

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 :

$$\sigma_{y|x}^2 = (1 - \rho^2) \sigma_y^2$$

- 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. The outcomes (y) are independent.
2. For a specified value of x_i , $y_i \sim N(\alpha + \beta x_i, \sigma^2)$.
3. Straight line relationship holds.
4. The variance σ^2 is constant across all values of x .

- 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.

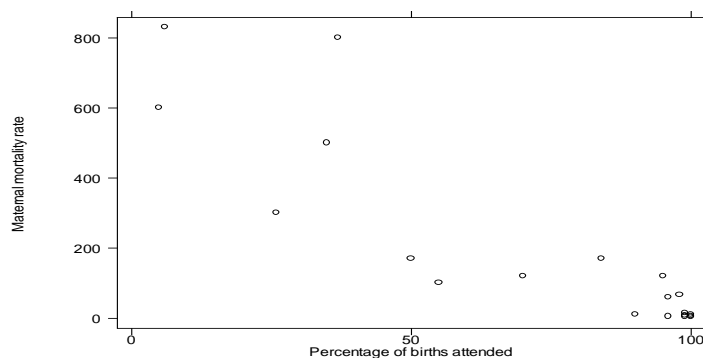
- Evaluation of the model
 - Coefficient of determination (R^2) – percentage of total variability explained by the model
 - Residual plots – check for homoscedasticity and linearity
 - Transformations (circle of powers) – correct for non-linearity

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.

twoway (scatter maternal attend)



- What can you say about the relationship between maternal mortality rate and the percentage of attended births?

As the percentage of births attended increases, the maternal mortality rate tends to be decreased. So, there is an inverse relationship between those two variables.

- Calculate the Pearson coefficient of correlation in Stata.

pwcorr maternal attend

The sample correlation is -0.88.

- Now test whether these outcomes are linearly related using that measure. State your null and alternative hypotheses and p-value.

Null hypothesis: The correlation between the percentage of births attended and the maternal mortality rate is zero.

Alternative hypothesis: The correlation between the percentage of births attended and the maternal mortality rate is not zero.

pwcorr maternal attend, sig

The p-value is near 0.0000.

- What do you conclude?

Since the p-value is less than 0.05, we can reject the null so that we can conclude that there is a linear relation between the percentage of births attended and the maternal mortality rate. As the % of births attended increases, the maternal mortality rate decreases,

- 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.

spearman maternal attend

Spearman's rho is -0.89.

- How does the Spearman estimate compare with the Pearson estimate?

While the Pearson estimate is parametric, the Spearman estimate does not assume normality and is a nonparametric approach.

- Test the hypothesis that the unknown population correlation is equal to zero using Spearman's rank correlation coefficient.

Similarly, since the p-value of the Spearman's rank estimate is less than 0.05, we can reject the null so that we can conclude that there is a correlation between the percentage of births attended and the maternal mortality rate. As the % of births attended increased, the maternal mortality rate decreases.

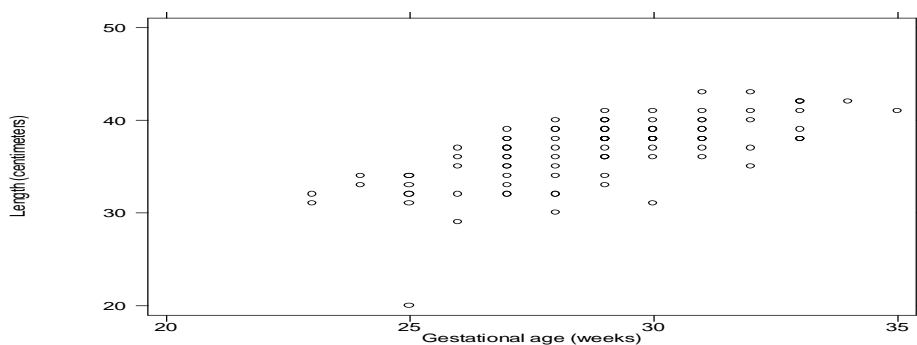
Example – Simple linear regression

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.

`twoway (scatter length gestage)`



As the gestational age (weeks) increases, the length tends to be increased. So, there is a linear relationship between those two variables. The vertical lines are due to rounding of gestational age in the dataset to the nearest week.

- What would be the equation for the *true* population regression line?

$$\mu_{\text{length}_i | \text{gestational age}_i} = \alpha + \beta \text{gestational age}_i$$

where length is the dependent variable and gestage is the independent variable

- Obtain the least squares regression line.

`regress length gestage`

Source	SS	df	MS	Number of obs = 100		
Model	575.73916	1	575.73916	F(1, 98) = 82.13		
Residual	687.02084	98	7.01041674	Prob > F = 0.0000		
Total	1262.76	99	12.7551515	R-squared = 0.4559		
				Adj R-squared = 0.4504		
				Root MSE = 2.6477		

length	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gestage	.9516035	.1050062	9.06	0.000	.7432221	1.159985
_cons	9.328174	3.045163	3.06	0.003	3.285148	15.3712

Thus, $\hat{\alpha} = 9.34$, $\hat{\beta} = 0.95$.

- What is the least squares *estimate* of the true population intercept ($\hat{\alpha}$)? Interpret this value in words.

$\hat{\alpha} = 9.34$. The expected value of length is 9.34 cm when gestational age is equal to zero. This is a nonsensical interpretation and shows the dangers of extrapolating beyond the range of the observed data.

- What is the least squares *estimate* of the true population slope ($\hat{\beta}$)? Interpret this value in words.

$\hat{\beta} = 0.95$. The expected change in length corresponding to a one week change in gestational age is 0.95 cm.

- 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.

Null: There is no linear relationship between the length and gestational age of low birth weight infants. ($\beta = 0$)

Alternative: There is a linear relationship between the length and gestational age of low birth weight infants. ($\beta \neq 0$)

Test stat: 9.06 (t distribution (98 degrees of freedom))

P-value: 0.000

Conclusion: since p-value is less than 0.05, we can reject the null and conclude that there is a significant linear relationship between the length and gestational age of a low birth weight infant. As gestational age increases, infant length increases on average 0.95 (0.74, 1.16) cm per week.

- Calculate a 95% confidence interval for the slope of the true population regression line.

95% CI for the slope is (0.7432, 1.1560)

- How does it reflect the result of your hypothesis test?

The 95% CI for the slope does not include 0. This implies that the slope is significant. This presents the same result as the hypothesis test.

- Now, let's recreate the scatter plot, this time including the fitted regression line.

`twoway (scatter length gestage)(lfit length gestage)`

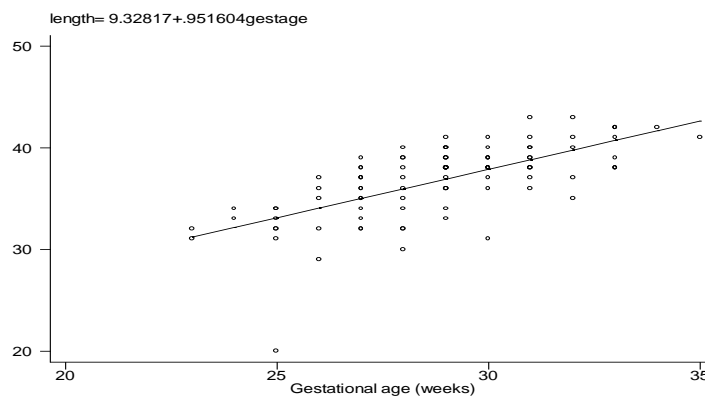
As an alternative, you can do the following:

1. Create the predicted value for each observation by typing

`predict yhat`

in the *Command* window. Note that a new variable, *yhat1*, appears in the variable window. This variable contains the predicted values based on Model 1.

2. To create the plot, choose *Graphics/Overlaid two-way graphs*. This command will create two graphs, one on top of the other. The first graph will be the scatter plot as above. The second will be a line plot of *yhat* vs. gestational age.
 - For “Plot 1”, choose “scatter” as the plot type, *gestage* as the X variable and *length* as the Y variable.
 - Click on the “Plot 2” tab. Choose “line” as the plot type, *gestage* as the X variable and *yhat1* as the Y variable. Click “OK”.



- What is the predicted mean length for all babies born at 29 weeks gestational age?

predict *yhat*

See the fitted value of the first row. Predicted value is 36.9246 at 29 weeks..

or $(9.34 + 0.95 * 25)$

- Produce a plot of the residuals versus the fitted values.

Quicker way: *rvfplot* command (just type *rvfplot* after fitting the regression)

```
predict ehat, resid
twoway (scatter ehat yhat)
```

There is no particular pattern between residual and predicted value at the zero value of residuals. This implies that the assumption of equal variances holds.

As an alternative, you can do the following:

- *Graphics/Regression diagnostic plots/Residuals-versus-fitted*. Click “OK”.

