A bit of regression: Quick and very applied[1]

STA312 Fall 2022

[1] See last slide for copyright information.

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \epsilon_i$$

- There are k explanatory variables.
- For each combination of explanatory variables, the conditional distribution of the response variable Y_i is normal, with constant variance σ^2.
- The conditional population mean of Y_i depends on the x values, as follows:

$$E(Y_i | x_{i,1}, \ldots x_{i,k}) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$$

“Control” means hold constant

- Regression model with four explanatory variables.
- Hold x_1, x_2 and x_4 constant at some fixed values.

$$\begin{aligned} E(Y|x_1, x_2, x_3, x_4) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ &= (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4) + \beta_3 x_3 \end{aligned}$$

- The equation of a straight line with slope β_3.
- Values of x_1, x_2 and x_4 affect only the intercept.
- So β_3 is the rate at which $E(Y|\mathbf{x})$ changes as a function of x_3 with all other variables held constant at fixed levels.
- *According to the model.*

More vocabulary

$E(Y|x_1, x_2, x_3, x_4) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4) + \beta_3 x_3$

- If $\beta_3 > 0$, describe the relationship between x_3 and (expected) y as "positive," controlling for the other variables. If $\beta_3 < 0$, negative.
- Useful ways of saying "controlling for" or "holding constant" include
  - Allowing for
  - Correcting for
  - Taking into account

Partitioning Sums of Squares

$$\begin{array}{ccccc} SST & = & SSR & + & SSE \\ \sum_{i=1}^{n}(Y_i - \overline{Y})^2 & = & \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 & + & \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \end{array}$$

$$R^2 = \frac{SSR}{SST}$$

Categorical Explanatory Variables

Unordered categories

- $X = 1$ 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 $E(Y|X = 0) = \beta_0$
- And β_1 is the difference between means, the average treatment effect.

More than Two Categories

Suppose a study has 3 treatment conditions. For example Group 1 gets Drug 1, Group 2 gets Drug 2, and Group 3 gets a placebo, so that the explanatory variable is Group (taking values 1,2,3) and there is some response variable Y (maybe response to drug again).

Why is $E(Y|X = x) = \beta_0 + \beta_1 x$ (with x = Group) a silly model?

Indicator Dummy Variables

With intercept

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
- Fill in the table.

Drug	x_1	x_2	$E(Y\|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Answer

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Drug	x_1	x_2	$E(Y\|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
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

- With an intercept in the model, need $p-1$ indicators to represent a categorical explanatory variable with p categories.
- If you use p dummy variables and an intercept, trouble.
- Regression coefficients are differences from the category that has no indicator.
- Call this the *reference category*.

What null hypotheses would you test?

Drug	x_1	x_2	$E(Y\|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

- Is the effect of Drug A different from the placebo? $H_0 : \beta_1 = 0$
- Is Drug A better than the placebo? $H_0 : \beta_1 = 0$
- Did Drug B work? $H_0 : \beta_2 = 0$
- Did experimental treatment have an effect? $H_0 : \beta_1 = \beta_2 = 0$
- Is there a difference between the effects of Drug A and Drug B? $H_0 : \beta_1 = \beta_2$

Now add a quantitative explanatory variable (covariate)

Covariates often come first in the regression equation

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $x_3 =$ Age
- $E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_1) + \beta_3 x_3$
B	0	1	$\mu_2 = (\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	0	0	$\mu_3 = \beta_0 + \beta_3 x_3$

Parallel regression lines.

More comments

Drug	x_1	x_2	$E(Y\|x_1,x_2,x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_1) + \beta_3 x_3$
B	0	1	$\mu_2 = (\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	0	0	$\mu_3 = \quad \beta_0 \quad + \beta_3 x_3$

- If more than one covariate, parallel regression planes.
- Non-parallel (interaction) is testable.
- "Controlling" interpretation holds.
- In an experimental study, quantitative covariates are usually just observed.
- Could age be related to drug if there is random assignment to drug?
- Good covariates reduce MSE, make testing of categorical variables more sensitive.

Hypothesis Testing

Standard tests when errors are normal

- Overall F-test for all the explanatory variables at once $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- t-tests for each regression coefficient: Controlling for all the others, does that explanatory variable matter? $H_0 : \beta_j = 0$
- Test a collection of explanatory variables controlling for another collection $H_0 : \beta_2 = \beta_3 = \beta_5 = 0$
- Example: 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?
- Most general: Testing whether sets of linear combinations of regression coefficients differ from specified constants. $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

t-tests and confidence intervals

$\widehat{\beta}_j$ are normally distributed.

$$t = \frac{\widehat{\beta}_j - \beta_j}{se_{\widehat{\beta}_j}} \sim t(n-k-1)$$

Full versus Restricted Model

Restricted by H_0

- 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*, or the *Unrestricted Model*.
- Fit a model with just A: Call it the *Restricted Model*. $R_F^2 \geq R_R^2$.
- The F-test is a likelihood ratio test (exact).

When you add the r additional explanatory variables in set B, R^2 can only go up

By how much? Basis of the F test.

$$
\begin{aligned}
F &= \frac{(R_F^2 - R_R^2)/r}{(1 - R_F^2)/(n-k-1)} \\
&= \frac{(SSR_F - SSR_R)/r}{MSE_F} \overset{H_0}{\sim} F(r, n-k-1)
\end{aligned}
$$

General Linear Test of $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$

$\mathbf{L}$ is $r \times p$, rows linearly independent

$$
\begin{aligned}
F &= \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})^\top (\mathbf{L}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{L}^\top)^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{r\, MSE_F} \\
&\stackrel{H_0}{\sim} F(r, n-k-1)
\end{aligned}
$$

Equal to full-restricted formula.

Copyright Information