Correlation and Linear regression

Prof. Dr. Akın Pala

Various tests

- Student's t-test
- Correlation
- Pearson correlation coefficient
- Spearman's rank correlation coefficient
- Regression analysis
- Analysis of variance (ANOVA)
- Chi-square test
- Factor analysis
- Repeated Measures/Time series analysis

Correlation

- The correlation coefficient is commonly used to quantify the degree of association between two variables.
- Finding of a correlation between two variables does not imply that a change in one variable causes a corresponding change in another
  - that's why you still need experiments.

Correlation

- Measures of association are often useful in characterizing the broad relationship between two variables
  - They do not always reflect the underlying and significant variation in the association between specific cases.
- Lung cancer is strongly associated with lifelong smoking
  - Some non smokers get lung cancer
  - Some heavy smokers live to be 100.
  - There is a broad relationship but there are a lot of exceptions.

Correlation

- If you have a very large population, and unless the association between two variables is 100.00%, exceptions will always occur.
- In the case of cancer, there are other causes of cancer beyond direct smoking (such as passive smoking) and also some protective measures for certain individuals that are genetic.
- You might be able to construct a model of the relationships between two variables in the large, but it’s always possible to refine this model until all variation is explained.

Pearson's product–moment correlation coefficient, r

- The more the relationship between two variables is similar to a straight line, the more correlated the variables are.
- If the values are only weakly scattered around a straight line in the scatter plot, the variables are said to be weakly correlated;
  - Weak scattering means that the values are generally clustered around the values predicted by a straight line, but where there is significant deviation or error from the predicted values.
- Correlation can be either positive or negative

Rho

- r is an estimate of the population parameter rho (ρ)->Greek letter for r.
- In the population, if $\rho = 0$ for two variables, the two variables are said to be independent.
- The formula for rho:

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

Correlation

- The sample correlation coefficient is calculated as the sum of products of the two standardized variables divided by the *degrees of freedom.*

$$r = \frac{\sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})}{\sigma_x}\right)\left(\frac{(y_i - \overline{y})}{\sigma_y}\right)}{n-1}$$

Correlation

- The correlation is derived from the sum of products of standardized deviations from the mean for the (x, y) set of coordinates.
- Why do you need to standardize the variables? It's the same old question of comparing apples and oranges
- If you divide the deviations of each case from the mean by the standard deviation, you can express everything in std deviations as a unit of measure.

Correlation

- For example, you may want to measure the relationship between
- Height and weight, measured in centimeters and kilograms, respectively.
- Measure egg size (jumbo, large, medium, small) and egg yield per year (300, 250 etc.)
- Percent milk fat (3 percent) and milk yield per lactation (750 kg, 900 kg etc)
- Cycle threshold (15–40) and number of genes (5 or 6) investigated.

Correlation

- To compute r:
- For each (x, y) set of coordinates, subtract the mean from each observation for x and y.
- Divide by the corresponding standard deviation.
- Multiply the two results together.
- The result is then added to a sum.
- The sum is divided by the degrees of freedom, n – 1.

Correlation

- r always ranges in value from -1 to 1
  - with values close to zero representing weak relationships, and high values representing strong relationships
    - either strongly negative or strongly positive.
- r=1.00 means that the two values are completely or perfectly positively correlated
- r=-1.00 means perfectly negatively correlated
- r=0.00 means that there is no <u>linear</u> relationship between two variables.

Correlation example

- A psychologist is interested in the relationship between net wealth and attractiveness, rated by an expert panel of five judges with a combined score out of 10.

	Wealth ($m)	Attractiveness
1.	1.21	2.44
2.	2.24	5.73
3.	1.20	2.93
4.	2.39	5.69
5.	1.10	2.74
6.	1.45	4.26
7.	2.29	5.11
8.	2.33	5.58
9.	1.13	2.42
10.	2.39	5.52

$$r = \frac{\sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sigma_x}\right)\left(\frac{(y_i - \bar{y})}{\sigma_y}\right)}{n-1}$$

Correlation

- The mean and standard deviation of wealth (*x) and attractiveness (y) are 1.773* and 0.593765, and 4.242 and 1.454203 respectively. *Computation of deviations and normalization:*

	$x-\bar{x}$	$\dfrac{x-\bar{x}}{\sigma_x}$	$y-\bar{y}$	$\dfrac{y-\bar{y}}{\sigma_y}$
1.	-0.563	-0.94819	-1.802	-1.23917
2.	0.467	0.786506	1.488	1.023241
3.	-0.573	-0.96503	-1.312	-0.90221
4.	0.617	1.039132	1.448	0.995734
5.	-0.673	-1.13345	-1.502	-1.03287
6.	-0.323	-0.54399	0.018	0.012378
7.	0.517	0.870715	0.868	0.596891
8.	0.557	0.938082	1.338	0.920092
9.	-0.643	-1.08292	-1.822	-1.25292
10.	0.617	1.039132	1.278	0.878832

Correlation

- Computation of products:

1. 1.174961
2. 0.804786
3. 0.87066
4. 1.034699
5. 1.170699
6. -0.00673
7. 0.519721
8. 0.863121
9. 1.356812
10. 0.913222

- Σ = 8.701

$$r = \frac{\sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sigma_x}\right)\left(\frac{(y_i - \bar{y})}{\sigma_y}\right)}{n-1}$$

Correlation

- Thus:
- $r = \Sigma / (n - 1)$
- $= 8.701/9$
- $= 0.967$
- the two variables wealth and attractiveness are highly positively correlated, meaning that people with low wealth tend to be less attractive than individuals with high wealth.
    - wealthy individuals tend to be more attractive than individuals with lower wealth.

Correlation

- A large number of variables in the physical world are correlated, but the relationship is not meaningful or causative.
- If you find a highly correlated measure that is fast, easy, and cheap to obtain, and your only interest is prediction, select whichever correlated variable gives you the highest *r.*
- You may look at many variables and their correlations, and start checking on the relationships you find significant.

Correlation, test for significance

- The significance of the correlation coefficient can be evaluated by using the *t* statistic.
- For example, if you believe there is a relationship between two variables in the population, you can test against a null hypothesis that $\rho = 0$, using the *t* statistic and the estimate of *r:*

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Correlation

- For the association between wealth and attractiveness, we can test the null hypothesis that they are independent:

$$t = \frac{0.967}{\sqrt{\frac{1-0.967^2}{10-2}}} = \frac{0.967}{\sqrt{\frac{0.065}{8}}} = 10.73$$

- Using the tables for t, t=10.73 with df=8
- P<0.001=highly significant

Coefficient of Determination

- The correlation coefficient alone provides an indication of the strength and direction of relationship.
- But it doesn’t directly tell you the proportion of variation in one variable that can be accounted for by the other.

Coefficient of Determination

- Nor can relative *r values be* directly compared in proportion, e.g., you can't say that $r = 0.2$ *represents double* the correlation of $r = 0.1$.
- The square of the correlation coefficient provides this measure, and is known as the *coefficient of determination.*
- *Thus, in the emotional* intelligence test example above, $r2 = 0.\ 9\times 0.9 = 0.81$. *Hence, 81% of the variation* in wealth can be directly attributed to attractiveness and vice versa

Coefficient of Determination

- Using linear regression, the unique variance accounted for by multiple explanatory variables can often be combined to account for the majority of variation in a response variable.
- If attractiveness uniquely explains 81% of the variation in wealth
- Level of education uniquely explains 10% of the variation
- Combining the two explanatory variables together to form a model for variation in wealth allows you to explain 91% of the variation in the variable.

Linear regression vs ANOVA

- Linear regression is an extremely valuable technique, which is often used for prediction in models where no experimental control has been applied to the collection of data.
  - For example, you may want to determine the relationship between training and performance in athletics.

Linear regression vs ANOVA

- However, where appropriate experimental design is in place, then it may be better to use the Analysis of Variance (ANOVA) technique
- ANOVA, like the *t-test, can be used to determine the* likelihood of different samples being drawn from the same population
  - Uses a hypothesis testing framework, where there is a hypothesized difference between two groups.
  - For example, you may want to determine whether athletes who have taken a mineral supplement perform better than athletes who have taken a placebo.

GLM

- Linear regression and ANOVA are both based on the *general linear model, which,* in simple terms, is the geometry and algebra of straight lines.
- The applications of the general linear model are widespread
  - Linear models have often been found to be useful approximations to more difficult real-world relationships.
- GLM includes both linear regression and ANOVA.

GLM

- Functional mathematical relationship between an independent (or explanatory) variable x and a dependent (or response) variable y:
- $y = f(x)$ means that you can calculate any value of y if you know the value of x.
- The function f( ) can be any valid mathematical function:
- $y = x$ means that the value of y is always the same as the value of x
    - $(x, y) = (1, 1), (2, 2), (3, 3)$

GLM

- $y = a*x$ means that the value of y is always some multiple of the value of x, determined by the constant a
  - where $a = 2$, $(x,y) = (1, 2), (2, 4), (3, 6)$
- $y = ax + b$ means that the value of y is always some multiple of the value of x, determined by the constant a, plus the value of a constant b
  - where $a = 2$, $b = 1$, $(x,y) = (1, 3), (2, 5), (3, 7)$
- $y = x^2$ means that the value of y is always the value of x multiplied by itself
  - where $(x,y)= (1, 1), (2, 4), (3, 9)$, etc.

GLM

- The general linear model can always be described by $y = ax + b$.
- There are usually more than one independent factor, such as gender, year, breed, age.
- If multivariate, $y1, y2, y3 = ax + b$ can be written, though each can be assigned its own slope (b1, b2, b3 etc.)

Why not just correlation

- You can run correlation analysis between two variables, see if the correlation coefficient is different than zero.
- However, the underlying model that relates the two variables may be nonlinear, so relying on correlation alone may not suggest any relationship.

Linear regression

- You have a relationship between two variables (x, y) that is described by the model $y = ax + b$.
- If the observed data perfectly matches the model, then a correlation of $r = 1.00$ will be observed
- However, there may be any number of reasons why the observed data may not actually match that predicted.

Linear regression

- There may be other independent variables that independently account for variation in y, other than x.
  - For example, athletic performance may be influenced by a range of factors, including coaching strategies, number of hours and their impact, etc.

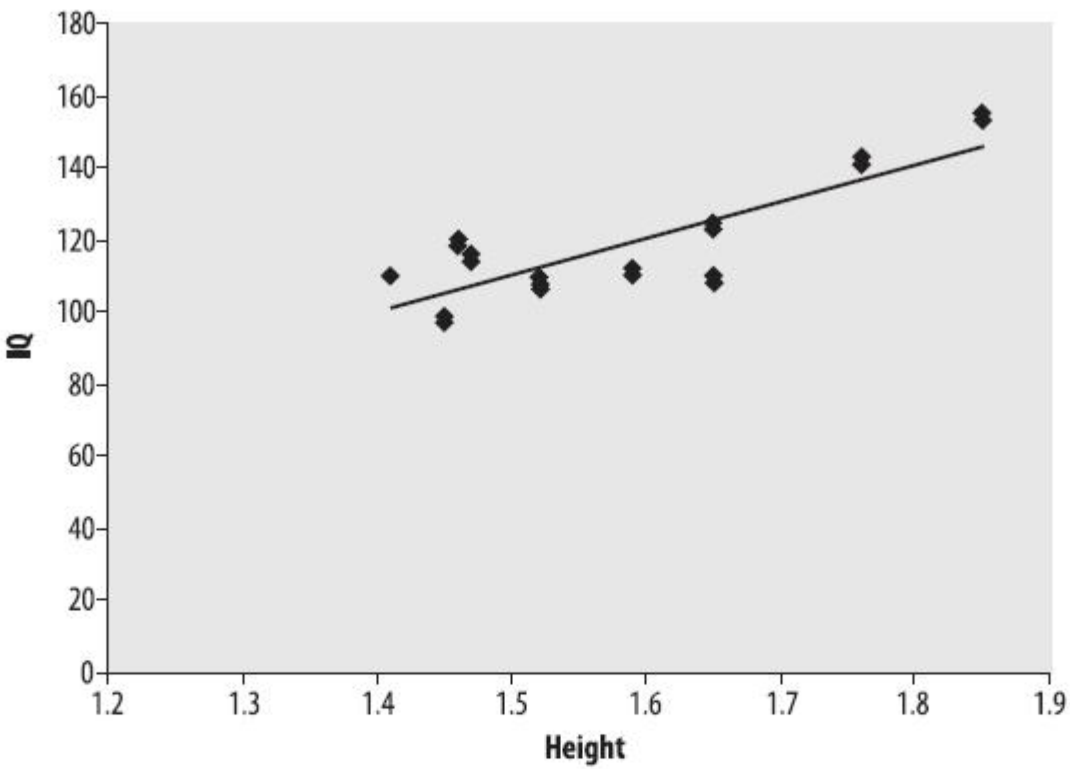
Linear regression

- There may be some element of random variation in the observations, such as bias or systematic error.
  - For example, two clinicians measuring tumor growth may use slightly different procedures to arrive at a different volume measurement from the same observation.
- The variables may represent phenomena that are not fixed in value over time.
  - For example, in a study of attitudes, participants' attitudes may change during the experimental period.

Linear regression, example

- A psychologist is interested in the relationship between height and intelligence (IQ).
- The psychologist believes that there is good evidence for a causal relationship between height and IQ
  - IQ is a dependent variable, and height is an independent variable.
- Some example measures relating IQ to height are shown in the next slide
- The correlation is high, at $r = 0.83$

IQ and height



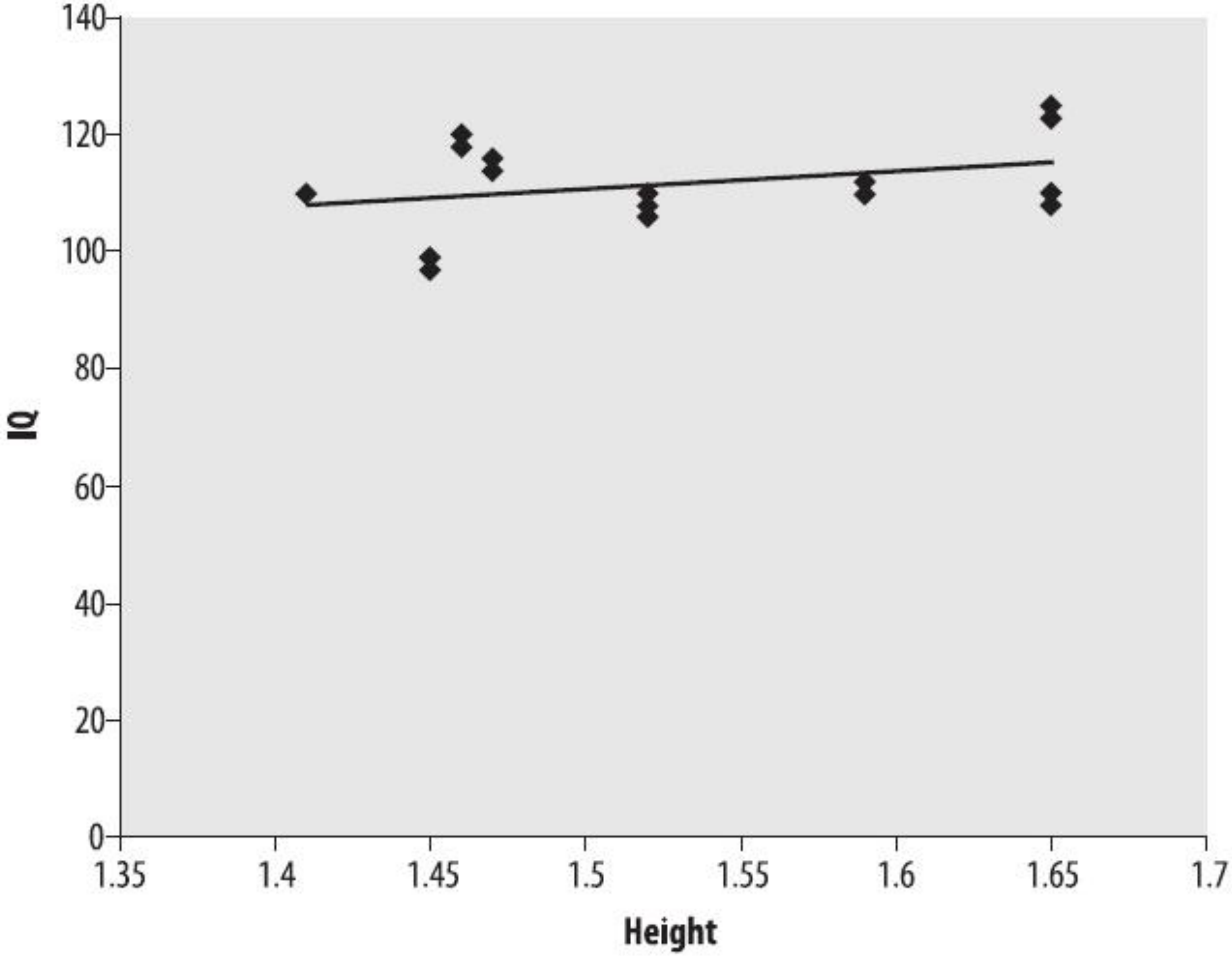
Linear regression, example

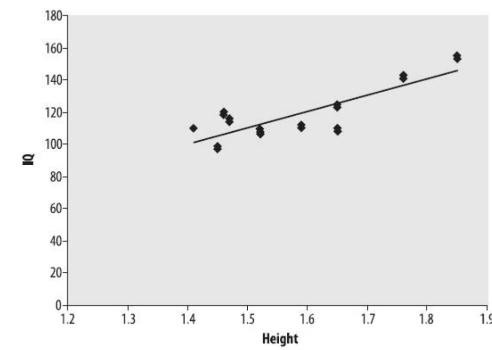
- Using the coefficient of determination, more than 67% of the variability in IQ can be attributed to height.
- Looking at the plot of the data, there is a clear linear relationship between IQ and height
  - As height rises, IQ also rises correspondingly.
- However, notice that there are also many cases where IQ $\approx$ 110, for a range of different heights (1.4m–1.65m).

Linear regression, example

- There is obviously a strong effect of four observations of large heights on the quantified association
  - If the psychologist had good reasons for considering these cases as outliers, and they were removed from the analysis, then a weak correlation would be observed $r = 0.33$, with $r^2 = 0.11$, as shown in the next slide.
- In this situation, the psychologist must examine whether the first or second relationship really represent, the true state of nature.

IQ and Height in a restricted range





Linear regression

- Linear regression can be used to go one step further than correlation.
- In this case, an imaginary straight line can be drawn on the graph, representing the hypothesized linear relationship between the two variables.
- The deviations from the imaginary line can be used to calculate the correlation coefficient
  - The linear model can be used to predict all dependent variable values along the straight line, based on values of the independent variable.

Linear regression

- This type of prediction is useful with a certain level of accuracy, from the coefficient of determination—how accurate your prediction is going to be.
- Using a regression line also means that you don’t need to experimentally measure every possible value of an independent variable to predict the corresponding value of the dependent variable.

Linear regression

- As seen from the first figure, extrapolation and prediction in ranges where you haven't measured may be dangerous to rely upon.
  - For example, if the psychologist had never measured any heights greater than 165cm, he would never have realized how strongly correlated the two variables were across the range of measurable height.

Linear regression

- Regression involves picking the “best” line that fits the observed data
  - where “best” means that the deviations between each data point of the linear model and the observed data are minimized.
- To ensure that the best overall line is selected as the “best fit,” you need to have some way of averaging across all of the observations to ensure that no single deviation results in an inappropriate line being selected.

Linear regression

- In linear regression, the deviations are measured against the dependent variable for each observation of the independent variable.
- It is usual to utilize the squared deviations, and sum them all together:

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Linear regression

- SSE=sum of squared errors
- N=# of cases
- y_i=the observed independent variable
- $\hat{y}_i$ =the expected value of the observation, as predicted from the model.

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Linear regression

- The goal of least-squares is to find the line that minimizes SSE for all observations.
- Error here means the quantitative difference between the expected and observed values.
- To most effectively minimize the SSE, in terms of a model fit, it is more desirable to reduce larger than smaller differences.
- Large differences are weighted more highly than smaller differences by squaring the deviations.

Formulas for calculation

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - b - ax_i)^2$$

$$\sigma_x^2 = \sum x^2 - \frac{(\sum x)^2}{N} \qquad \sigma_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{N}$$

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \qquad b = \frac{\sum y}{N} - a\frac{\sum x}{N}$$

IQ calculations

- Assume:

$$\sum x = 33.25$$

$$\sum y = 2486$$

$$\sum x^2 = 53.01$$

$$\sum y^2 = 299676$$

Calculations:

$$\bar{x} = \frac{33.25}{21} = 1.58$$

$$\bar{y} = \frac{2486}{21} = 118.38$$

IQ calculations

$$\sigma_x^2 = \sum x^2 - \frac{(\sum x)^2}{N} \quad \sigma_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{N}$$

$$\sigma_x^2 = 53.01 - \frac{(33.25 * 33.25)}{21} = 0.36$$

$$\sigma_{xy} = 3973.04 - \frac{(33.25 * 2486)}{21} = 36.87$$

IQ calculations

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \qquad b = \frac{\sum y}{N} - a\frac{\sum x}{N}$$

$$a = \frac{36.87}{0.36} = 102.35$$

$$b = 118.38 - (102.35 * 1.58) = -43.67$$

IQ calculations

- The model that best fits:
  - $Y = ax + b = 102.35x - 43.67$
- So, to predict IQ based on height:
  - $IQ = 102.35 * height - 43.67$

IQ calculations

- Accuracy of the model can be estimated using coefficient of determination.
- The **coefficient of determination** (R^2) is the proportion of the variance in the dependent variable that is predictable from the independent variable.
  - If low, you need more parameters for accurate estimation
  - If $R^2=1$ then strength of the predictions is very high.

Inference

- There have been no experimental controls or manipulations
- So the relationship between IQ and height is purely predictive and not causative
- Simply developing a mathematical model and using least-squares regression and correlation analysis to determine the strength of relationship between two variables cannot be used to infer causation.

Which one is more useful?

- The correlation coefficient is useful in determining the strength of association between two variables
- Once established, the slope of the regression can be used to estimate by how much one variable rises (or falls) as a function of another.
- If the relationship is weak, then the slope of a regression line is meaningless
  - Both measures are useful, and have an important role in characterizing and predicting relationships between two variables respectively.

Correlation vs. Regression

- Correlation: X and Y relationship
- Regression: Y on X