Multivariable Regression

Part 1 - one quantitative and one categorical predictor

Ryan Miller

Grinnell College
Statistics

Introduction

Previously we learned about **simple linear regression** as a way of relating a quantitative explanatory variable with a quantitative outcome:

$$\text{Theoretical Model:} \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\text{Fitted Model:} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

In **multivariable linear regression**, we propose that the outcome is determined by a linear combination of explanatory variables and random error:

$$\text{Theoretical Model:} \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$
$$\text{Fitted Model:} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Introduction

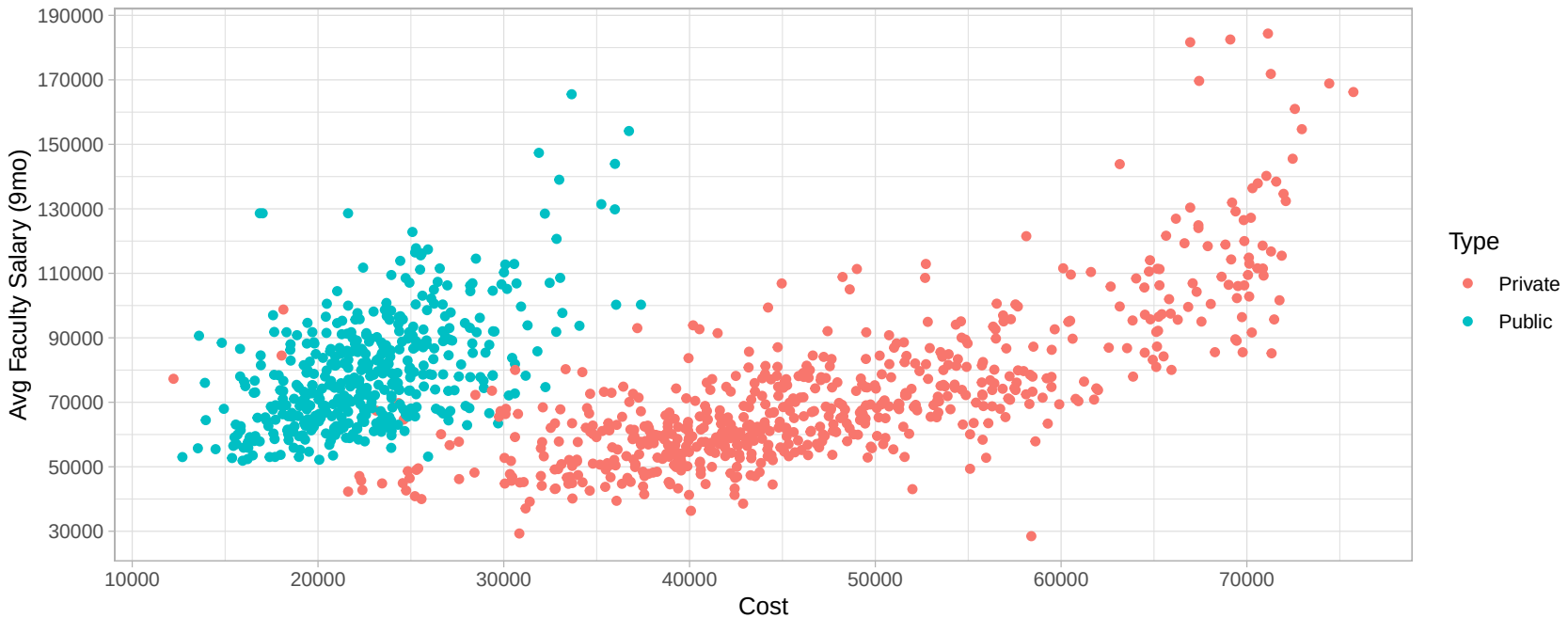
Multivaraible regression allows us to find *adjusted effects* that account for *simultaneous changes* that occur in other explanatory variables:

- **Marginal effects**: how X and Y are related without considering any other factors (ie: the relationship averaged across all other potential predictors)
- **Adjusted effects**: how X and Y are related after adjusting for imbalances in another factor (ie: the relationship when another predictor is held constant)

So, unless the explanatory variables we're considering are independent, we expect simple linear regression and multivariable regression to show different relationships.

Example

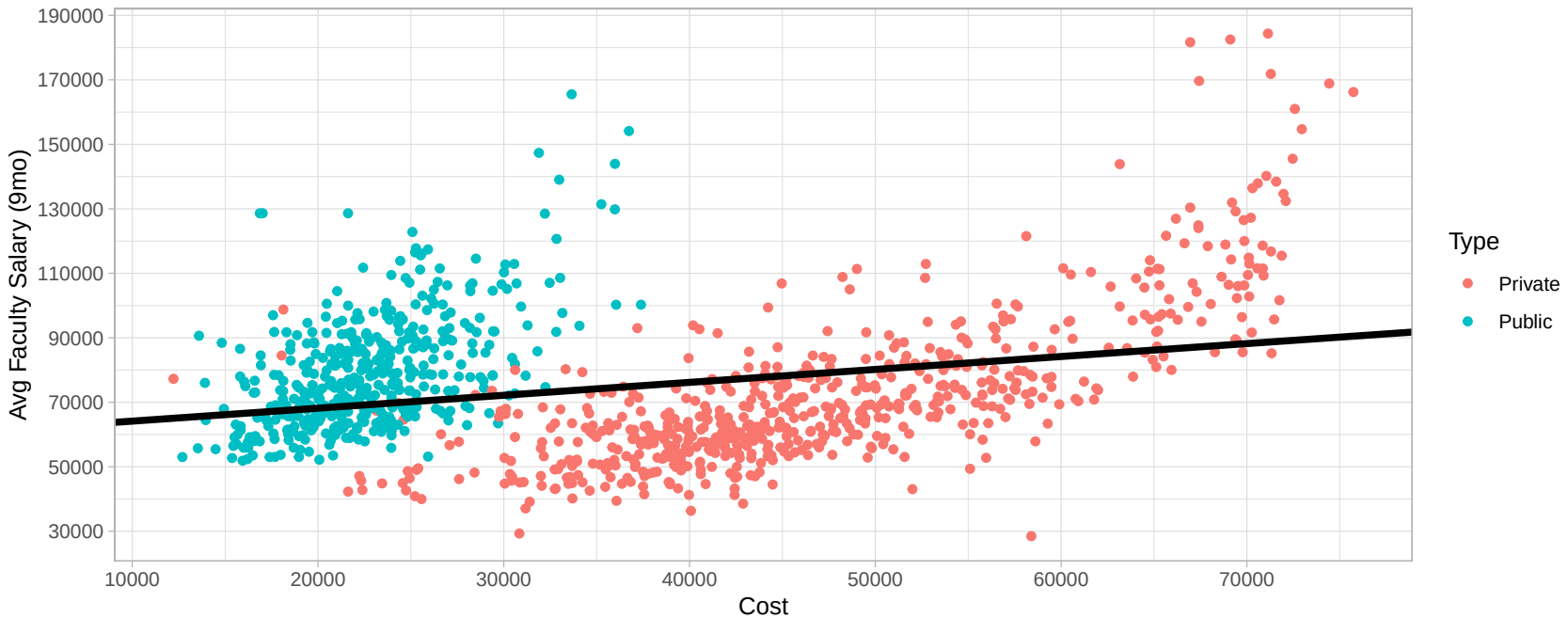
Consider the relationship between two explanatory variables, type of institution and cost, and a college's average 9-month faculty salary:



1. Do both explanatory variables appear related to the outcome?
2. Do both explanatory variables appear related to each other?

Example (cont.)

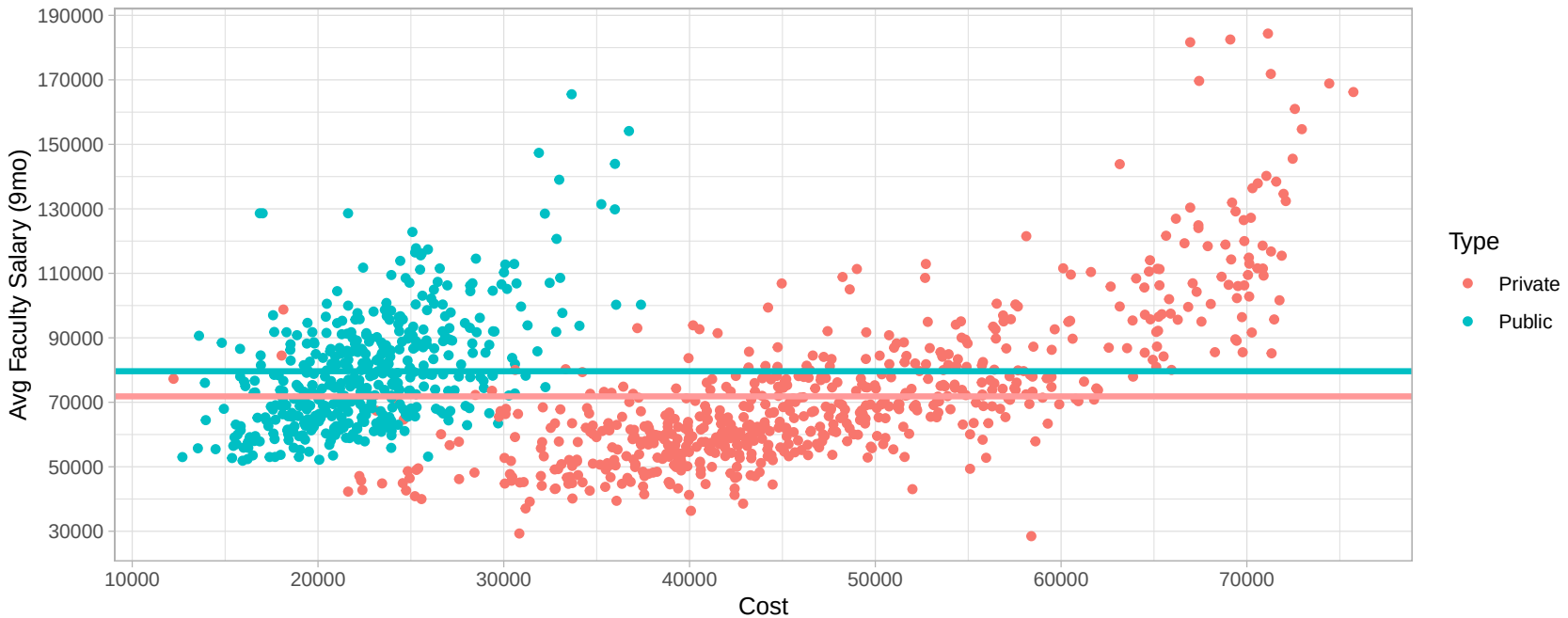
If we only consider cost (simple linear regression) we get the fitted equation $\widehat{\text{Salary}} = 60150 + 0.4 \cdot \text{Cost}$:



Is the slope of this line a *marginal* or *adjusted* effect? Does the line seem like a good model for the data?

Example

If we only consider `Type` (one-way ANOVA), we get:
$\widehat{\text{Salary}} = 71836 + 7800 \cdot (\text{Type} = \text{“Public”})$:



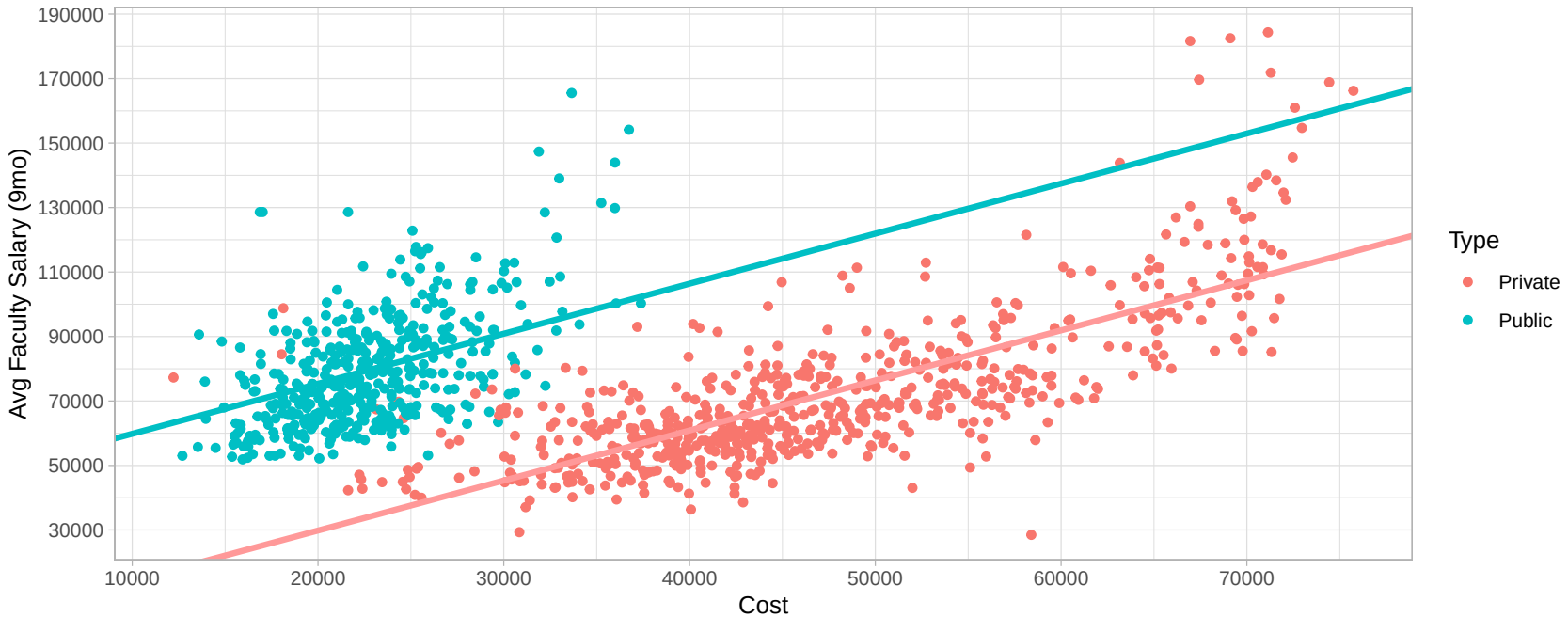
Is the gap between these lines a *marginal* or *adjusted* effect? Does it seem to adequately represent the trends in the data?

Challenges

- Public and private schools tend to have very different costs with minimal overlap in their cost distributions
- Public schools tend to cost less, and higher costs are associated with higher salaries, so a simple comparison of private and public colleges is partly reflecting differences in cost rather than differences in the type of institution
- Similarly, a simple comparison of cost and salary is reflecting differences in type of institution rather than differences in cost

Adjusted Effects

If we fit a model using *both* type and cost, we get:
$\widehat{\text{Salary}} = -1229 + 45529 \cdot (\text{Type} = \text{“Public”}) + 1.55 \cdot \text{Cost}$:



How do we interpret the coefficient of `Type='Public'`? How does this differ from the previous model (without `Cost`)?

Model Coefficient Interpretations

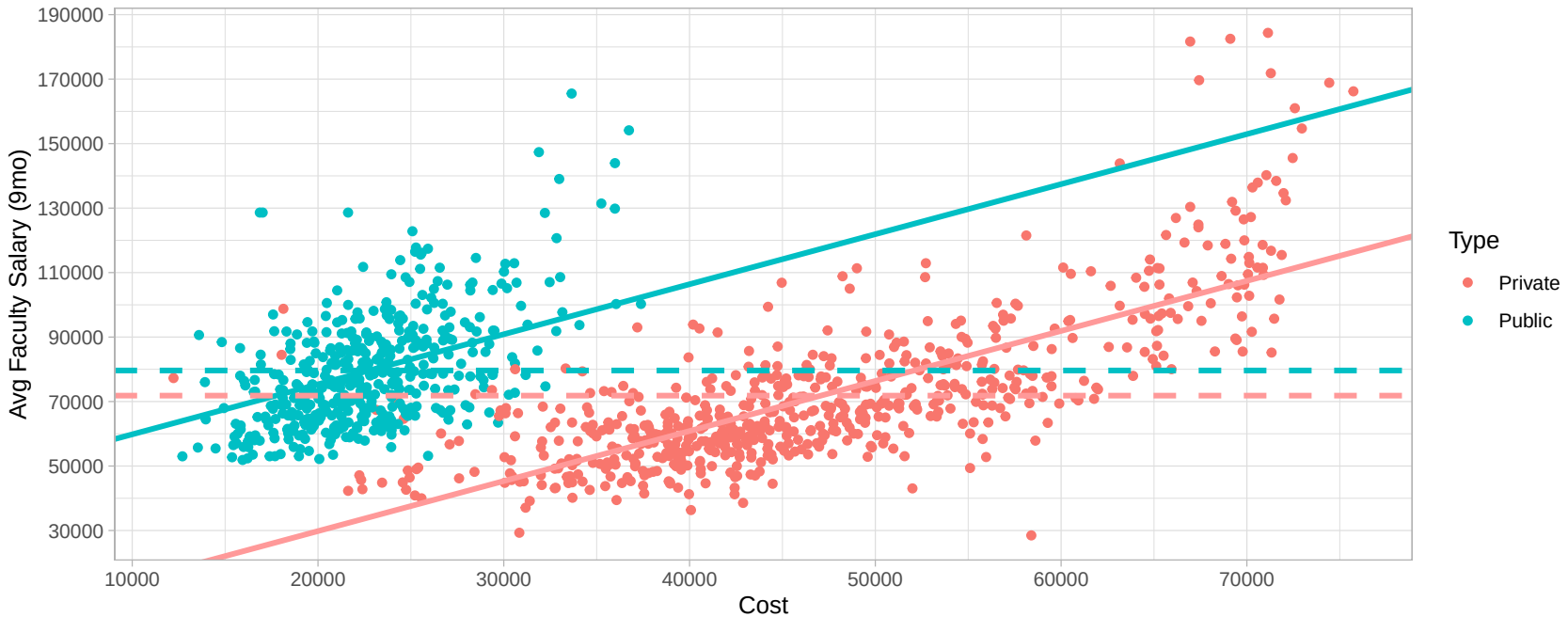
- Cost only model: $\widehat{\text{Salary}} = 60150 + 0.4 * \text{Cost}$
  - Averaging across both types (private and public), each \$1 increase in cost is expected to increase average faculty salary by \$0.4
  - This is the **marginal effect** of cost
- Cost and type model:
  $\widehat{\text{Salary}} = -1229 + 45529 * (\text{Type} = \text{'Public'}) + 1.55 * \text{Cost}$
  - For *colleges of the same type*, each \$1 increase in cost is expected to increase average faculty salary by \$1.55
  - This is the **adjusted effect** of cost, adjusting for differences in type

Model Coefficient Interpretations (cont.)

- Type only model: $\widehat{\text{Salary}} = 71836 + 7800 \cdot (\text{Type} = \text{`Public'})$
  - Averaging across colleges of all costs, faculty salaries are expected to be \$7800 higher for public colleges than private colleges
  - This is the **marginal effect** of type
- Cost and type model:
  $\widehat{\text{Salary}} = -1229 + 45529 * (\text{Type} = \text{`Public'}) + 1.55 \cdot \text{Cost}$
  - For colleges of the *same cost*, faculty salaries are on average \$45529 higher for public colleges than private colleges
  - This is the **adjusted effect** of type, adjusting for differences in cost

Is an Adjusted Effect Appropriate?

Suppose you're giving career advice to a recent PhD graduate. Is the marginal or adjusted effect of `Type` more useful in helping them decide which colleges they should apply to?



Is an Adjusted Effect Appropriate? (cont.)

- The adjusted effect seemingly suggests that public colleges pay better and should be favored in this individual's job search
  - However, it isn't useful to adjust for cost in this situation, as the PhD graduate isn't concerned with differences in cost between private and public colleges
- The marginal effect is more useful here, as it reflects the expected difference in salary, averaged across all private and public schools (irrespective of their cost)

When to use Adjusted Effects vs. Marginal Effects?

Example #1:

- Outcome: SAT score
- Predictor: Attending a private vs. a public high school
  - Marginal effect: private school students on average score 80 points higher
  - Adjusted effect: adjusting for household income, private school students are expected to score only 10 points higher

Questions:

1. Which estimate is more relevant to *parents*? Which is more relevant to *policymakers*?
2. Why are the marginal and adjusted effects so different?

When to use Adjusted Effects vs. Marginal Effects?

Example #2:

- Outcome: Sale price of a home
- Predictor: Size of the home
  - Marginal effect: each additional 100 sq ft increases the expected price by $35,000
  - Adjusted effect: for homes in the same neighborhood, each 100 sq ft only increases the expected price by $12,000

Questions:

1. Which estimate is more relevant to *home buyer* comparing two houses in the same neighborhood?
2. Why are the marginal and adjusted effects so different?

Conclusion

We've now seen two types of effects:

- **Marginal effects**: how X and Y are related without considering any other factors (ie: the relationship averaged across all other potential predictors)
- **Adjusted effects**: how X and Y are related after adjusting for imbalances in another factor (ie: the relationship when another predictor is held constant)

Regression can be used to estimate *adjusted effects*. In some situations the adjusted effect is more meaningful than the marginal effect, but this isn't universally true, variables should be adjusted for only if doing so helps answer the question of interest.