Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

Statistical **MODEL**

- There are *p-1* explanatory variables
- For each *combination* of explanatory variables, the conditional distribution of the response variable *Y* is normal, with constant variance
- The conditional population mean of *Y* depends on the *x* values, as follows:

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

Control means hold constant

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\frac{\partial}{\partial x_3} E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_3$$

So β_3 is the rate at which $E[Y|\boldsymbol{x}]$ changes as a function of x_3 with all other variables held constant at fixed levels.

Increase x_3 by one unit holding other variables constant

$$
\begin{array}{rlll}
 & \beta_0 + \beta_1 x_1 + \beta_2 x_2 & +\beta_3 (x_3 + 1) & +\beta_4 x_4 \\
- & (\beta_0 + \beta_1 x_1 + \beta_2 x_2 & +\beta_3 \; x_3 & +\beta_4 x_4) \\
= & \beta_3 (x_3 + 1) - \beta_3 x_3 & & \\
= & \beta_3 & &
\end{array}
$$

So β_3 is the amount that $E[Y|\boldsymbol{x}]$ changes when x_3 is increased by one unit and all other variables are held constant at fixed levels.

It's model-based control

To "hold x_5 constant" at some particular value, like x_5=14, you don't even need data at that value.

Statistics b estimate parameters beta

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Categorical IVs

- X=1 means Drug, X=0 means Placebo
- Population mean is $E[Y|X = x] = \beta_0 + \beta_1 x$
- For patients getting the drug, mean response is $E[Y|X = 1] = \beta_0 + \beta_1$
- For patients getting the placebo, mean response is $E[Y|X = 0] = \beta_0$

Sample regression coefficients for a binary IV

- X=1 means Drug, X=0 means Placebo
- Predicted response is $\widehat{Y} = b_0 + b_1 x$
- For patients getting the drug, predicted response is

$$\widehat{Y} = b_0 + b_1 = \overline{Y}_1$$

- For patients getting the placebo, predicted response is

$$\widehat{Y} = b_0 = \overline{Y}_0$$

Regression test of b_1

- Same as an independent t-test
- Same as a oneway ANOVA with 2 categories
- Same t, same F, same p-value.

Drug A, Drug B, Placebo

- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Fill in the table

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Drug A, Drug B, Placebo

- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

Regression coefficients are *contrasts* with the category that has no indicator – the *reference* category

Indicator dummy variable coding with intercept

- Need p-1 indicators to represent a categorical explanatory variable with p categories
- If you use p dummy variables, trouble
- Regression coefficients are **contrasts** with the category that has no indicator
- Call this the **reference category**

Now add a quantitative variable (covariate)

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

Effect coding

- *p-1* dummy variables for *p* categories
- Include an intercept
- Last category gets -1 instead of zero
- What do the regression coefficients mean?

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1 x_1+\beta_2 x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

Meaning of the regression coefficients

Group	x_1	x_2	$E[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

$$\mu = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \beta_0$$

The grand mean

With effect coding

- Intercept is the *Grand Mean*
- Regression coefficients are deviations of group means from the grand mean.
- They are the non-redundant *effects*.
- Equal population means is equivalent to zero coefficients for all the dummy variables
- Last category is not a reference category

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1 x_1+\beta_2 x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

Add a covariate: Age = x_1

Group	x_2	x_3	$E[Y\|\boldsymbol{X}=\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = \beta_0 + \beta_2 \quad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \quad + \beta_1 x_1$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_2 - \beta_3 + \beta_1 x_1$

Regression coefficients are deviations from the average conditional population mean (conditional on x_1).

So if the regression coefficients for all the dummy variables equal zero, the categorical explanatory variable is unrelated to the response variable, controlling for the covariate(s).

Effect coding is very useful when there is more than one categorical explanatory variable and we are interested in *interactions* --- ways in which the relationship of an explanatory variable with the response variable *depends* on the value of another explanatory variable.

Interaction terms correspond to products of dummy variables.

Analysis of Variance

And testing

Analysis of Variance

- Variation to explain: **Total Sum of Squares**

$$\text{SSTO} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- Variation that is still unexplained: **Error Sum of Squares**

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

- Variation that is explained: **Regression** (or Model) **Sum of Squares**

$$\text{SSR=SSTO-SSE} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$$

ANOVA Summary Table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	$p-1$	SSR	$MSR = SSR/(p-1)$	$F = \frac{MSR}{MSE}$	p-value
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTO$			

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

Proportion of variation in the response variable that is explained by the explanatory variables

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}}$$

Hypothesis Testing

- Overall F test for all the explanatory variables at once,
- T-tests for each regression coefficient: Controlling for all the others, does that explanatory variable matter?
- Test a collection of explanatory variables controlling for another collection,
- Most general: Testing whether sets of linear combinations of regression coefficients differ from specified constants.

Controlling for mother's education and father's education, are (any of) total family income, assessed value of home and total market value of all vehicles owned by the family related to High School GPA?

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5$$

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

(A false promise because of measurement error in education)

Full vs. Reduced Model

- You have 2 sets of variables, A and B
- Want to test B controlling for A
- Fit a model with both A and B: Call it the **Full Model**
- Fit a model with just A: Call it the **Reduced Model**

$$R_F^2 \geq R_R^2$$

When you add explanatory variables, R^2 can only go up

- By how much? Basis of F test.
- Same as testing H_0: All betas in set B (there are *d* of them) equal zero
- General H_0: **Lβ** = **h** (**L** is dxp, row rank d)

$$F = \frac{(SSR_F - SSR_R)/d}{MSE_F}$$

$$= \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{d\,MSE_F}$$

When you add explanatory variables to a model (with observational data)

- Statistical significance can appear when it was not present originally
- Statistical significance that was originally present can disappear
- Even the signs of the b coefficients can change, reversing the interpretation of how their variables are related to the dependent variable.
- Another version of Simpson's paradox

Copyright Information