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* independent variables
- For each *combination* of IVs, the conditional distribution of the dependent 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|**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

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

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

Statistics b estimate parameters beta

$$
\begin{aligned}
E[Y|\boldsymbol{X} = \boldsymbol{x}] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\
\widehat{Y} &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4
\end{aligned}
$$

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

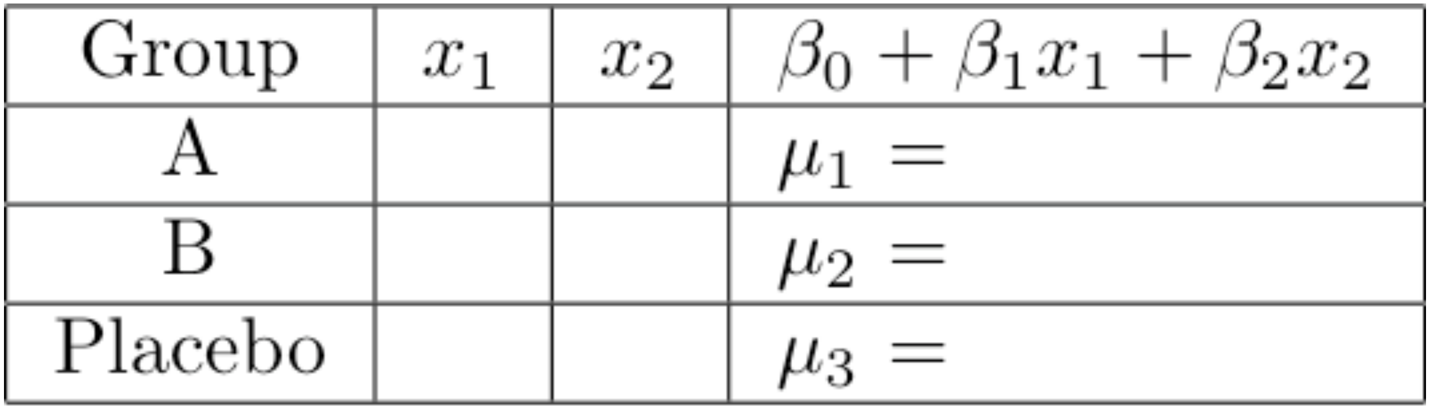
$$\widehat{Y} = b_0 = \overline{Y}_0$$

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

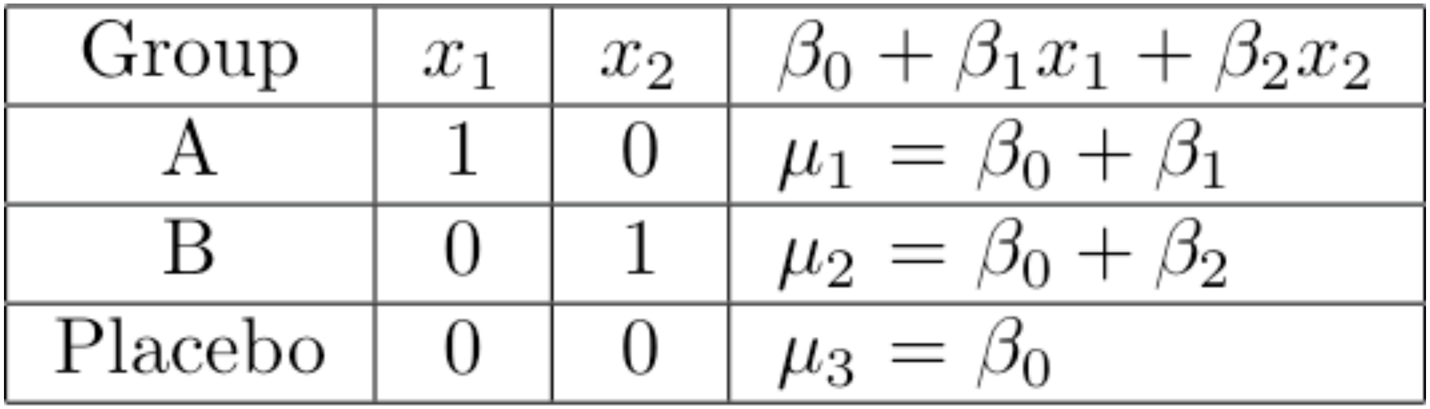
- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Fill in the table



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	$\mu_3 = \beta_0$

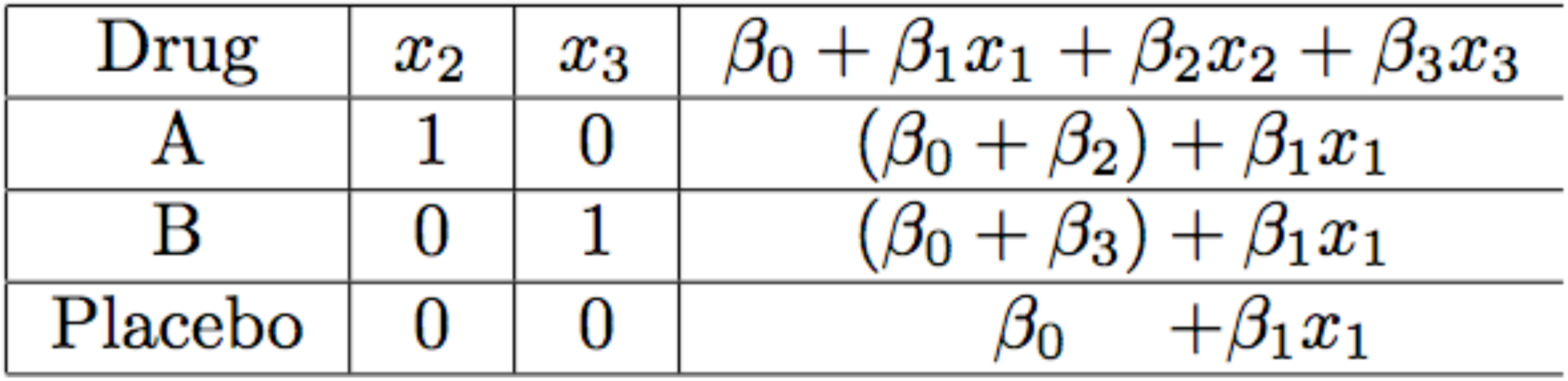
Regression coefficients are *contrasts* with the category that has no indicator – the *reference* category

Indicator dummy variable coding with intercept

- Need p-1 indicators to represent a categorical IV 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
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

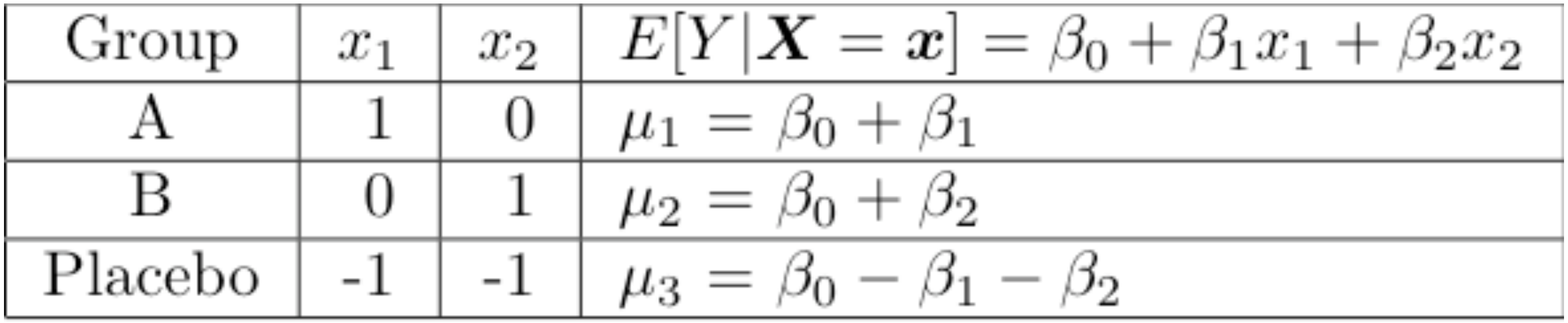


Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad +\beta_1 x_1$

Parallel slopes, ANCOVA

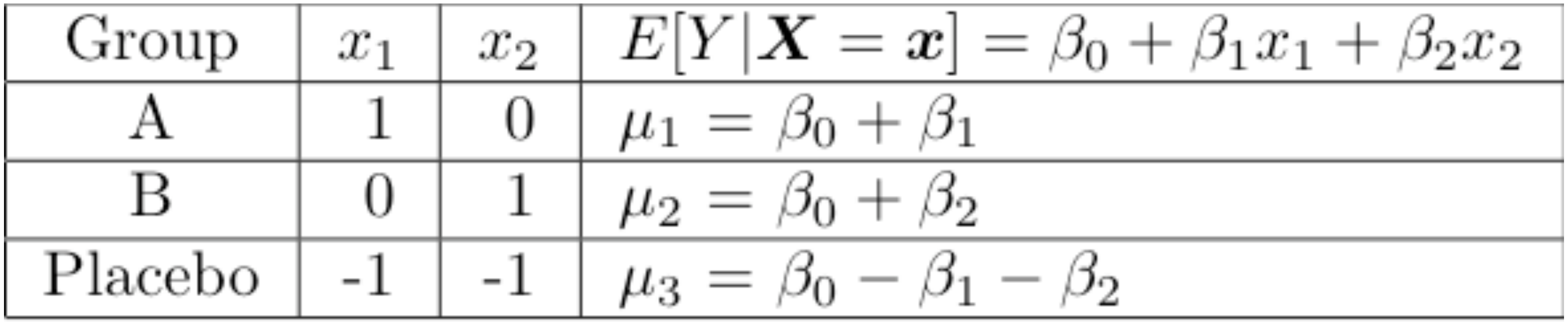
Effect coding

- *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



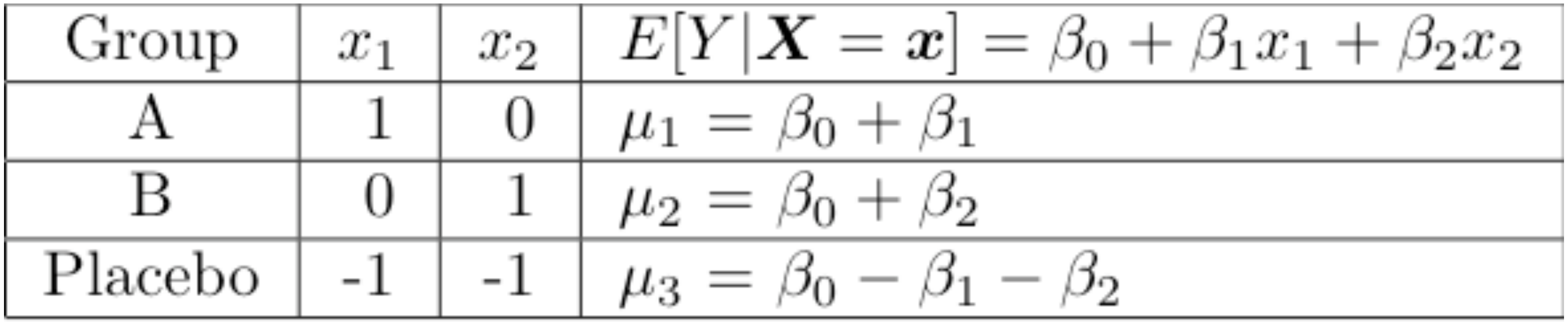
Group	x_1	x_2	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

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

The grand mean

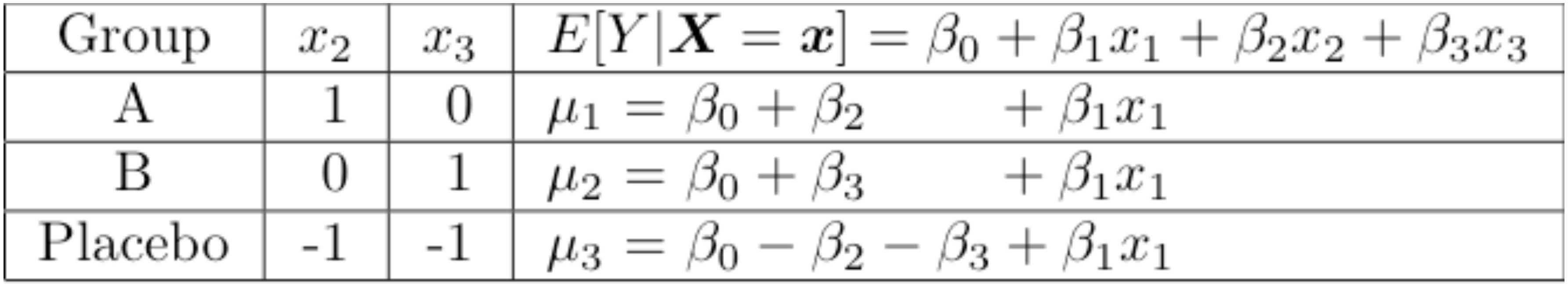
With effect coding

- 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	$\mu_1 = \beta_0 + \beta_2 \qquad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \qquad + \beta_1 x_1$
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 IV is unrelated to the DV, controlling for the covariate(s).

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

Interaction terms correspond to products of dummy variables.