BME 2200: BME Biostatistics and Research Methods

Lecture 7: Linear Regression

Lucas C. Parra
Biomedical Engineering Department
City College of New York

parra@ccny.cuny.edu
Content, Schedule

1. Scientific literature:
   - Literature search
   - Structure biomedical papers, engineering papers, technical reports
   - Experimental design, correlation, causality.

2. Presentation skills:
   - Report – Written report on literature search (individual)
   - Talk – Oral presentation on biomedical implant (individual and group)

3. Graphical representation of data:
   - Introduction to MATLAB
   - Plot formats: line, scatter, polar, surface, contour, bar-graph, error bars. etc.
   - Labeling: title, label, grid, legend, etc.
   - Statistics: histogram, percentile, mean, variance, standard error, box plot

4. Biostatistics:
   - Basics of probability
   - t-Test, ANOVA
   - Linear regression
   - Error analysis
   - Confidence intervals, sensitivity, specificity
Linear Trends

Is there a linear trend in these two variables $x$ and $y$?

$$(x_i, y_i) \quad i = 1, \ldots, N$$
And if we agree that there is one, which one exactly is the trend? Here are 3 examples of *linear models* of the data that are optimal in some sense:
Assume that $x$ is the *independent variable* (which we can choose freely) and $y$ is the *dependent variable* (which we observe as a result). A sensible linear model is one that predicts the outcome $y$ from $x$ with minimum error.

$$\begin{align*}
(x_i, y_i) & \quad i = 1, \ldots, N \\
\text{Prediction error} & \quad \text{or 'noise'} n \\
\hat{y}_i & = bx_i + a
\end{align*}$$
Least squares regression

The simplest possible relation between two graded variables $x$ and $y$ is a (noisy) linear dependency:

$$y = bx + a + e$$

where $b$ is the slope, $a$ is the $y$-axis intercept and $e$ is some error.

If we are given samples $x_i, y_i$ we can find optimal parameters $a, b$ that explain the dependence of $y$ from $x$ with minimum error, or equivalently, require that the estimate computed from $x_i$

$$\hat{y}_i = bx_i + a$$

is as close as possible to the observed $y_i$

$$\hat{a}, \hat{b} = \text{argmin}_{a,b} \sum_i (y_i - bx_i - a)^2$$

$$E(a,b)$$
Least squares regression

The total prediction error, \( E(a,b) \), is a function of the desired parameters. Its minimum is easily obtained as

\[
\hat{a} = \frac{\langle y \rangle \langle x^2 \rangle - \langle x \rangle \langle xy \rangle}{\langle x^2 \rangle - \langle x \rangle^2}
\]

\[
\hat{b} = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2}
\]

where the angular brackets represent the mean of its content, e.g.

\[
\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i
\]
Significance of trend?

The estimate $\hat{b}$ for the slope has been computed from a random sample. It is therefore a random variable. With a different sample we would have obtained a different estimate for the slope.

Therefor the question is: How significant is this trend?

Null Hypothesis: Assume that there is NO DEPENDENCE between the variables $x$ and $y$, and the true (unobservable) slope really is zero.

We will determine the likelihood that we would have observed the specific non-zero slope estimate by chance when truly the slope is zero.

If the likelihood $p$ is below 0.01 (1%) we will conclude that there is a reasonable chance (99% or higher) that $y$ really depends on $x$. We will then call the the trend statistically significant.

Statistically significant does not mean that the dependence is strong! It simply states that the dependance is detectable above chance.
The slope estimate $\hat{b}$ is a sum of independent random variables. Due to the central limit theorem it is therefore approximately Gaussian distributed. Its standard deviation (=standard error of the estimate) is given by

$$s_b = \frac{1}{\sqrt{N - 2}} \frac{s_e}{s_x} e_i = y_i - \hat{y}_i$$

where $s_e$ and $s_x$ are the standard deviation of the estimation error $e_i$ and the data $x$ respectively. Because in the estimation process we had 2 parameters to choose the normalization is now with $N - 2$.

Note that this definition is slightly different than Glantz who computes $s_e$ by normalizing with $N - 2$ and instead uses in $N - 1$ in the formula for $s_b$ above. The result for $s_b$ is the same in both cases.
t-test for trend

The test for significance is the same than for any other Gaussian variable. In this case the second “class” is the null hypothesis, which postulates that the mean is 0, and since we are certain about this mean the corresponding standard error is 0 as well.

The $t$-statistics simplifies then to (with $\nu = N - 2$ degrees of freedom)

$$ t = \frac{\hat{b}}{s_{\hat{b}}} $$

Example: In this sample with $N = 17$ students

$$ \hat{b} = 1.4 \text{ Lb/in.} \quad s_{\hat{b}} = 1.5 \text{ Lb/in.} $$

which gives $t = 0.94$ and $p = 0.36$. Hence there is no significant dependence of weight on height in this sample.
Significance vs. Correlation

- A statistically significant slope however does not mean that the dependence of the variables is strong.
- In particular for large $N$ we may detect a statistically significant trend even though the dependence may be very weak.

Example:
Dotted line is the true slope ($b=1$ in all cases) and solid line is the estimated slope. The noise std. and number of samples changes: $s_e=1$, $s_e=5$ and $N=10$, $N=300$. 
Correlation – $r^2$ value

To measure strength of dependence between $x$ and $y$ consider the standard deviation of the residual $e$ and the standard deviation of the dependent variable $y$:

If $\hat{y}$ is a perfect model for $y$, then $s_e = 0$
If $x$ is not predictive of $y$ at all, then $s_e = s_y$

A sensible measure of how well $y$ can be predicted from $x$ is therefore the $r$ squared value

$$ r^2 = 1 - \frac{s_e^2}{s_y^2} $$

Strong dependency: $r^2 = 1$
No dependency: $r^2 = 0$
Correlation coefficient

Turns out that this $r$ value is exactly equal the **correlation coefficient** defined as

$$r = \frac{R_{yx}}{\sqrt{R_{yy} R_{xx}}}$$

with the **covariance**

$$R_{yx} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})$$

For the diagonal terms this simplifies to: $R_{xx} = s_x^2$, $R_{yy} = s_y^2$

The definition of the correlation coefficient is symmetric in $x$ and $y$. No variable has a preferred role as dependent or independent.

It is a normalized measure of how variables vary together without presupposing any sort of causal relationship.
Significance of Correlation?

Zero correlation means there is no linear dependence between two variables. But once again, when we measure a correlation from random measurements we may obtain a non-zero value just by chance. How large must the correlation be to be statistically significant?

**Null Hypothesis:** Assume there is no correlation. The true value is \( r = 0 \).

Under the Null hypothesis the measured correlation \( r \) follows again a \( t \)-statistics with \( N-2 \) degrees of freedom.

\[
t = \sqrt{(N-2)} \frac{r}{\sqrt{1-r^2}}
\]

We can use once again the \( t \)-test to determine significance of a non-zero correlation.
Assignment

Assignment 8:

1. Reproduce the figure on slide number 10 using the data posted online (it may differ from the data shown in class).

2. Compute the optimal slope $b$, the axes intercept $a$, the $t$ statistics, the $p$ value, and the $r^2$ valued. Is there a statistically significant trend in the data?

3. Compare your results with the matlab function `regress()` (remember that $F = t^2$).

4. Read chapter 8 in Glantz.
Which regression?

If the noise is in the variable $x$ and we can set $y$ exactly it may be more reasonable to estimate the dependency of $x$ from $y$. If both variables are noisy maybe we should minimize the total noise.
Principal component

When there is no preferred dimension it is more reasonable to minimize the error in all variables. Here now we use a different notation (axes are called \( x_1 \), \( x_2 \), and vectors are in bold \( \mathbf{x} = [x_1 \ x_2] \))
Principal component

Given $N$ data points $x_1, ..., x_N$, the optimization problem is to find the orientation $w$ that minimizes the total reconstruction error:

$$ w = \arg \min_{\|w\|=1} \sum_{i=1}^{N} \| n_i \|^2 $$

One can show that the solution is given by the $w$ that solves the eigenvalue equation for the data covariance matrix $R_{xx}$

$$ R_{xx} \ w = \lambda \ w $$

$$ R_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T $$