

LAB 12

I. Correlation

- Correlation measures the strength of the linear association between two *continuous* variables.
- Population correlation: ρ
 - $-1 \leq \rho \leq 1$
 - $\rho \approx 0$ implies no linear relationship (a nonlinear relationship may exist)
 - $\rho > 0$ implies a positive linear relationship (positive correlation)
 - $\rho < 0$ implies a negative linear relationship (negative correlation)
- Pearson's correlation coefficient (r)
 - Estimates correlation of the population (ρ)
 - Can test $H_0: \rho=0$
 - Test assumes that the pairs of observations, X and Y, are obtained randomly
 - Test assumes X and Y are normally distributed
 - Test statistic follows a t distribution with $n-2$ degrees of freedom
 - Pearson's correlation coefficient is sensitive to outliers.
- Spearman's rank correlation coefficient (r_s)
 - More robust estimate of correlation
 - Since this is based on the ranks, it is considered a nonparametric statistic
 - Can also test whether $\rho=0$
 - Test assumes that the pairs of observations, X and Y, are obtained randomly
 - Test statistic follows a t distribution with $n-2$ degrees of freedom
- Remember: A large correlation does not imply causality.
- If a test of $H_0: \rho=0$ is not rejected, it does not imply that the variables are independent (think of curvilinear relationships).

II. Simple Linear Regression

- Regression measures the association between two continuous variables when one variable is treated as the *response* and the other as the *explanatory* variable
- The objective of regression is to predict the value of the response associated with a fixed value of the explanatory variable. That is, we are examining the relationship between two continuous variables, specifying how a change in the explanatory variable affects a change in the response variable.
- Review the equation for a line and interpretation of the slope and intercept
$$y=a+bx$$
- Regression concepts
 - μ_y = mean of y and σ_y = standard deviation of y
 - $\mu_{y|x}$ = mean of y given x and $\sigma_{y|x}$ = standard deviation of y given x
 - The relationship between $\sigma_{y|x}$ and ρ and σ_y :

$$\sigma^2_{y|x} = (1 - \rho^2)\sigma^2_y$$

- Since $-1 \leq \rho \leq 1$, $\sigma_{y|x} \leq \sigma_y$
- Confidence intervals for the mean value of y given a value of x vs. confidence intervals for the mean value of y

- Linear regression model

- Assumptions

1.

2.

3.

4.

- Population regression line: $\mu_{y|x} = \alpha + \beta x$
- Least squares estimated regression line: $\hat{y} = \hat{\alpha} + \hat{\beta}x$

Residuals:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\alpha} + \hat{\beta}x_i) \end{aligned}$$

- Method of Least Squares:

- Sum of squared residuals:
$$\begin{aligned} \sum e_i^2 &= (y_i - \hat{y}_i)^2 \\ &= (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \end{aligned}$$

- MLS finds the values of $\hat{\alpha}$ and $\hat{\beta}$ that minimize the sum of the squared residuals.

- Inference for regression coefficients

- If $\beta=0$ then $\mu_{y|x} = \mu_y \rightarrow$ no linear relationship between y and x
 - Test for $H_0: \beta=0$

$$t = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})} \sim t_{n-2}$$

- Confidence interval for β

$$\hat{\beta} \pm t_{n-2} \hat{se}(\hat{\beta})$$

- Inference for predicted values

- \hat{y} = predicted mean value (a.k.a., fitted values)
 - \tilde{y} = predicted individual value
 - $\sigma_{\hat{y}} \leq \sigma_{\tilde{y}}$
 - CI for \hat{y} narrower than CI for \tilde{y}
 - Do not extrapolate beyond range of data.

- Evaluation of the model
 - Coefficient of determination (R^2) – percentage of total variability explained by the model
 - Residual plots – check for homoscedasticity and linearity
 - Transformations (circle of powers) – correct for non-linearity

Example – Correlation

The data are contained in the data set “maternal mortality.dta”. Open this in Stata.

1. Suppose we wish to examine the relationship between the percentage of births attended by trained healthcare personnel -- including physicians, nurses, midwives, and other health care workers -- and the mortality rate per 100,000 live births. The values for a random sample of 20 countries appear in the table, below.
- Construct a scatter plot of the data, placing percentage of births attended on the horizontal axis and maternal mortality rate on the vertical axis.
 - What can you say about the relationship between maternal mortality rate and the percentage of attended births?
 - Calculate the Pearson coefficient of correlation in Stata.
 - Now test whether these outcomes are linearly related using that measure. State your null and alternative hypotheses and p-value.

- What do you conclude?
- If we are interested in calculating a more robust measure of association between two variables, we can order the sets of outcomes x and y from smallest to largest and compute the rank correlation coefficient instead. Spearman's rank correlation is simply Pearson's r calculated using ranks rather than actual observations. Calculate this measure of association in Stata using the following command.
- How does the Spearman estimate compare with the Pearson estimate?
- Test the hypothesis that the unknown population correlation is equal to zero using Spearman's rank correlation coefficient.

Example – Simple linear regression

The data are contained in the data set “low+birth+weight+infants-1.dta”. Open this in Stata.

2. Suppose that we are interested in the relationship between length and gestational age for the population of low birth weight infants, defined as those weighing less than 1500 grams. We will perform our analysis using the data from a sample of 100 low birth weight infants born in Boston, Massachusetts.

- Create a scatter plot of length versus gestational age. What can you say about the relationship between length and gestational age? Does the relationship appear linear? Explain the vertical lines in the scatterplot.
- What would be the equation for the *true* population regression line?
- Obtain the least squares regression line.
- What is the least squares *estimate* of the true population intercept ($\hat{\alpha}$)? Interpret this value in words.
- What is the least squares *estimate* of the true population slope ($\hat{\beta}$)? Interpret this value in words.

- Test if there is a significant linear relationship between the length and gestational age of a low birth weight infant. State the null and alternative hypotheses, calculate the test statistic, state the distribution of your test statistic, state the p-value, draw a conclusion.
- Calculate a 95% confidence interval for the slope of the true population regression line.
- How does it reflect the result of your hypothesis test?
- Now, let's recreate the scatter plot, this time including the fitted regression line.
- What is the predicted mean length for all babies born at 29 weeks gestational age?
- Produce a plot of the residuals versus the fitted values.