Regression analysis
Two variables
(Montgomery and Runger: ch 11
Brani Vidakovic: ch 14)
Reminder
Covariance Defined

Covariance is a number quantifying average dependence between two random variables.

The covariance between the random variables $X$ and $Y$, denoted as $\text{cov}(X, Y)$ or $\sigma_{XY}$ is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$  \hfill (5-14)

The units of $\sigma_{XY}$ are units of $X$ times units of $Y$.

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$. 

Correlation is “normalized covariance”

• Also called: Pearson correlation coefficient

\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]

is the covariance normalized to be \[-1 \leq \rho_{XY} \leq 1\]

Karl Pearson (1852–1936)
English mathematician and biostatistician
Covariance and Scatter Patterns

Figure 5-13 Joint probability distributions and the sign of cov(X, Y). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.
Regression analysis

• Many problems in engineering and science involve sample in which two or more variables were measured. They may not be independent from each other and one (or several) of them can be used to predict another

• Everyday example: in most samples height and weight of people are related to each other

• Biological example: in a cell sorting experiment the copy number of a protein may be measured alongside its volume

• **Regression analysis** uses a sample to build a model to predict protein copy number given a cell volume
Sir Francis Galton, (1822 -1911) was an English statistician, anthropologist, proto-geneticist, psychometrician, eugenicist, (“Nature vs Nurture”, inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

Invented both correlation and regression analysis when studied heights of fathers and sons

Found that fathers with height above average tend to have sons with height also above average but closer to the average. Hence “regression” to the mean
Two variable samples

- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don’t know how dirty was the air to begin with

Table 11-1 Oxygen and Hydrocarbon Levels

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>Hydrocarbon Level x(%)</th>
<th>Purity y(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>90.01</td>
</tr>
<tr>
<td>2</td>
<td>1.02</td>
<td>89.05</td>
</tr>
<tr>
<td>3</td>
<td>1.15</td>
<td>91.43</td>
</tr>
<tr>
<td>4</td>
<td>1.29</td>
<td>93.74</td>
</tr>
<tr>
<td>5</td>
<td>1.46</td>
<td>96.73</td>
</tr>
<tr>
<td>6</td>
<td>1.36</td>
<td>94.45</td>
</tr>
<tr>
<td>7</td>
<td>0.87</td>
<td>87.59</td>
</tr>
<tr>
<td>8</td>
<td>1.23</td>
<td>91.77</td>
</tr>
<tr>
<td>9</td>
<td>1.55</td>
<td>99.42</td>
</tr>
<tr>
<td>10</td>
<td>1.40</td>
<td>93.65</td>
</tr>
<tr>
<td>11</td>
<td>1.19</td>
<td>93.54</td>
</tr>
<tr>
<td>12</td>
<td>1.15</td>
<td>92.52</td>
</tr>
<tr>
<td>13</td>
<td>0.98</td>
<td>90.56</td>
</tr>
<tr>
<td>14</td>
<td>1.01</td>
<td>89.54</td>
</tr>
<tr>
<td>15</td>
<td>1.11</td>
<td>89.85</td>
</tr>
<tr>
<td>16</td>
<td>1.20</td>
<td>90.39</td>
</tr>
<tr>
<td>17</td>
<td>1.26</td>
<td>93.25</td>
</tr>
<tr>
<td>18</td>
<td>1.32</td>
<td>93.41</td>
</tr>
<tr>
<td>19</td>
<td>1.43</td>
<td>94.98</td>
</tr>
<tr>
<td>20</td>
<td>0.95</td>
<td>87.33</td>
</tr>
</tbody>
</table>
\[ Y = \beta_0 + \beta_1 X + \epsilon \]

_Figure 11-1_ Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

\[ Y = 75 + 15 \cdot X + \epsilon \]
Linear regression

The **simple linear regression model** is given by

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

\( \varepsilon \) is the **random error term**

slope \( \beta_1 \) and intercept \( \beta_0 \) of the line are called **regression coefficients**

**Note:** \( Y \), \( X \) and \( \varepsilon \) are random variables

The minimal assumption: \( E(\varepsilon \mid x) = 0 \) \( \Rightarrow \)

\[ E(Y \mid x) = \beta_0 + \beta_1 x + E(\varepsilon \mid x) = \beta_0 + \beta_1 x \]
$Y = \beta_0 + \beta_1 X + \epsilon \quad E(\epsilon|x) = 0 \quad \forall x$

How does one find $\beta_0$ & $\beta_1$?

$\text{Cov}(Y, X) = \text{Cov}((\beta_0 + \beta_1 X + \epsilon), X) = \text{Cov}(\beta_0, X) + \beta_1 \text{Cov}(X, X) + \text{Cov}(\epsilon, X)$

$\text{Cov}(\beta_0, X) = 0$ since $\beta_0$ is constant

$\text{Cov}(X, X) = E(X^2) - E(X)^2 = \text{Var}(X)$

$\text{Cov}(\epsilon, X) = E(\epsilon \cdot X) - E(\epsilon) \cdot E(X) = E(\epsilon \cdot X) = \sum_{all \ x} x \cdot E(\epsilon|x) = 0$

Thus

$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \beta_0 = E(Y) - \beta_1 E(X)$
Method of least squares

- The **method of least squares** is used to estimate the parameters, $\beta_0$ and $\beta_1$ by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

**Figure 11-3** Deviations of the data from the estimated regression model.
Traditional notation

**Definition**

The **least squares estimates** of the intercept and slope in the simple linear regression model are

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}
\]

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right) \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)}{\sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2} = \frac{S_{xy}}{S_{xx}} \tag{11-8}
\]

where \( \bar{y} = (1/n) \sum_{i=1}^{n} y_i \) and \( \bar{x} = (1/n) \sum_{i=1}^{n} x_i \).
11-2: Simple Linear Regression

Definition

The least squares estimates of the intercept and slope in the simple linear regression model are

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

(11-7)

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left( \frac{\left( \sum_{i=1}^{n} y_i \right)\left( \sum_{i=1}^{n} x_i \right)}{n} \right) n^2}{\sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 n^2}
\]

(11-8)

where \( \bar{y} = (1/n) \sum_{i=1}^{n} y_i \) and \( \bar{x} = (1/n) \sum_{i=1}^{n} x_i \).
11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

(11-24)

Symbolically,

\[ SS_T = SS_R + SS_E \]  

(11-25)
11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

**VERY COMMONLY USED**

- The quantity

\[ R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \]

is called the **coefficient of determination** and is very commonly used to judge the adequacy of a regression model.

- \(0 \leq R^2 \leq 1\);
- We often refer (loosely) to \(R^2\) as the amount of variability in the data explained or accounted for by the regression model.
11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R²)

• For the oxygen purity regression model,
  \[ R^2 = \frac{SS_R}{SS_T} \]
  \[ = \frac{152.13}{173.38} \]
  \[ = 0.877 \]

• Thus, the model accounts for 87.7% of the variability in the data.
Estimating $\sigma^2_{\varepsilon}$

An unbiased estimator of $\sigma^2_{\varepsilon}$ is

$$\hat{\sigma}^2_{\varepsilon} = \frac{SSE}{n - 2} \quad (11-13)$$

where $SS_E$ can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1S_{xy} \quad (11-14)$$
11-3: Properties of the Least Squares Estimators

- Slope Properties
  \[ E(\hat{\beta}_1) = \beta_1 \]
  \[ V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{s_{xx}} = \frac{\hat{\sigma}^2}{n \cdot \hat{c}^2} \]

- Intercept Properties
  \[ E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \frac{\hat{\sigma}^2}{n} \left[ \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right] = \frac{\hat{\sigma}^2}{n} \left[ 1 + \frac{\mu^2}{\hat{c}^2} \right] \]
11-4: Hypothesis Tests in Simple Linear Regression

Figure 11-5  The null hypothesis $H_0: \beta_1 = 0$ is not rejected.
11-4: Hypothesis Tests in Simple Linear Regression

The null hypothesis $H_0: \beta_1 = 0$ is rejected.

Figure 11-6 The null hypothesis $H_0: \beta_1 = 0$ is rejected.
11-4.1 Use of Z-tests for large $n$

An important special case of the hypotheses of Equation 11-18 is

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the significance of regression. Failure to reject $H_0$ is equivalent to concluding that there is no linear relationship between $X$ and $Y$. 

\
11-4: Hypothesis Tests in Simple Linear Regression

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

\[ Z = \frac{\hat{\beta}_1}{\hat{\sigma}_\varepsilon \sqrt{1/n}} \]

Choose \( \alpha \) (e.g., \( \alpha = 0.05 \))

Confidence in rejecting \( H_0 \)

Reject \( H_0 \) if \( |Z| > Z_{\alpha/2} = 1.96 \)
11-4.1 Use of $t$-tests for smaller $n$.

The number of degrees of freedom is $n-2$.

One can always fit a straight line through two points so one needs $n \geq 3$. 
11-4: Hypothesis Tests in Simple Linear Regression

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

\[ T = \frac{\hat{\beta_1}}{\hat{\sigma}/\sqrt{n}} \]

Choose \( \alpha \)
(e.g. \( \alpha = 0.05 \) for 95% confidence
in rejecting \( H_0 \))

\[ t_{\alpha/2, n-2} \]

is such that
\[ 1 - \frac{\alpha}{2} = \text{cdf} (t_{\alpha/2, n-2}) \]

Reject \( H_0 \) if \( |T| > t_{\alpha/2, n-2} \)
Human T cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2010
- All samples are from T cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene’s expression level in a given sample as measured by the microarray technology

A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations I Corresponding author


Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this ‘expression space’ is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and
“Let’s Make a Deal” show with Monty Hall aired on NBC/ABC 1963-1986
Gene Expression “Wheel of Fortune”

- Each group gets a pair of genes that are known to be correlated.
- Each group also gets a random pair of genes selected by the “Wheel of Fortune”. They may or may not be correlated.
- Download (log-transformed) `expression_table.mat`
- Run command `fitlm(x,y)` on assigned and random pairs
- Record $\beta_0$, $\beta_1$, $R^2$, P-value of the slope $\beta_1$ and write them on the blackboard
- Validate Matlab result for $R^2$ using your own calculations
- Look up gene names (see `gene_description` in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?
Correlated pairs
plausible biological connection based on short description

\[ g_1=1994; \quad g_2=188; \quad \text{group 1} \]
\[ g_1=2872; \quad g_2=1269; \quad \text{group 2} \]
\[ g_1=1321; \quad g_2=10; \quad \text{group 3} \]
\[ g_1=886; \quad g_2=819; \quad \text{group 4} \]
\[ g_1=2138; \quad g_2=1364; \quad \text{group 5} \]

no obvious biological common function

\[ g_1=1+\text{floor}(\text{rand.}*3000); \quad g_2=1+\text{floor}(\text{rand.}*3000); \]
\[ \text{disp}([g_1, \; g_2]) \]
Random pairs

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1;
disp([g1,g2]);
Matlab code

- load expression_table.mat
- g1=2907; g2=288;
- x=exp_t(g1,:); y=exp_t(g2,:);
- figure; plot(x,y,'ko');
- lm=fitlm(x,y)
- y_fit=lm.Fitted;
- hold on; plot(x,lm.Fitted,'r-');
- SST=sum((y-mean(y)).^2);
- SSR=sum((y_fit-mean(y)).^2);
- SSE=sum((y-y_fit).^2);
- R2=SSR./SST
- disp([gene_names(g1), gene_names(g2)]);
- disp(gene_description(g1)); disp(gene_description(g2));