

Quantitative Research Methods: Introduction to correlation and regression

Brendan Halpin, Sociology, University of Limerick

January 2018



Formula for multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_k X_k + e$$

$$e \sim N(0, \sigma)$$

- Interpretation of β_j
 - How much \hat{Y} changes for a 1-unit in X_j holding all other values constant
 - The estimated effect on Y of a 1-unit change in X_j , "controlling for" or "taking account" of all the other X s



Residuals

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_k X_k$$

$$Y = \hat{Y} + e$$

$$e \sim N(0, \sigma)$$

- Mean of zero
- Standard deviation of σ (RMSE)
- Normally distributed
- Should have no structured relationship to X variables



R^2

- R^2 : coefficient of multiple determination
- TSS = sum of squared deviation from the mean = $\sum(Y_i - \bar{Y})^2$
- RSS = sum of squared deviation from the regression prediction = $\sum(Y_i - \hat{Y}_i)^2$
- $R^2 = \frac{TSS - RSS}{RSS}$
- Range: 0 (no relationship) to 1 (perfect linear relationship)
- PRE: Proportional Reduction in Error



R^2 and correlation

- In bivariate regression, R^2 is the square of the correlation coefficient between Y and X
- In multiple regression, it is the square of the correlation between Y and \hat{Y}
- (In bivariate regression the correlation between X and \hat{Y} is 1)



Hypothesis testing: one parameter at a time

- t-test: $abs(\beta_j / se_j) > t$
- Interpretation:
 - Null: population value of β is 0; this variable has no influence once the other variables are taken account of



Example

```
. reg income age i.sex
```

Source	SS	df	MS	Number of obs =	959
Model	33922983.9	2	16961492	F(2, 956)	= 45.72
Residual	354670636	956	370994.389	Prob > F	= 0.0000
Total	388593620	958	405630.083	R-squared	= 0.0873
				Adj R-squared	= 0.0854
				Root MSE	= 609.09

	income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age		-3.144945	1.083398	-2.90	0.004	-5.271057 -1.018833
sex						
female		-352.678	39.51326	-8.93	0.000	-430.2208 -275.1353
_cons		1035.878	54.58935	18.98	0.000	928.7494 1143.007



Hypothesis testing: all parameters together

- F-test:
 - $\beta_1 = \beta_2 \dots = \beta_k = 0$
- Null hypothesis: no X variable has an effect once the others are taken care of.
- A "global" test: the null is that there is no relevant variable in the model
- Calculation based on TSS and RSS, but also number of cases and number of parameters estimated
- Uses F distribution (two df parameters: k and $n-k-1$, k is number of parameters, n the number of cases)



Hypothesis testing: additional parameters

- Delta F-test compares "nested" models
 - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_g X_g$
 - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_g X_g + \beta_h X_h \dots + \beta_k X_k$
- Null hypothesis: $\beta_h = \dots = \beta_k = 0$
- That is, given the variables already in the model, the additional variables contribute no explanatory power.
- Useful when adding multi-category variables, or related groups of variables

Dummy variables

With a two-category variable, we represent it a 0/1 and interpret it as the effect of being in the second category (e.g., female) compared with the first.

With more than two categories we create a set of binary variables, "indicator variables" or "dummy variables":

	d1	d2	d3	d4
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1

For m categories, m-1 dummy variables are sufficient.

We interpret the parameter as the estimated effect of being in that category relative to the omitted or "reference" category.

Example

```
. reg income age i.sex i.qual
```

Source	SS	df	MS		
Model	85960604.5	5	17192120.9	Number of obs =	959
Residual	302633015	953	317558.253	F(5, 953)	= 54.14
Total	388593620	958	405630.083	Prob > F	= 0.0000
				R-squared	= 0.2212
				Adj R-squared	= 0.2171
				Root MSE	= 563.52

	income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age		-.3897295	1.04777	-0.37	0.710	-2.445933 1.666474
sex						
female		-336.9623	36.75947	-9.17	0.000	-409.1011 -264.8234
qual						
A-level, other sub-d..		-459.9208	78.54165	-5.86	0.000	-614.0554 -305.7862
O-level, commercial...		-701.695	77.16016	-9.09	0.000	-853.1185 -550.2716
Sub-O-level, no qual		-864.9695	76.41768	-11.32	0.000	-1014.936 -715.0032
_cons		1563.508	81.83797	19.10	0.000	1402.904 1724.111