Applied Statistics

for Computer Science BSc, Exam

Probability Theory and Mathematical Statistics

for Computer Science Engineering BSc, Term grade

István Fazekas

University of Debrecen

2020/21 fall

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus

Computer tool: Matlab

Book:

Yates, Goodman:

Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

Lecture 13

Regression

Fitting functions in science

Galilei's law of free fall
https://en.wikipedia.org/wiki/Free_fall

$$x = \frac{g}{2}t^2,$$

where g is acceleration due to gravity ($g = 9.81 m/s^2$ near the surface of the earth),
t is the time elapsed,
x is the distance
The result was that the distance is a quadratic function of the time.

Fitting functions in science

Newton's law of universal gravitation
https://en.wikipedia.org/wiki/Newton's_law_of_universal_gravitation
Every point mass attracts every single other point mass by a force acting along the line intersecting both points. The force is proportional to the product of the two masses and inversely proportional to the square of the distance between them:

$$F = G\frac{m_1 m_2}{r^2}$$

G is the gravitational constant ($G = 6.674 \times 10^{-11} m^3 kg^{-1} s^{-2}$).
The task was to find the proper function for the force.

Linear regression

Example 1.
We measured the heights and the weighs of 10 students.
The results are the following

$$x_1, \ldots, x_{10} = 160,\ 166,\ 170,\ 175,\ 178,\ 179,\ 180,\ 185,\ 190,\ 198,$$

$$y_1, \ldots, y_{10} = 55,\ 70,\ 65,\ 80,\ 75,\ 85,\ 80,\ 92,\ 85,\ 97$$

x_i is the height and y_i is the weight of the ith student.
Find the linear approximation of the weight by the height.
Find a general method of the approximation.

Linear regression...

Example 1...

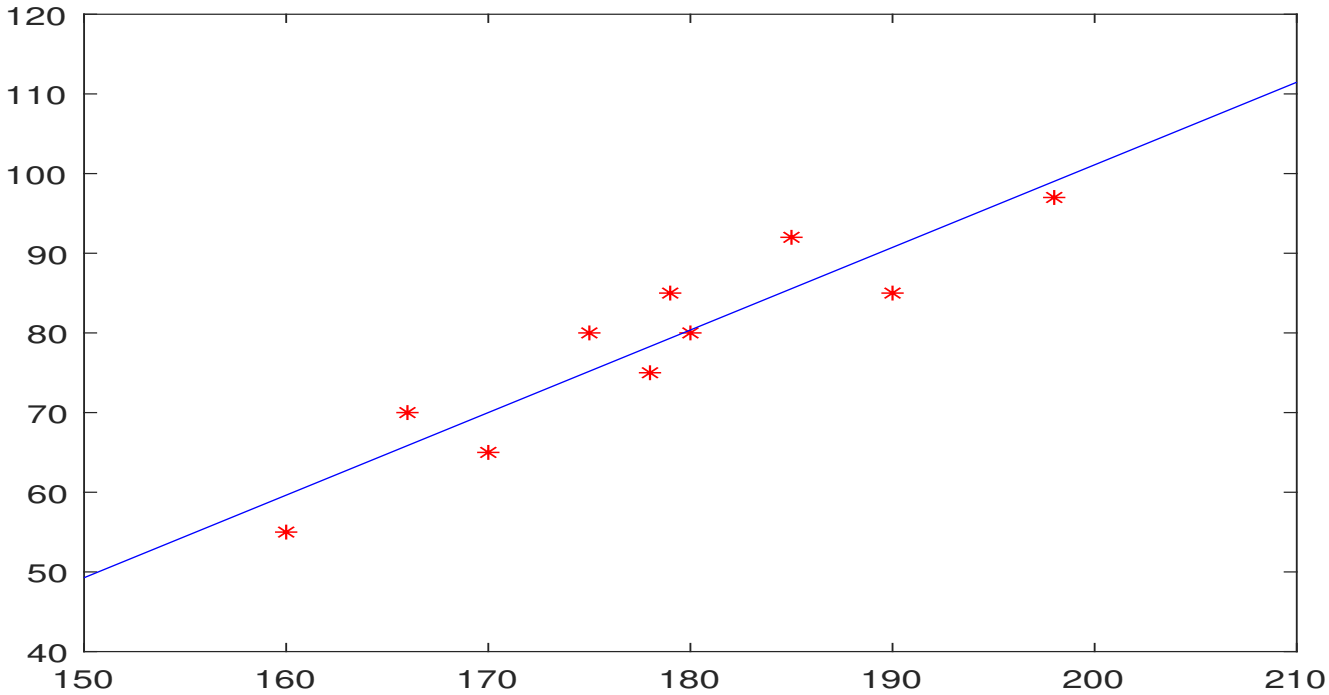

Figure: The approximation of the weight (y) by the height (x) is $y \approx 1.0366x - 106.2210$

Linear regression in probability theory

Approximate the r.v. Y by a function of another r.v. X.
We say that $f_0(X)$ the **best approximation (prediction)** of Y in a class of functions Λ, if $f_0 \in \Lambda$ and

$$\mathbb{E}(Y - f_0(X))^2 = \min_{f \in \Lambda} \mathbb{E}(Y - f(X))^2. \tag{1}$$

If Λ is the class of linear functions, then we arrive at the notion of **best linear approximation (prediction)**.
The principle in (1) is the **method of least squares**.

Linear regression in probability theory...

Theorem 1. Let $0 < \mathrm{Var}X, \mathrm{Var}Y < \infty$.
Then the best linear prediction of Y based on X is

$$\hat{Y} = \mathrm{corr}(X, Y)\frac{\sqrt{\mathrm{Var}Y}}{\sqrt{\mathrm{Var}X}} \cdot X + \mathbb{E}Y - \mathrm{corr}(X, Y)\frac{\sqrt{\mathrm{Var}Y}}{\sqrt{\mathrm{Var}X}} \cdot \mathbb{E}X. \quad (2)$$

Linear regression in probability theory...

Proof. We have to find the minimum of the function

$$g(a,b) = \mathbb{E}[Y - (aX + b)]^2 =$$

$$= \mathbb{E}Y^2 - 2a\mathbb{E}(XY) - 2b\mathbb{E}Y + a^2\mathbb{E}X^2 + 2ab\mathbb{E}X + b^2.$$

The partial derivatives are

$$\frac{\partial g(a,b)}{\partial a} = -2\mathbb{E}(XY) + 2a\mathbb{E}X^2 + 2b\mathbb{E},$$

$$\frac{\partial g(a,b)}{\partial b} = -2\mathbb{E}Y + 2a\mathbb{E}X + 2b.$$

Linear regression in probability theory...

Proof... The above partial derivatives are zeros if

$$a = \mathrm{cov}(X, Y)/\mathrm{Var}X,$$

$$b = -a\mathbb{E}X + \mathbb{E}Y$$

From these we obtain (2). As the matrix of the second derivatives is

$$\begin{pmatrix} 2\mathbb{E}X^2 & 2\mathbb{E}X \\ 2\mathbb{E}X & 2 \end{pmatrix},$$

so g has minimum.

Remark. If $\mathrm{corr}(X, Y) = \pm 1$, then (2) gives Y itself.

Linear regression in probability theory...

Exercise. Roll two fair dice.

Let X_1 denote the number shown by the first die, and X_2 the number shown by the second one.

Let $X = X_1$ and $Y = X_1 + X_2$.

Find the linear regression of Y on X.

Now $\mathbb{E}X = 3.5$, $\mathbb{E}Y = 7$.

Let $\mathrm{Var}(X) = \sigma^2$.

Then $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 2\sigma^2$ and

$$\mathrm{cov}(X, Y) = \mathrm{cov}(X_1, X_1) + \mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_1, X_1) = \mathrm{Var}(X_1) = \sigma^2$$

because $\mathrm{cov}(X_1, X_2) = 0$ by independence. So

$$a = \frac{\mathrm{cov}(X, Y)}{\mathrm{Var}(X)} = 1, \quad b = -a\mathbb{E}X + \mathbb{E}Y = 3.5$$

Therefore

$$Y \approx X + 3.5$$

Linear regression in probability theory...

Exercise. Roll two fair dice.
Let X_1 denote the number shown by the first die, and X_2 the number shown by the second one.
Let $X = \max\{X_1, X_2\}$ and $Y = \min\{X_1, X_2\}$.
Find the linear regression of Y on X.

Linear regression in probability theory...

Exercise.
Let the distribution of (X, Y) be two-dimensional normal with expectation vector and covariance matrix

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

Show that the best linear approximation is

$$Y \approx 0.5X + 1.5$$

Linear regression in statistics

Let

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

be a two-dimensional sample.
Find a line

$$y = ax + b$$

so that

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 \tag{3}$$

is minimal. It is again the method of least squares.

Linear regression in statistics...

Let

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

be the empirical means and the empirical variances,

$$m_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \quad c = \frac{m_{xy}}{s_x \cdot s_y}$$

be the empirical covariance and the empirical correlation coefficient, finally let

$$\widehat{R} = c \cdot \frac{s_y}{s_x}$$

be the regression coefficient.

Linear regression in statistics...

Theorem 2. The least squares estimators $\widehat{a}$, $\widehat{b}$ of a, b are

$$\widehat{a} = \frac{m_{xy}}{s_x^2} = c \cdot \frac{s_y}{s_x} = \widehat{R},$$

$$\widehat{b} = \overline{y} - \overline{x} \cdot \frac{m_{xy}}{s_x^2} = \overline{y} - \overline{x}\widehat{R}.$$

Therefore the linear regression is

$$y = c \cdot \frac{s_y}{s_x} x + \overline{y} - \overline{x} \cdot c \cdot \frac{s_y}{s_x} = \widehat{R}(x - \overline{x}) + \overline{y}.$$

This equation in standardized form is

$$\frac{y - \overline{y}}{s_y} = c \cdot \frac{x - \overline{x}}{s_x}.$$

Linear regression in statistics...

Proof of Theorem 2. Proof 1. Let

$$g(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

We should minimize this function. We calculate partial derivatives and find the minimum like in the proof of Theorem 1.

Linear regression in statistics...

Proof of Theorem 2...

Proof 2. We can apply Theorem 1 to the two-dimensional random variable (X, Y) with distribution

$$P(X = x_i, Y = y_i) = 1/n, \qquad i = 1, 2, \ldots, n.$$

We should find the minimum of $\mathbb{E}[Y - (aX + b)]^2$. But now

$$\mathbb{E}[Y - (aX + b)]^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$

so the task is the same as in (3). Applying the result of Theorem 1 we can see that now

$$\mathbb{E}X = \overline{x}, \ \mathbb{E}Y = \overline{y}, \ \mathrm{Var}(X) = s_x^2, \ \mathrm{Var}(Y) = s_y^2,$$

and $\mathrm{corr}(X, Y)$ is the empirical correlation coefficient of the two samples. So Theorem 1 gives Theorem 2.

Linear regression in statistics...

Proof of Theorem 2...

Proof 3.

Apply orthogonal projection.

Proof 4.

Apply the forthcoming Theorem 3.

Linear regression in statistics...

Example 1...

The Matlab solution is the following

```
x=[160 166 170 175 178 179 180 185 190 198];
y=[ 55 70 65 80 75 85 80 92 85 97];
mx=mean(x); my=mean(y);
vx=mean(x.^2)-mx^2; vy=mean(x.^2)-my^2;
mxy=mean(x.*y)-mx*my;
a=mxy/vx; b=my-mx*a;
figure
plot(x,y,'r*'); hold on
t=150:0.01:210; plot(t,a*t+b,'b-');
```

The general linear model

$$\boxed{\mathbf{Y} = X\beta + \varepsilon} \tag{4}$$

is the linear model, where
$\mathbf{Y}$ is the n-dimensional observation vector (response variable),
X is an $n \times p$ known non-random matrix, the matrix of explanatory variables,
β is a p-dimensional vector of unknown parameters,
ε is a not observable n-dimensional random vector (random error).
Usually $n \gg p$, because p is the number of explanatory variables, n is the number of observed objects.

The general linear model. The OLS method...

If $\mathbb{E}\varepsilon = \mathbf{0}$ and $\mathrm{var}\varepsilon = \sigma^2 I$ (σ^2 is an unknown parameter), then the model is **homoscedastic**.
Then we apply the OLS=Ordinary Least Squares method to find the estimator of β.
The estimator will be denoted by $\widehat{\beta}$.
So let $\widehat{\beta}$ be the minimizing point of

$$\|\mathbf{Y} - X\beta\|^2$$

(Here $\|\,.\,\|$ is the norm in $\mathbb{R}^n$.)

Theorem. $\widehat{\beta}$ is the OLS estimator $\iff$ $\widehat{\beta}$ is the solution of the **normal equation**

$$\boxed{X^\top \mathbf{Y} = X^\top X\beta}$$

The linear model. The OLS method...

Proof.

When will be $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ minimal?

If $\mathbf{Y} - X\boldsymbol{\beta}$ is the orthogonal complement of $\mathbf{Y}$ to the subspace generated by the column vectors of X.

That is $\mathbf{Y} - X\boldsymbol{\beta}$ is orthogonal to each column of X, that is

$$X^\top \mathbf{Y} - X^\top X\boldsymbol{\beta} = \mathbf{0},$$

so

$$X^\top X\widehat{\boldsymbol{\beta}} = X^\top \mathbf{Y}.$$

Remark. $X^\top X$ is invertible $\iff$ $\text{rank}X = p$.

Remark. If $\text{rank}X = p$, then

$$\boxed{\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}}.$$

The linear model. The OLS method...

Theorem. Let $\mathbb{E}\varepsilon = \mathbf{0}$, $\mathrm{var}\varepsilon = \sigma^2 I$ and $\mathrm{rank}X = p$.
Then $\widehat{\beta} = (X^\top X)^{-1} X^\top \mathbf{Y}$ is an unbiased estimator of β,
and $\mathrm{var}\widehat{\beta} = \sigma^2 (X^\top X)^{-1}$.
Proof. If $\mathrm{rank}X = p$, then $(X^\top X)$ is invertible. Then

$$\mathbb{E}\widehat{\beta} = (X^\top X)^{-1} X^\top \mathbb{E}\mathbf{Y} = \beta,$$

as $\mathbb{E}\mathbf{Y} = X\beta$. And

$$\mathrm{var}(\widehat{\beta}) = (X^\top X)^{-1} X^\top (\mathrm{var}(\mathbf{Y})) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1},$$

because $\mathrm{var}(\mathbf{Y}) = \mathrm{var}(\varepsilon) = \sigma^2 I$.

The linear model. The Gauss-Markov theorem

In the homoscedastic case $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}$ is
BLUE=Best Linear Unbiased Estimator.
That is
Theorem. If $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathrm{var}\boldsymbol{\varepsilon} = \sigma^2 I$
and $\mathrm{rank} X = p$,
then $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}$
is the best linear unbiased estimator of the parameter vector $\boldsymbol{\beta}$.

SST, SSR, SSE

Statistical programs for the linear model calculate the following quantities.

Sum of Squares Total, denoted by SST, is the squared differences between the observed dependent variable and its mean. Another notation for the SST is TSS i.e. Total Sum of Squares:
$\sum_{i=1}^{n}(y_i - \bar{y})^2$

Sum of Squares due to Regression, SSR, is the sum of the differences between the predicted value and the mean of the dependent variable. Another notation for the SSR is ESS i.e. Explained Sum of Squares:
$\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$

SST=SSR+SSE

Sum of Squares Error, or SSE is the difference between the observed value and the predicted value. It is also known as RSS or residual sum of squares (residual as in: remaining or unexplained). We should minimize it.
$\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$
For linear regression with intercept and for the linear model with intercept the following result is true.
The total variability of the data set is equal to the variability explained by the regression line plus the unexplained variability, known as error: SST=SSR+SSE, i.e.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$$

RMSE, R-squared

Root Mean Squared Error, RMSE is a measure of the differences between values predicted by a model or an estimator and the values observed:
$\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}$
Small RMSE shows that our model describes well the phenomenon.

Coefficient of determination, R^2 or r^2 "R-squared" is the proportion of the variance in the dependent variable that is predictable from the independent variable(s):

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 is between 0 and 1. If it is close to 1, then our model fits well to the data.
Adjusted R^2 is $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

F and t statistics

Assuming normal distribution, F and t statistics are used.
F-statistic is used to test our model versus the constant model. At small p-value we reject the constant model.

t-statistic is used for testing whether a parameter equals zero. For small p-value we reject that the parameter equals zero.