SOC 3811/5811:
BASIC SOCIAL STATISTICS

Three Variable Relationships and Multiple Regression

Multiple Regression Analysis

Multiple Regression Analysis

> "a statistical technique for estimating the relationship between a continuous dependent variable and two or more continuous or discrete independent, or predictor, variables"

For today, we will limit ourselves to…

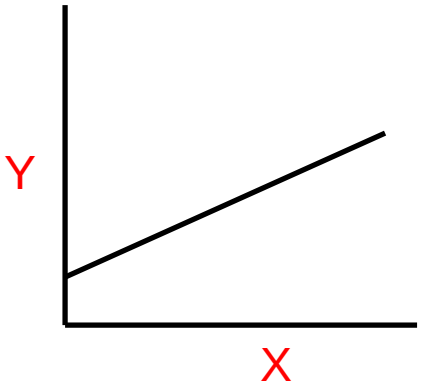
- …two predictor variables
- …continuous predictor variables

Extensions to 3+ predictor variables and to discrete predictor variables will be natural extension of what we cover today

Multiple Regression Analysis

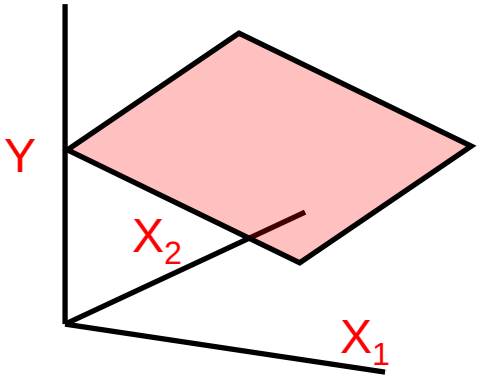
The bivariate regression prediction equation describes a 2-dimensional line

$$\hat{Y}_i = a + b_1 X_{1i}$$



The multivariate (2 independent variable) prediction equation describes a 3-dimensional plane

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i}$$



Multiple Regression Analysis

The ordinary least squares (OLS) method is used to estimate a, b_1, and b_2 … again, this method minimizes the sum of the squared residuals (or prediction errors)

To compute a, b_1, and b_2 we only need the sample means, the standard deviations, and the correlations

$$b_1 = \left(\frac{s_Y}{s_{X_1}}\right)\frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

$$b_2 = \left(\frac{s_Y}{s_{X_2}}\right)\frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

$$a = \overline{Y} - (b_1\overline{X}_1 + b_2\overline{X}_2)$$

Interpreting Multiple Regression Coefficients

How are a, b_1, and b_2 interpreted in the equation:

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i}$$

Intercept a:

The predicted value of Y when both X_1 and X_2 equal 0

Multiple regression coefficient (or slope) b_1:

The expected change in Y associated with a one unit increase in X_1, *controlling for X_2*

Multiple regression coefficient (or slope) b_2:

The expected change in Y associated with a one unit increase in X_2, *controlling for X_1*

Coefficient of Determination

As in the bivariate case we can use R^2 to express the proportion of variation in Y that is accounted for by the predictor variables

Computationally, in the model with two predictors:

$$R^2_{Y \bullet X_1 X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2 r_{YX_1} r_{YX_2} r_{X_1 X_2}}{1 - r^2_{X_1 X_2}}$$

Testing Hypotheses about $\rho^2_{Y \bullet X1X2}$

Hypothesis Testing in 6 Steps

1. State the null (H_0) and alternative (H_1) hypotheses
2. Check that the sample data conform to basic assumptions; if they do not, then do not go any further
3. Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis
4. Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level
5. Calculate the test statistic … F
6. Compare the test statistic to the critical value

Testing Hypotheses about $\rho^2_{Y \bullet X1X2}$

The hypothesis test for $\rho^2_{Y \bullet X1X2}$ is an F test with df_{NUM}=2 (the number of predictors in the model) and df_{DENOM}=N—3 (N-1 minus the number of predictors in the model)

Testing Hypotheses about $\rho^2_{Y \bullet X1X2}$

Calculate the test statistic

The F statistic when there are two predictors is

$$F_{2,N-3} = \frac{SS_{REGRESSION}/2}{SS_{ERROR}/N-3} = \frac{MS_{REGRESSION}}{MS_{ERROR}}$$

Computationally:

$$SS_{TOTAL} = (s_Y^2)(N-1)$$

$$SS_{REGRESSION} = (R^2_{Y \bullet X_1 X_2})(SS_{TOTAL})$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGRESSION}$$

Testing Hypotheses about β_1 & β_2

Hypothesis Testing in 6 Steps

1. State the null (H_0) and alternative (H_1) hypotheses
2. Check that the sample data conform to basic assumptions; if they do not, then do not go any further
3. Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis
4. Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level
5. Calculate the test statistic … t
6. Compare the test statistic to the critical value

Testing Hypotheses about β_1 & β_2

The hypothesis test for β_1 and β_2 are t tests with N-3 degrees of freedom (because MS_{ERROR} has N-3 degrees of freedom when there are two predictor variables)

Testing Hypotheses about β_1 & β_2

Calculate the test statistic

The t statistic for β_1 is

$$t_{N-3} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1 - 0}{\sqrt{\frac{MS_{ERROR}}{(s_{X_1}^2)(N-1)(1 - R_{X_1 \bullet X_2}^2)}}}$$

The t statistic for β_2 is:

$$t_{N-3} = \frac{b_2 - 0}{s_{b_2}} = \frac{b_2 - 0}{\sqrt{\frac{MS_{ERROR}}{(s_{X_2}^2)(N-1)(1 - R_{X_2 \bullet X_1}^2)}}}$$

Partial Correlation

Earlier we talked about the correlation coefficient, r, as a measure that describes the strength and direction of the association between two continuous variables

If r_{YX1} represents the bivariate correlation between Y and X_1, then $r_{YX1 \bullet X2}$ represents the partial correlation between Y and X_1 that persists after controlling for X_2

In the context of a regression model with two explanatory variables, the partial correlation between Y and X_1 is

$$r_{YX_1 \bullet X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{YX_2}}\sqrt{1 - r^2_{X_1X_2}}}$$

Testing Hypotheses about $r_{YX1 \bullet X2}$

Hypotheses tests about partial correlation coefficients are identical to hypothesis tests for the corresponding regression coefficient

If we reject the hypothesis that β_1 equals zero in the population, we are simultaneously rejecting the null hypothesis that $\rho_{YX1 \bullet X2}$ equals zero

Likewise, if we reject the hypothesis that β_2 equals zero in the population, we are simultaneously rejecting the null hypothesis that $\rho_{YX2 \bullet X1}$ equals zero

	X_1	X_2	Y	Mean	Variance
X_1	1.000			8.040	2.108
X_2	0.350	1.000		6.112	4.103
Y	0.150	0.220	1.000	4.063	5.125

$$b_1 = \left(\frac{s_Y}{s_{X_1}}\right)\frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

$$= \left(\frac{\sqrt{5.125}}{\sqrt{2.108}}\right)\frac{.150 - (.350)(.220)}{1 - .350^2} = .130$$

$$\hat{Y} = 1.716 + .130X_1 + .213X_2$$

$$b_2 = \left(\frac{s_Y}{s_{X_2}}\right)\frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

$$= \left(\frac{\sqrt{5.125}}{\sqrt{4.103}}\right)\frac{.220 - (.150 \cdot .350)}{1 - .350^2} = .213$$

$$a = \bar{Y} - (b_1\bar{X}_1 + b_2\bar{X}_2)$$

$$= 4.063 - (.130 \cdot 8.04) + (.213 \cdot 6.112) = 1.716$$

	X_1	X_2	Y	Mean	Variance
X_1	1.000			8.040	2.108
X_2	0.350	1.000		6.112	4.103
Y	0.150	0.220	1.000	4.063	5.125

$$R^2_{Y \bullet X_1 X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2 r_{YX_1} r_{YX_2} r_{X_1 X_2}}{1 - r^2_{X_1 X_2}}$$

$$= \frac{.150^2 + .220^2 - 2(.150)(.220)(.350)}{1 - .350^2}$$

$$= .054$$

	X_1	X_2	Y	Mean	Variance
X_1	1.000			8.040	2.108
X_2	0.350	1.000		6.112	4.103
Y	0.150	0.220	1.000	4.063	5.125

$$SS_{TOTAL} = (s_Y^2)(N-1)$$

$$= 5.125(99) = 507.375$$

$$F^* = F_{2,97} = 3.12$$

$$SS_{REGRESSION} = (R^2_{Y \bullet X_1 X_2})(SS_{TOTAL})$$

$$= (.054)(507.375) = 27.398$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGRESSION}$$

$$= 507.375 - 27.398 = 479.977$$

$$F_{2,N-3} = \frac{SS_{REGRESSION}/2}{SS_{ERROR}/N-3} = \frac{MS_{REGRESSION}}{MS_{ERROR}}$$

$$= \frac{27.398/2}{479.977/97} = 2.79$$

	X_1	X_2	Y	Mean	Variance
X_1	1.000			8.040	2.108
X_2	0.350	1.000		6.112	4.103
Y	0.150	0.220	1.000	4.063	5.125

$$t_{N-3} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1 - 0}{\sqrt{\dfrac{MS_{ERROR}}{(s^2_{X_1})(N-1)(1 - R^2_{X_1 \bullet X_2})}}}$$

$$t_{N-3} = 1.99$$

$$= \frac{.150 - 0}{\sqrt{\dfrac{749.977/97}{(2.108)(97)(1-.350^2)}}} = .791$$

$$t_{N-3} = \frac{b_2 - 0}{s_{b_2}} = \frac{b_2 - 0}{\sqrt{\dfrac{MS_{ERROR}}{(s^2_{X_2})(N-1)(1 - R^2_{X_2 \bullet X_1})}}} = 1.808$$

	X_1	X_2	Y	Mean	Variance
X_1	1.000			8.040	2.108
X_2	0.350	1.000		6.112	4.103
Y	0.150	0.220	1.000	4.063	5.125

$r_{YX_1} = .15$

$$r_{YX_1 \bullet X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r_{YX_2}^2}\sqrt{1 - r_{X_1X_2}^2}}$$

$$= \frac{(.150) - (.220)(.350)}{\sqrt{1 - .220^2}\sqrt{1 - .350^2}} = .08$$