy} \\
\Leftrightarrow \quad & b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n}
\end{aligned}
$$

# Relationship between correlation and regression

$$
\begin{aligned}
b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n} \quad \Leftrightarrow \quad & b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\
\Leftrightarrow \quad & b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \\
\Leftrightarrow \quad & b = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2\right)} \\
\Leftrightarrow \quad & b = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x^2} \\
\Leftrightarrow \quad & b = \left( \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma_x}\right) \left(\frac{y_i - \overline{y}}{\sigma_y}\right) \right) \cdot \frac{\sigma_y}{\sigma_x} \\
\Leftrightarrow \quad & b = r \frac{\sigma_y}{\sigma_x}
\end{aligned}
$$

# Another relation between correlation and regression

$$
\begin{aligned}
\frac{\text{explained variance}}{\text{total variance}} &= \frac{\sum_{i=1}^{n}((a + bx_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{\sum_{i=1}^{n}((\overline{y} - b\overline{x} + bx_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= \frac{\sum_{i=1}^{n} b^2(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= b^2 \cdot \left(\frac{\sigma_x}{\sigma_y}\right)^2 \\
&= r^2 \left(\frac{\sigma_y}{\sigma_x}\right)^2 \cdot \left(\frac{\sigma_x}{\sigma_y}\right)^2 \\
&= r^2
\end{aligned}
$$

# Standard error for the regression slope and intercept

$$SE_b = \frac{sd_r}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$SE_a = sd_r \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$