Regression Review

Review/Background - Introduction to Causal Inference

Kevin Li

Linear Regression

For explanatory variables $X_1, X_2, \dots, X_p$, an outcome variable Y, for units $i = 1, 2, \dots n$:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}_{\mathbb{E}[Y_i|X_i]} + \varepsilon_i$$

- β_0 is the intercept - the expected value of Y when all $X = 0$.
- $\beta_j \in \{\beta_1, \dots, \beta_p\}$ are coefficients. For every one unit increase in X_j, there is an expected β_j unit change in Y, holding all other explanatory variables constant.

Fitted values (predictions, estimated best-fit line):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$

Vector Form

We can also write the same regression as:

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

Where vector $\mathbf{X}_i$ (usually lower-case in statistics, but uppercase in causal inference) and vector $\boldsymbol{\beta}$:

$$\mathbf{X}_i = \begin{pmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

You can multiply this out to check it equals initial equation.

Matrix Form

We can also write the same regression as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Where vector $\mathbf{y}$, matrix $\mathbf{X}$, vector $\boldsymbol{\beta}$, and vector $\boldsymbol{\varepsilon}$ are:

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} \; \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ 1 & X_{31} & X_{32} & \dots & X_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \; \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

(You can multiply out the vectors to see it equals the original regression for each individual i in the sample)

OLS Estimator

We want to estimate $\hat{\boldsymbol{\beta}}$. How to do this? - by minimising the error (more specifically, minimise the sum of squared errors for any chosen values of $\hat{\boldsymbol{\beta}}$).

- What is error? It is the difference between the actual Y_i value in our data, and the estimated $\hat{Y}_i$ by our fitted values.

$$\text{minimise:} \quad SSE(\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$$

1. Multiple out the equation above.
2. Take the derivative in respect to vector $\hat{\boldsymbol{\beta}}$.
3. Solve for $\hat{\boldsymbol{\beta}}$ to get the OLS estimates that minimise the SSE:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Estimators and Uncertainty

We usually only have a sample of individuals from the population, when we estimate $\hat{\beta}$.

- What if we had a different sample with different individuals? We would get different $\hat{\beta}$ estimate.

So we have to account for **sampling uncertainty**. This is done with a sampling distribution.

- Sampling distribution is basically - imagine you take a sample, estimate $\hat{\beta}$. then take another hypothetical sample, and another. The distribution of estimates is the sampling distribution.

Standard deviation of a sampling distribution is the **standard error** of the estimate.

Unbiasedness and Variance

Our goal in statistics is to use our sample estimator $\hat{\beta}$ to estimate the true population β (we do not know the value of).

If the expected value of the sampling distribution $\mathbb{E}[\hat{\beta}]$ is equal to the true population value β (that we do not know), then the estimator is considered **unbiased**.

- That means on average, any estimate we run will have an expected value of the true population value.
- Thus, we want an unbiased estimator, since any specific estimate with any specific sample will be on average, correct.

We generally prefer unbiased estimators that have low variance.

- Low variance means estimates are less spread apart. If our estimator is unbiased, and the variance is low, that means any individual estimate is close to the true population value.

Gauss-Markov: Unbiasedness

Gauss-Markov theorem (at least part of it) states that OLS is an unbiased estimator of the true $\hat{\beta}$ under the following conditions.

1. **Linearity in parameters**: The true relationship between X and Y can be represented by some form of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
2. **Random Sampling**: Random sample from a population.
3. **No Perfect Multicollineraity**: No 100% (exact) linear correlations between explanatory variables.
4. **Strict Exogeneity**: Formally defined as $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$. This implies that $Cov(\varepsilon, X_j) = 0$ for any explanatory variable X_j.

Violations to exogeneity are often caused by omitted confounders. Thus, ommitted confounders = biased (bad) estimates (important!)

Variance (Heteroscedasticity)

In causal inference, we almost always assume heteroscedasticity. The variance of the OLS estimator under heteroscedasticity:

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \underbrace{\begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \dots & \ddots & 0 \\ 0 & 0 & \dots & \sigma_i^2 \end{pmatrix}}_{\varepsilon \text{ variance matrix}} \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$$

- Where σ_i^2 is the variance of ε_i term.

For clustered standard errors (and other standard errors), we simply modify the variance matrix used.

Standard Errors (Heteroscedasticity)

We know the variance of OLS (last slide). But there is an issue - we do not know the value of σ_i^2. Thus, we will estimate it with the residuals squared $\hat{\varepsilon}_i^2$ - the difference between our sample Y_i and estimated $\hat{Y}_i$ squared.

$$\widehat{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \underbrace{\begin{pmatrix} \hat{\varepsilon}_1^2 & 0 & ... & 0 \\ 0 & \hat{\varepsilon}_2^2 & ... & 0 \\ \vdots & ... & \ddots & 0 \\ 0 & 0 & ... & \hat{\varepsilon}_i^2 \end{pmatrix}}_{\varepsilon \text{ variance matrix}} \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$$

The estimate of the standard error of $\hat{\beta}$, $\widehat{se}(\hat{\beta})$, is just the square root of our estimated variance.

Hypothesis Testing

There is uncertainty in our estimates of $\hat{\beta}$. How can we be confident the true β is different from 0 with just our estimate?

- We run a hypothesis test - Null hypothesis $H_0 : \beta = 0$.

We calculate a t-test statistic (why t not z? because we estimate σ_i^2 with $\hat{\varepsilon}_i^2$, which introduces uncertainty).

$$t = \frac{\hat{\beta}}{\widehat{se}(\hat{\beta})}$$

Then, we use the t-test statistic and a t-distribution with $n - p - 1$ degrees of freedom to calculate the **p-value**.

- P-value is the probability the null hypothesis is true ($\beta = 0$), given our estimate $\hat{\beta}$. If this probability is lower than 5%, we consider the null hypothesis unlikely, and reject the null and accept the alternate hypothesis.