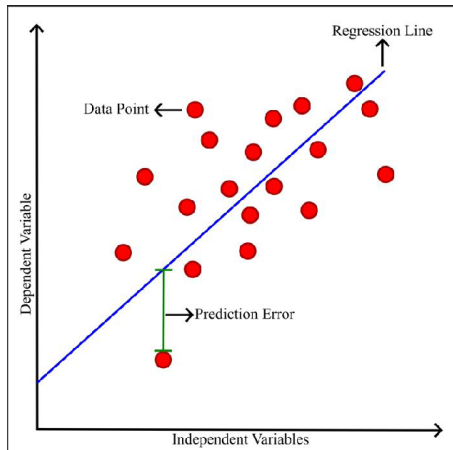# Regression Review
Review/Background - Introduction to Causal Inference

Kevin Li

Linear Regression

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{best-fit line}} + \varepsilon_i$$

β_0 is the **intercept** of the line: expected value of Y when $X = 0$.

β_1 is the slope/**coefficient**: for a one unit increase in X, there is an expected β_1 change in Y.

ε_i is the **error**: points will not be exactly on the line, and this represents how far individual i's Y value is from the line.

Multiple Linear Regression

We can have multiple independent variables $X_1, X_2, \ldots, X_p$:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}_{\text{regression best-fit hyperplane (p-dimensional)}} + \varepsilon_i$$

▶ For any $\beta_j = \beta_1, \ldots, \beta_p$: for a one unit increase in X_j, there is an expected β_j change in Y, *holding all other independent variables constant*.

We can also write the same regression in matrix form in two ways:

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i \iff \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
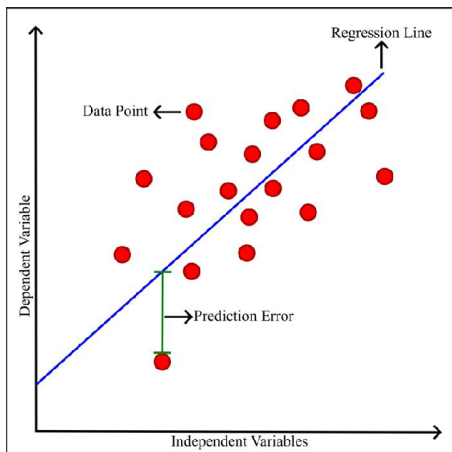
Fitted Values

The previous two slides introduced the **population model** - i.e. the *true* relationship between independent and dependent variables in the *population*.

▶ We often do not observe the population, so we have to use our **sample** to create estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\varepsilon}_i$.

▶ Once we estimate with our sample, we can generate our "fitted" model (fitted values), which is our sample-estimated best-fit prediction line.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip}$$

Estimating $\hat{\beta}_0, \ldots, \hat{\beta}_p$

$$SSE = \sum(Y_i - \hat{Y}_i)^2$$

We want to find values for $\hat{\beta}_0, \ldots, \hat{\beta}_p$ that minimise the **sum of squared** (prediction) **errors**.

Makes sense: we want our fitted line to have as little prediction error as possible to capture the true relationships.

Why squared? We don't care about positive or negative errors, just the size of errors.

OLS Estimator

We want to find $\hat{\beta}_0, \dots, \hat{\beta}_p$ values that minimise the sum of squared errors:

$$\text{minimise:} \quad \text{SSE}(\hat{\beta}) = \sum (Y_i - \hat{Y}_i)^2 = \underbrace{(\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})}_{\text{linear algebra form}}$$

Do some linear algebra, take the derivative in respect to $\hat{\beta}$, and you will get the solutions:

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

Vector $\hat{\beta}$ will contain estimated values for $\hat{\beta}_0, \dots, \hat{\beta}_p$.

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**.

▶ **Sampling distribution**: 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

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 with any random sample 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.

▶ 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 $\mathbf{y} = \mathbf{X}\beta + \varepsilon$.

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 $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$. This implies that $\text{Cov}(\varepsilon, X_j) = 0$ for any explanatory variable X_j. Violations to exogeneity are often caused by omitted confounders.

Variance (Heteroscedasticity)

The estimated variance of the OLS estimator (under heteroscedasticity):

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

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

▶ Sometimes, we will use other standard errors, like clustered standard errors. These have a different ε variance matrix.

Hypothesis Testing

There is uncertainty in our estimates of any coefficient $\hat{\beta}_j$. How do we know the true β is **not** 0 with just our sample estimate (our null: $H_0 : \beta_j = 0$)? We calculate a t-test statistic:

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

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 is true ($\beta_j = 0$), given our estimate $\hat{\beta}_j$. If this is lower than 5%, the null is unlikely, so we reject the null and conclude there is significant relationship between X_j and Y.