MS&E 226: Fundamentals of Data Science

Lecture 14: Regression analysis of experiments

Ramesh Johari

$\widehat{\text{ATE}}$ from a randomized experiment

Suppose we run a completely randomized design (CRD) experiment, and compute the following estimate of the average treatment effect ATE:

$$\widehat{\text{ATE}} = \frac{1}{n_1} \sum_{i:W_i=1} Y_i(1) - \frac{1}{n_0} \sum_{i:W_i=0} Y_i(0).$$

In the last lecture, we noted that $\widehat{\text{ATE}}$ is an unbiased estimator for ATE.

Regression analysis

Notice that $\widehat{\text{ATE}}$ is the *difference* between two groups: those where $W = 1$, and those where $W = 0$.

We saw earlier in the course that we can also compute such differences using OLS linear regression, with Y as the outcome and W as the sole feature.

Regression analysis

In particular, suppose we use OLS to fit the following model:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_W W.$$

In a randomized experiment, $W_i = 0$ or $W_i = 1$ for every observation.

Therefore:

- $\hat{\beta}_0$ is the average outcome in the control group.
- $\hat{\beta}_0 + \hat{\beta}_W$ is the average outcome in the treatment group.
- So $\hat{\beta}_W = \widehat{\text{ATE}}$!

An example in R

I constructed a synthetic "experiment" where $n_1 = n_0 = 1000$, and:

$$Y_i = 1 + W_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$. (Question: what is the true ATE?)

```
lm(formula = Y ~ 1 + W, data = df)
...
Coefficients:
            Estimate Std. Error
(Intercept) 10.01399    0.03157
W            0.96854    0.04464
...
```

The estimated standard error on $\hat{\beta}_1 = \widehat{\text{ATE}}$ is the same as the estimated standard error we would obtain using the direct formula $\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_0^2/n_0}$ in the last lecture.

Using covariates

Adding covariates to the regression

In a randomized experiment, we often observe additional *covariates* $\mathbf{X}_i$ for each individual *i before* treatment assignment (e.g., age, gender, baseline measurements).

Adding these covariates to the regression can improve the precision of the estimate of ATE, i.e., lower the variance (and thus the standard error).

We say that we are *controlling* for these covariates (the added covariates are referred to as *regression controls*).

Important note: The regression controls must be collected *pre-treatment*, otherwise you can introduce confounding!

Important warning

The covariates $\mathbf{X}_i$ must be observed *pre-treatment*!

Why? If X is affected by the treatment, then controlling for X can introduce bias.

Example: In a drug trial, if X is a post-treatment measurement of blood pressure after taking the drug, then some of the difference between treatment and control outcomes may be mistakenly attributed to variation in the (post-treatment) X rather than the actual treatment W.

Always ensure covariates are measured *before* treatment assignment.

OLS regression with covariates

We can adapt the preceding regression by adding covariates:

$$Y \sim \hat{\beta}_W W + (\hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p).$$

As long as we randomized assignment, in this model we can *still* interpret $\hat{\beta}_W$ as an estimate of the ATE.

This is because randomization ensures that W is uncorrelated with both X and the potential outcomes.

Note: This interpretation is valid *even* if the true population model is *not* linear in the covariates $\vec{X}$; if you choose to interpret the coefficients on the covariates as well, then as usual assumptions (A1)-(A3) are needed for valid inference.

Regression controls: Interpretation

For now suppose there is a single continuous covariate X. In the regression $Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X$, we can interpret the model as follows:

For an individual with covariate X:

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_X X$ (baseline outcome under control)
- $Y(1) \approx \hat{\beta}_0 + \hat{\beta}_W + \hat{\beta}_X X$ (outcome under treatment)
- Individual treatment effect $\approx \hat{\beta}_W$

In this model, X helps explain variation in the *baseline* outcome $Y(0)$, but the treatment effect is assumed constant across individuals.

Regression controls: An example

I created a synthetic experiment where $n_0 = n_1 = 1000$.

For each individual i, $X_i \sim \mathcal{N}(0, 1)$ is a pre-existing covariate, and W_i is the treatment indicator.

I constructed Y_i as:

$$Y_i = 10 + W_i + 2X_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$.

In this example:

- The true ATE is 1.0 –it does not vary depending on X.
- However, some of the variation in Y_i is explained by X as well.

Controlling for observables: An example

Suppose we regress Y on the treatment indicator W alone:

```
lm(formula = Y ~ 1 + W, data = df)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.9606     0.0719  138.53   <2e-16 ***
W             1.1041     0.1017   10.86   <2e-16 ***
...
```

The standard error on W is 0.1017.

Controlling for observables: An example

Now suppose we include the covariate X in the regression:

```
lm(formula = Y ~ 1 + W + X, data = df)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.95913    0.03218  309.50   <2e-16 ***
W            1.02709    0.04552   22.57   <2e-16 ***
X            2.00650    0.02246   89.32   <2e-16 ***
...
```

Notice that the standard error is *much smaller* on the coefficient of W: it dropped from 0.1017 to 0.04552.

This is because X explains much of the variation in Y, reducing residual variance.

(Note: (A1)-(A3) hold in this synthetic example; if you are not sure if (A3) holds, you should use the bootstrap or robust standard errors.)

Why does controlling help?

By including X in the regression, we account for variation in outcomes that is due to X, not due to treatment.

This reduces the *residual variance* $\hat{\sigma}^2$ (i.e., the sample variance of the residuals), which in turn reduces the standard error of $\hat{\beta}_W$.

Even though randomization ensures unbiasedness, controlling for pre-treatment covariates improves *precision*.

Imperfect randomization

Controlling for observed covariates can potentially offer another benefit:

If the randomization was less than perfect, regression controls can mitigate selection bias.

How this works:

- Suppose, e.g., individuals with higher X were more likely to receive treatment in the experiment.
- Ignoring this creates *omitted variable bias*: part of the variation in Y is explained by X, not by treatment.
- Controlling for X can remove this bias.

But a warning: Of course, if randomization was imperfect, then we also can't be sure that just the covariates we observe are sufficient to remove selection bias!

We will have more to say about the use of observed covariates to remove confounding in the next lecture.

Interacted effects regression

Interacted effects regression

Controlling for covariates shows that we can reduce variance compared to the simple difference-in-means estimator $\widehat{\text{ATE}}$. Can we do better?

For simplicity, let's continue to assume a single continuous covariate. Consider the *interacted effects (IE) regression*:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X + \hat{\beta}_{WX} W \times (X - \bar{X}).$$

In this regression, we add an *interaction term* between treatment and *centered* covariate $(X - \bar{X})$.

This is a more flexible specification: It also allows for the possibility that now the treatment W also changes the *slope* on the covariate X.

Why center the covariates?

In the IE regression:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X + \hat{\beta}_{WX} W \times (X - \bar{X}),$$

why do we use $(X - \bar{X})$ instead of just X in the interaction?

Centering ensures that $\hat{\beta}_W$ directly estimates the ATE.

Without centering:

- $\hat{\beta}_W$ would estimate the treatment effect when $X = 0$.
- This may not be meaningful (e.g., if X is age, $X = 0$ is not in our population).

With centering:

- Informally, $\hat{\beta}_W$ estimates the treatment effect at the average value of X.
- Since the ATE can be written as: $\mathbb{E}_X[\mathbb{E}_Y[Y(1) - Y(0)|X]]$, this intuition can be used to show that $\hat{\beta}_W$ converges to ATE as the sample size grows (i.e., that it is *consistent*).

Efficiency of IE regression

In fact, it can be shown that if the IE regression is used to estimate ATE, then it has several desirable properties as the sample size increases; asymptotically:

- It is *consistent* for ATE, even if the true model is not linear in X, or if errors are heteroskedastic (i.e., (A1)-(A3) need not hold);
- It has variance *at least as small* as the variance of the estimate of ATE from the simple regression $Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X$, and can be *strictly smaller*.

In practice, therefore, the implication is that when analyzing data from randomized experiments you should *always* use the IE regression instead of simple regression.

SEs for IE regression

When using the IE regression, it is important to note that *the estimated standard error in the standard OLS regression table will be incorrect*.

This is because we have centered the covariate X_i using a sample mean that is computed using the *same* data that we use to fit the regression; this favorably reduces the variation in the coefficient estimate $\hat{\beta}_W$.

As a result, in general when using IE regression, you should use either the bootstrap or robust standard errors. (See R Markdown notebook for details.)

Robust standard errors for IE regression in R

We briefly discussed robust standard errors in Lecture 9.

The following R code shows how to implement the IE regression with robust standard errors using the `sandwich` package:

```
library(sandwich)
library(lmtest)

# Fit IE regression with centered covariate
model_ie <- lm(Y ~ W + X + W:I(X - mean(X)), data = df)

# Display coefficients with robust SEs (HC3 type)
coeftest(model_ie, vcov = vcovHC(model_ie, type = "HC3"))
```

Robust standard errors for IE regression in R

We obtain the following regression output in the previous example; note that the SE on $\hat{\beta}_W$, 0.045558, is similar to the SE on $\hat{\beta}_W$ in the simple regression, 0.04552.

This is because in this case, there is no heterogeneity in the treatment effect with varying X.

```
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  9.959148   0.032070 310.5441   <2e-16 ***
W            1.027095   0.045558  22.5449   <2e-16 ***
X            1.985550   0.031946  62.1535   <2e-16 ***
W:X_centered 0.041811   0.045280   0.9234   0.3559
```

Heterogeneous treatment effects

Beyond the average treatment effect

So far we've focused on estimating the *average* treatment effect (ATE) across the entire population.

But what if run a randomized experiment, and the treatment effect varies across individuals?

These are called *heterogeneous* treatment effects (HTE).

Formalizing heterogeneous treatment effects

The *conditional average treatment effect* (CATE) given covariates $\vec{X}$ is:

$$\text{CATE}(\vec{X}) = \mathbb{E}[Y(1) - Y(0)|\vec{X}].$$

This measures the average treatment effect among individuals with covariate value $\vec{X}$.

In particular, it allows the possibility that the causal effect differs across individuals.

Estimating CATE via IE regression

Again suppose only a single continuous covariate X. Recall the IE regression we introduced earlier:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X + \hat{\beta}_{WX} W \times (X - \bar{X}).$$

This model actually directly allows us to estimate the CATE, *if* we believe the true population model is linear in X (i.e., (A1) holds).

Specifically, for an individual with covariate X:

- $Y(0) \approx \hat{\beta}_0 + \hat{\beta}_X X$; and
- $Y(1) \approx \hat{\beta}_0 + \hat{\beta}_W + (\hat{\beta}_X + \hat{\beta}_{WX})X - \hat{\beta}_{WX}\bar{X}$; so
- $\widehat{\text{CATE}}(X) = \hat{\beta}_W + \hat{\beta}_{WX}(X - \bar{X})$

Note that $\hat{\beta}_W$ is the estimated CATE at $X = \bar{X}$, i.e., an estimate of the ATE.

Interpreting the interaction coefficient

In the model $Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X + \hat{\beta}_{WX} W \times X$:

- $\hat{\beta}_W$ is the estimated treatment effect when $X = 0$.
- $\hat{\beta}_{WX}$ measures how the treatment effect *changes* as X increases by one unit.
- If $\hat{\beta}_{WX} > 0$, the treatment effect is larger for individuals with higher X.
- If $\hat{\beta}_{WX} < 0$, the treatment effect is smaller for individuals with higher X.
- If $\hat{\beta}_{WX} \approx 0$, there is little evidence of heterogeneous treatment effects.

Interactions: Example with no HTE

Recall our earlier example where the true model is:

$$Y_i = 10 + W_i + 2X_i + \epsilon_i.$$

There is no heterogeneity here: the treatment effect is 1.0 for everyone.

What happens if we estimate a model with interactions?
When we estimated a model with interactions, we obtained $\hat{\beta}_{WX} = 0.041811$ and a p-value of 0.3559, which is as expected.

Interactions: Example with HTE

Now suppose we change the model so the treatment effect *does* vary with X:

$$Y_i = 10 + (1 + 2X_i)W_i + 2X_i + \epsilon_i.$$

Now the treatment effect is $1 + 2X_i$, which increases with X.

What happens when we estimate a model with interactions on this data?

Interactions: Example with HTE

Here is what we obtain if we estimate the IE regression with robust standard errors:

```
...
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  9.999627   0.031007 322.495 < 2.2e-16 ***
W            1.043747   0.044590  23.408 < 2.2e-16 ***
X            1.997663   0.031575  63.268 < 2.2e-16 ***
W:X_centered 1.976603   0.044862  44.060 < 2.2e-16 ***
...
```

Now the interaction coefficient is large (1.98) and significant.

This correctly captures that the treatment effect increases with X:

$$\widehat{\text{CATE}}(X) = 1.04 + 1.98X.$$

Testing for heterogeneous treatment effects

In the IE regression:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X + \hat{\beta}_{WX} W \times (X - \bar{X}),$$

how do we test whether treatment effects are heterogeneous?

Run a hypothesis test on the interaction coefficient:

- H_0: $\beta_{WX} = 0$ (homogeneous treatment effects)
- H_1: $\beta_{WX} \neq 0$ (heterogeneous treatment effects)

This is a standard hypothesis test on $\hat{\beta}_{WX}$, using the reported t-statistic.

Benefits of modeling interactions

Including interaction terms in regression analysis of experiments can provide several benefits:

- *Efficiency*: Always at least as efficient as simple regression.
- *Robustness*: Consistent for ATE even under model misspecification.
- *Heterogeneity*: Understand *which* individuals or subgroups benefit most (or least) from treatment.
- *Targeting*: Identify individuals who would benefit most from treatment for policy or business decisions.

More flexible models

Limitations of OLS regression approaches

A key limitation of the OLS regression approachees is that the set of pre-treatment features must be “small” compared to the sample size.

Otherwise, as we have seen previously in this course, OLS can *overfit* to the features.

In the worst case, if there are *more* features than samples, OLS will not even have a unique solution.

This is odd, since we should be able to use any additional features to obtain an even more precise estimate of ATE...

From OLS to machine learning

Consider the IE regression, but with uncentered X:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_X X + \hat{\beta}_{WX} WX.$$

This can be equivalently written as two *separate* linear regressions:

- For control group ($W = 0$): $Y \sim \hat{\mu}_0(X)$, where $\hat{\mu}_0(X) = \hat{\beta}_0 + \hat{\beta}_X X$
- For treatment group ($W = 1$): $Y \sim \hat{\mu}_1(X)$, where $\hat{\mu}_1(X) = (\hat{\beta}_0 + \hat{\beta}_W) + (\hat{\beta}_X + \hat{\beta}_{WX})X$

The ATE estimate is then: $\hat{\beta}_W = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$. (This is the same estimate we would obtain from the centered IE regression.)

Natural question: If we're fitting two separate models anyway, why restrict ourselves to linear regression? Why not use more flexible machine learning models?

The T-learner

Now suppose there are many features. The preceding discussion motivates the following approach, called the *T-learner* ("T" for "two"):

1. Fit a regression model $\hat{\mu}_1(\vec{X})$ to predict Y using only the *treatment* group data ($W = 1$).
2. Fit a regression model $\hat{\mu}_0(\vec{X})$ to predict Y using only the *control* group data ($W = 0$).

The T-learner (continued)

3. For each individual i, estimate their CATE:

$$\widehat{\text{CATE}}(\mathbf{X}_i) = \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i).$$

4. Estimate the average treatment effect by averaging CATE over all individuals:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\text{CATE}}(\mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^{n} [\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)].$$

The models $\hat{\mu}_1$ and $\hat{\mu}_0$ can be *any* regression method: linear regression, random forests, neural networks, etc.

Note: Computing standard errors is not possible in closed form for such methods; usually resampling approaches like the bootstrap are used.

Looking ahead: Causal inference from observational data

The T-learner is one example of a modern, machine-learning based approach to causal inference.

In the next lecture, we will discuss how such methods can be useful even in settings where there is possible selection bias, *if* we believe that observed features account for the selection bias.