Multivariable Linear Regression

Part 1 - Categorical Predictors

Ryan Miller

Grinnell College
Statistics

Introduction

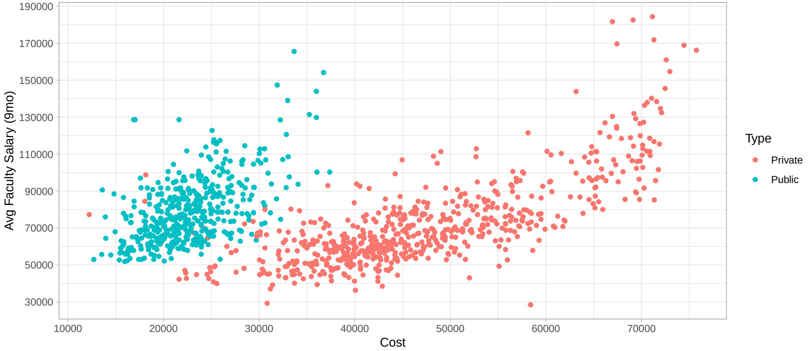
The theoretical framework of regression allows us to relate several explanatory variables with a response variable simultaneously:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

We'll begin our study of these models with the simplest case: one quantitative and one categorical explanatory variable.

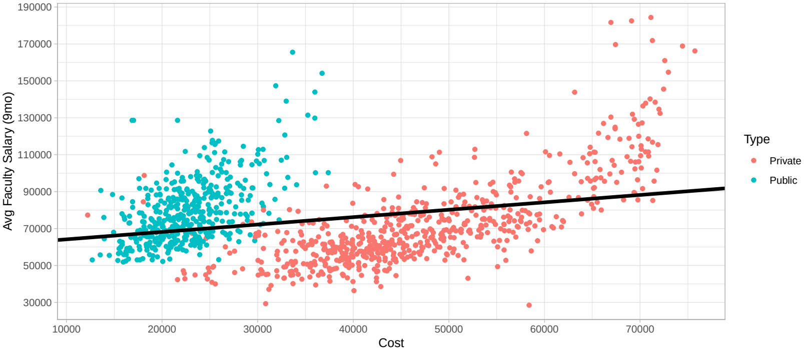
Application

Shown below are 3 variables describing primarily undergraduate colleges:



Model #1 - `Avg_Fac_Salary ~ Cost`

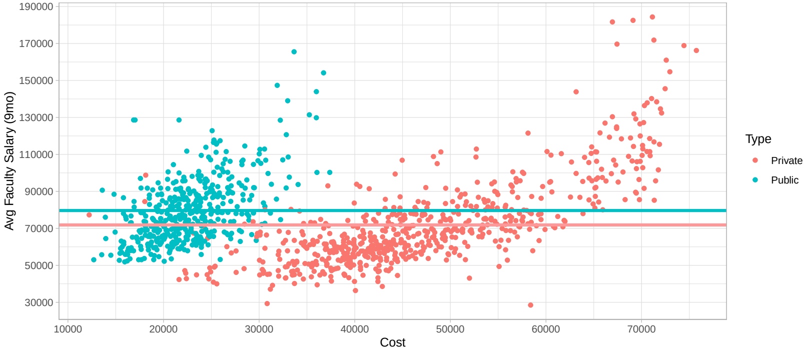
Fitted model: $\hat{y} = 60150 + 0.4 * \text{Cost}$



Grinnell College
Statistics

Model #2 - Avg_Fac_Salary ~ Type

Fitted model: $\hat{y} = 71836 + 7800 * (\text{Type} = \text{`Public'})$



Grinnell College
Statistics

One-hot Encoding

Regression equations involve numeric inputs, so the categorical variable "Type" is mapped to a **dummy variable**: `Type = 'Public'` using **one-hot encoding**:

College	Type
Grinnell College	"Private"
University of Iowa	"Public"
University of Minnesota	"Public"
Middlebury College	"Private"
Carlton College	"Private"

College	Type = "Public"
Grinnell College	0
University of Iowa	1
University of Minnesota	1
Middlebury College	0
Carlton College	0

One category defines the **reference group**, private colleges in this example.

One-hot Encoding

One-hot encoding can handle categorical variables with arbitrarily many categories:

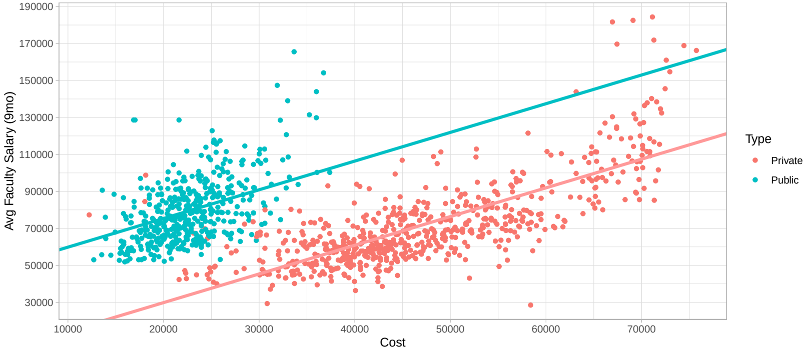
College	State
Grinnell College	IA
University of Iowa	IA
University of Minnesota	MN
Middlebury College	VT
Carlton College	MN

College	State = "MN"	State = "VT"
Grinnell College	0	0
University of Iowa	0	0
University of Minnesota	1	0
Middlebury College	0	1
Carlton College	1	0

Note that the category "IA" defines the reference group in this example.

Model #3 - `Avg_Fac_Salary ~ Cost + Type`

Fitted model: $\hat{y} = -1229 + 45529 * (\text{Type} = \text{`Public'}) + 1.55 * \text{Cost}$



Grinnell College
Statistics

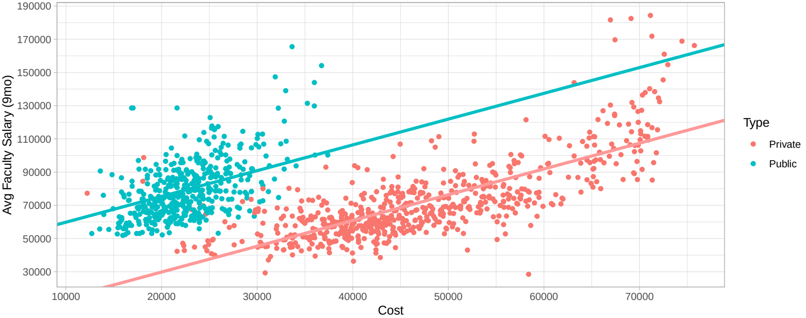
Adjusted Effects (example #1)

Compare the coefficient of `Cost` in Model #1 and Model #3:

- Model #1: $\hat{y} = 60150 + 0.4 * \text{Cost}$
  - Averaging across both types (private and public), each \$1 increase in a college's cost is expected to increase its average faculty salary by \$0.4
- Model #3: $\hat{y} = -1229 + 45529 * (\text{Type} = \text{`Public'}) + 1.55 * \text{Cost}$
  - Within colleges of the same type, each \$1 increase in cost is expected to increase average faculty salary by \$1.55

Adjusted Effects (example #1)

The slope of Model #3 is much steeper because this model has the flexibility to find a separate intercept for private and public colleges:



This allows the model to account for the large number of private colleges with comparatively high costs and low salaries.

Grinnell College
Statistics

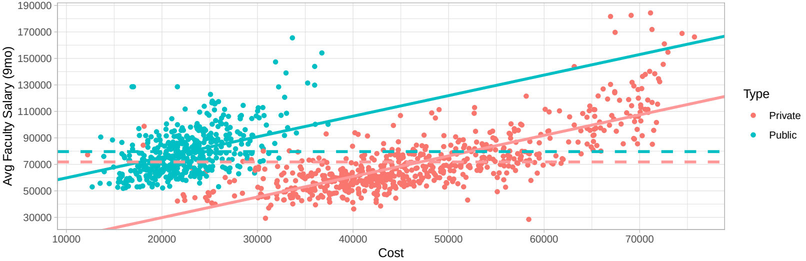
Adjusted Effects (example #2)

Compare the coefficient of `(Type = 'Public')` in Model #2 and Model #3:

- Model #2: $\hat{y} = 71836 + 7800 * (\text{Type} = \text{`Public'})$
  - Averaging across colleges of all costs, faculty salaries are expected to be \$7800 higher for public colleges than private colleges
- Model #3: $\hat{y} = -1229 + 45529 * (\text{Type} = \text{`Public'}) + 1.55 * \text{Cost}$
  - Within colleges of the *same cost*, faculty salaries are on average \$45529 higher for public colleges than private colleges

Adjusted Effects (example #2)

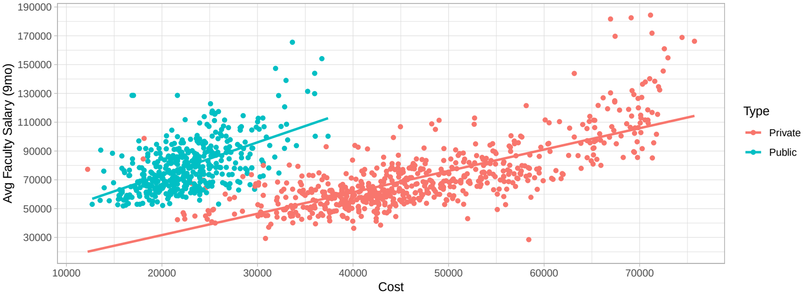
- The suspiciously large effect from Model #3 illustrates a common misuse of regression
  - Because there's very little overlap in the distributions of cost for private and public colleges, we may want to rely upon Model #2



If you're giving career advice, which model offers a more useful portrayal of the role of `Type`?

Stratification

Model #3 forced the same slope (in the `Cost` dimension) for both private and public colleges. We could allow for different slope using *stratification*:



- Among private colleges: $\hat{y} = 1952.14 + 1.485 * \text{Cost}$
- Among public colleges: $\hat{y} = 28025.86 + 2.267 * \text{Cost}$

Grinnell College
Statistics

Conclusion

- Categorical variables are represented in regression models via one-hot encoding
  - This designates one category as the reference group, and the estimated coefficients of dummy variables describe expected differences from this group
- Regression can be used to estimate *adjusted effects*, such as the effect of cost within colleges of the same type
  - We should be mindful of whether an adjusted effect or a marginal effect is more relevant to our specific analysis