MATH423 Notes

Chany Ahn

Contents

1 Introduction

Regression is about quantitative, predictive relationships.

- **Prediction**: make statements about unobserved data (future data).
- **Inference**: make statements about the unknown data generating mechanism.

2 Optimal Linear Predictor

Predicting a random variable from its distribution. Given a single random variable Y, where the distribution is known, the optimal way of predicting Y is:

$$m = \mathbb{E}(Y) \tag{1}$$

This is optimal because it *minimizes the mean square error.*

The criteria to measure goodness of a prediction: let m be the prediction of Y:

- $Y - m$: prediction error
- $(Y - m)^2$: squared prediction error.

Use *mean squared error*:

$$MSE(m) = \mathbb{E}[(Y - m)^2] \tag{2}$$

as the criterion to measure the prediction performance. Since we want to minimize the MSE, we need to solve for the following relation:

$$\min_m MSE(m) = \min_m \mathbb{E}[(Y - m)^2] \tag{3}$$

Let m^* represent the value of m that minimizes $MSE(m)$, then we get

$$m^* = \arg\min_m \mathbb{E}[(Y - m)^2] \tag{4}$$

where m^* is the minimizer. If x^* is the minimizer of $f(x)$, then we need to solve for:

$$\left.\frac{df(x)}{dx}\right|_{x=x^*} = 0 \tag{5}$$

If we take $\mathbb{E}[|Y - m|]$, which is the mean absolute deviation (MAD), then

$$m^* = median(Y) \tag{6}$$

The advantage of this is it is more stable against outliers. The disadvantage is that it is computationally inefficient.

3 Predicting One Random Variable from Other Variables

Consider two types of variables:

- "Output": Y – measure of interest.
  - Outcome

  - Response variable
  - Dependent variable
- "Input": X – variable(s) that correlated to the "output".
  - Features
  - Covariates or factors
  - Predictors
  - Explanatory variables

To predict Y using the information of X

$$Y \approx m(X) \tag{7}$$

The prediction function $m(X)$ does not necessarily imply causality. Thus, we use the *regression function*:

$$m(x) = \mathbb{E}[Y|X = x] \tag{8}$$

This is the "optimal" prediction of Y, where we use MSE as the criterion to measure the prediction accuracy. Thus, if we use the MSE criterion, we get:

$$m^*(\cdot) = \arg\min_{m(\cdot)} \mathbb{E}_{X,Y}[(Y - m(X))^2] \tag{9}$$

where we can show that

$$m^*(x) = \mathbb{E}_{Y|X}[Y|X = x] = \mathbb{E}[Y|X = x] \tag{10}$$

Note on conditional expectation:

$$\mathbb{E}_{X,Y}[g(X,Y)] = \mathbb{E}_X\{\mathbb{E}_{Y|X}[g(Y,X)]\}$$

Proof. See hand-out. □

Note:

- If we use $\mathbb{E}[|Y - m(X)|]$ as the criterion, then the optimal function is $m^*(x) = median(Y|X = x)$.
- At no point was the marginal distribution of (X, Y) specified.
- At no point did we assume the fluctuation of Y was Gaussian or symmetric.
- At no time did we assume that X came before Y in time or that X causes Y.

4 Nearest-Neighbor Regression

We estimate the regression function through the equation

$$m^*(x) = \mathbb{E}(Y|X = x) \tag{11}$$

using the n observations, (x_i, y_i), where $x_i \in \mathbb{R}^p$. One way to estimate $m^*(x)$ is:

$$\hat{m}(x) = average(\{y_i : x_i = x\}) \tag{12}$$

The above can be difficult because realistically there will be few points at exactly x_i. So, we cannot estimate $m^*(x)$ directly. So, we can just relax the definition:

$$\begin{aligned}\hat{m}(x) &= average(\{y_i : x_i \text{ equal to or very close to } x\}) \\ &= \frac{1}{k} \sum_{x_i \in N_k(x)} y_i\end{aligned}$$

$N_k(x)$ is the neighborhood of x defined by k closest points to x_i in the training sample. There are different distance metrics to compute the nearest neighbours, usually the Euclidean distance).

4.1 Theoretical Guarantee of KNN

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. KNN is good for small p (i.e. $p \leq 4$ and large N). Under mild regularity conditions on joint probability distributions of $P(X, Y)$, one can show that as $N, k \to \infty$, $\frac{k}{N} \to 0$, then

$$\hat{m}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} \to \mathbb{E}(Y|X = x) \tag{13}$$

5 Curse of Dimensionality

For small p, we can always find a fairly large neighborhood of observations close to the target x and average them. However, for a large p, if we consider the KNN inputs, X, uniformly distributed in a p-dimensional unit hypercube, and we want to predict a target x, we must set up a neighborhood around x to capture a fraction R of the observation.

The expected edge length (radius):

$$L_p(R) = R^{\frac{1}{p}} \tag{14}$$

Example: If we let $p = 10$, to capture 1% or 10% of the data to form a local average, we must cover $L_{10}(0.01) = 0.63$ and $L_{10}(0.1) = 0.8$ of the range of each input. Such neighbourhoods are no longer "local" (there are a lot of good videos on this topic).

6 Optimal Linear Prediction

In general, $m(x) = \mathbb{E}(Y|X = x)$ might have a complicated form:

- KNN: $\mathbb{E}(Y|X = x) = ave\{y_i : x_i \in N_k(x)\}$
- Linear Regression: $\mathbb{E}(Y|X = x)x^\top\beta$
- Additive Model: $\mathbb{E}(Y|X = x) = f_1(x_1) + \ldots f_p(x_p)$, $x \in \mathbb{R}^p$
- Decision Tree: $\mathbb{E}(Y|X = x) = T(x)$
- Random Forest/Gradient Boosting: $\mathbb{E}(Y|X = x) = \sum_{m=1}^{M} \beta_m T_m(x)$
- Deep Learning: $\mathbb{E}(Y|X = x) = (\ldots \sigma(W_2\sigma(W_1 x))$
- SVM: $\mathbb{E}(Y|X = x) = \sum_i \alpha_i k(x, x_i)$, where $k(x, x_i)$ is the kernel.

To simplify $m(X)$, we restrict

$$m(X) = \mathbb{E}(Y|X) = \beta_0 + \beta_1 X \tag{15}$$

NoteL we only have one predictor $X \in \mathbb{R}$.

The question we want to ask is: *what are the optimal values of β_0 and β_1 that minimizes MSE?*

We ASSUME the distribution of (X, Y) is known. Thus,

$$(\beta_0^*, \beta_1^*) = \arg\min_{(\beta_0, \beta_1)} \mathbb{E}_{X,Y}[(Y - m(x))^2] \tag{16}$$

$$= \arg\min_{(\beta_0, \beta_1)} \mathbb{E}_{X,Y}[(Y - (\beta_0 + \beta_1 X))^2] \tag{17}$$

We can come up with analytical solutions for both:

$$\beta_0^* = \mathbb{E}(Y) - \beta_1^* \mathbb{E}(X) \tag{18}$$

$$\beta_1^* = \frac{Cov(X,Y)}{Var(X)} \tag{19}$$

Proof. See in handout.

□

$m^*(X) = \beta_0^* + \beta_1^* X$ is called the "true" regression line (the optimal linear prediction function). If the assumption is not satisfied, then we may get a linear prediction for data without a linear trend.

Note: The optimal regression line goes through $(\mathbb{E}(X), \mathbb{E}(Y))$.

$$m^*(X) = \beta_0^* + \beta_1^* X \tag{20}$$

$$= \mathbb{E}(Y) - \beta_1^* \mathbb{E}(X) + \beta_1^* x \text{ (from (18))} \tag{21}$$

If we plugin $x = \mathbb{E}(X)$, we get $m^*(\mathbb{E}(X)) = \mathbb{E}(Y)$. But, $\mathbb{E}(Y|X = x)$ does not necessarily go through $(\mathbb{E}(X), \mathbb{E}(Y))$.

If X and Y are "centered", i.e. $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, the optimal regression line goes through $(0, 0)$ since $\beta_0^* = 0$.

Things to notice for β_1^*:

- $Cov(X, Y)$ incerases, then β_1^* increase.
- $Var(X)$ (or the spread of your input data increases), β_1^* decreases.

The optimal slope β_1^* doesn't change if we use $Y - c$ and $X - c'$. This is NOT true for the intercept β_0^*.

Note, non-linear patters cannot be appropriately modelled by this predictor. Imagine a true regression function:

$$\mathbb{E}(Y|X = x) = e^x \tag{22}$$

If we do a Taylor expansion at $X = x_0$, then we get:

$$e^x = e^{x_0} + \frac{de^x}{dx}\bigg|_{x=x_0} (x - x_0) + \frac{1}{2}\frac{d^2e^x}{dx^2}\bigg|_{x=x_0} (x - x_0)^2 + \dots \tag{23}$$

$$= e^{x_0} + e^{x_0}(x - x_0) + \frac{1}{2}e^{x_0}(x - x_0)^2 + \dots \tag{24}$$

The quadratic term must be dominated by the linear term for it to have a significant influence.

$$\frac{1}{2}e^{x_0}|x - x_0|^2 << e^{x_0}|x - x_0| \tag{25}$$

$$\frac{|x - x_0|^2}{|x - x_0|} << \frac{2e^{x_0}}{e^{x_0}} \tag{26}$$

$$|x - x_0| << 2 \tag{27}$$

7 Estimating Optimal Linear Prediction using Data (Plug-in Estimator)

To estimate the optimal linear prediction from n observations of data, $(x_1, y_1) \dots (x_n, y_n)$, we use the estimator:

$$\beta_1^* \approx \hat{\beta}_1 = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{28}$$

where $\overline{y}, \overline{x}$ are the sample means of the output and input, respectively.

$$\beta_0^* \approx \hat{\beta}_0 = \hat{\mathbb{E}}(Y) - \hat{\beta}_1 \hat{E}(X) = \overline{y} - \hat{\beta}_1 \overline{x} \tag{29}$$

Thus, the fitted regression line is

$$m^*(x) \approx \hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{30}$$

$$= \left[\overline{y} - \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \overline{x} \right] + \left[\frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \right] x \tag{31}$$

Proof. See handout.

□

The fitted regression line, $\hat{m}(x)$ goes through $(\overline{x}, \overline{y})$ (do the calculations, they're very simple).

If the data is centered, then the fitted line goes through $(0, 0)$ (do the calculations, they're very simple).

The slope does not change under the shift of the data ($y_i' = y_i - c$, $X_i' = x_i - c'$).

8 Simple Linear Regression

In order to do inference, we have to make assumptions and believe our data is generated from the SLR model.

Model Assumptions:

1. (Arbitrary input) The distribution X_1 is arbitrary (X_1 can be even nonrandom).
2. (Linear function and additive error)
   $$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
   We want to distinguish between deterministic ($Y = \beta_0 + \beta_1 X_1$) and non-deterministic functions ($Y = \beta_0 + \beta_1 X_1 + \varepsilon$).
   The noise variable ε may represent:
   - Other factors not considered in the model (but associated with fluctuation in Y).
   - Measurement errors.
   - Combination of both.
3. (Zero mean and constant variance error): $\mathbb{E}(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 > 0$ (doesn't change with X_1).
   Note, if ε has nonzero mean ($\mathbb{E}(\varepsilon) = c$), we can find a random variable ε' with $\mathbb{E}(\varepsilon') = 0$ and $Var(\varepsilon') = \sigma^2$, s.t. $\varepsilon = \varepsilon' + c \Rightarrow \varepsilon' = \varepsilon - c$

The original model can be re-written as:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon && (32)\\ &= \beta_0 + \beta_1 X_1 \varepsilon' + c && (33)\\ &= (\beta_0 + c) + \beta_1 X_1 + \varepsilon' \quad (\beta_0 + c = \beta_0') && (34)\\ &= \beta_0' + \beta_1 X_1 + \varepsilon' && (35) \end{aligned}$$

4. (Independent error) $\varepsilon \perp\!\!\!\perp X_1$ ("statistically" independent). Thus, $f(\varepsilon, x_1) = f(\varepsilon)f(x_1)$ and $\mathbb{E}(\varepsilon|X_1) = \mathbb{E}(\varepsilon)$.

   We ignore the "intransitive" case. Most time, statistical independence bewteen two variables indicates that there is no (causal) relationship between two variables, but exceptions exist!

Intransitive Case: Let's consider X and Z which are two independent fair coins, and Y which are defined as such:

$$X = \begin{cases} 1 & \text{head} \\ 0 & \text{tail} \end{cases}, \; Z = \begin{cases} 1 & \text{head} \\ 0 & \text{tail} \end{cases}, \; Y = \begin{cases} 1 & \text{if } X = Z \\ 0 & \text{if } X \neq Z \end{cases}$$

X and Z are the causes of Y, but Y is "statistically" independent to X. To show this, we need to show that $P(Y = 1|X = 1) = P(Y = 1)$.

$$\begin{aligned} P(X = 1) = P(Z = 1) &= \frac{1}{2} \\ P(Y = 1|X = 1) = P(Y = 1|X = 0) &= \frac{1}{2} \\ P(Z = 1|X = 1) = P(Z = 1) &= \frac{1}{2} \\ P(Y = 1) &= P(Y = 1|X = 0)P(X = 0) + P(Y = 1|X = 1)P(X = 1) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} \\ &= P(Y = 1|X = 1) = P(Y = 1|X = 0) \end{aligned}$$

Therefore, Y and X are independent even if X causes Y.

The SLR assumptions actually implies that the true regression function is linear (i.e. true regression line).

$$\begin{aligned} \mathbb{E}(Y|X_1 = x_1) &= \mathbb{E}(\beta_0 + \beta_1 X_1 + \varepsilon|X_1 = x_1) && (36)\\ &= \mathbb{E}(\beta_0 + \beta_1 X_1|X_1 = x_1) + \mathbb{E}(\varepsilon|X_1 = x_1) && (37)\\ &= \beta_0 + \beta_1 x_1 && (38) \end{aligned}$$

where $\mathbb{E}(\varepsilon|X_1 = x_1) = \mathbb{E}(\varepsilon)$ since $\varepsilon \perp\!\!\!\perp X_1$. We note that Y has constant variance, thus

$$\begin{aligned} Var(Y|X_1 = x_1) &= Var(\beta_0 + \beta_1 X_1 + \varepsilon|X = x_1) && (39)\\ &= Var(\beta_0 + \beta_1 x_1) + Var(\varepsilon|X = x_1) && (40)\\ &= 0 + Var(\varepsilon) = \sigma^2 && (41) \end{aligned}$$

So, we summarize these results as:

$$Y|X_1 = x_1 \sim D(\beta_0 + \beta_1 x_1, \sigma^2) \tag{42}$$

Where D is a distribution, $\beta_0 + \beta_1 x_1$ is the expectation, and σ^2 is the variance.

We think of the SLR assumptions as a modeling decision that (we hope) will be useful, rather than the fact about the true underlying relationship.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \Rightarrow Y = f(X_1) + \varepsilon \tag{43}$$

where $\mathbb{E}(Y|X_1 = x_1) = f(x_1)$.

8.1 Parameter Interpretations

- Interpretation of β_0:

  It is the intercept, and the expected value of Y when $X_1 = 0$, i.e.

$$\mathbb{E}(Y|X_1 = 0) = \mathbb{E}(\beta_0 + \beta_1 \cdot 0|X_1 = 0) = \beta_0 \tag{44}$$

- Interpretation of β_1:

  It is the slope of the regression function.

$$\beta_1 = \mathbb{E}(Y|X_1 = x_1 + 1) - \mathbb{E}(Y|X_1 = x_1) \tag{45}$$
$$= \beta_0 + \beta_1(x_1 + 1) - (\beta_0 + \beta_1 x_1) \tag{46}$$

  If we select two sets of cases for (X_1, Y) distribution, where X_1 differs by 1, we expect the associated Y to differ by β_1 "on average" (not for the difference of Y).

  Note: Not to claim as the result of causality, but statistical association between X_1 and Y.

- Interpretation of σ^2:

  The variance of the noise around the regression line. It represents a typical distance of a point form the true regression line.

9 Model Setup for Multiple Data Points

We assume multiple data points, $(X_{11}, Y_1), (X_{21}, Y_2), \ldots, (X_{n1}, Y_n)$, are generated from the same model.

- $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$, for $i = 1, \ldots, n$
- $\mathbb{E}(\varepsilon_1) = 0$, for $i = 1, \ldots, n$
- $Var(\varepsilon_i) = \sigma^2$, for $i = 1, \ldots, n$
- $\varepsilon_i \perp\!\!\!\perp X_{i1}$, $\varepsilon_j \perp\!\!\!\perp X_{j1}$, and $\varepsilon_i \perp\!\!\!\perp \varepsilon_j$

The above assumptions implies the following results:

- $\mathbb{E}(Y_i|X_{i1} = x_{i1}) = \beta_0 + \beta_1 x_{i1}$, for $i = 1, \ldots, n$
- $Var(Y_i|X_{i1} = x_{i1}) = \sigma^2$, for $i = 1, \ldots, n$
- $Y_i \not\perp\!\!\!\perp Y_j$

$$Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 X_{i1} + \varepsilon_i, \beta_0 + \beta_1 X_{j1} + \varepsilon_j) \tag{47}$$
$$= Cov(\beta_1 X_{i1} + \varepsilon_i, \beta_1 X_{j1} + \varepsilon_j) \tag{48}$$
$$= \beta_1^2 Cov(X_{i1}, X_{j1}) + \beta_1 Cov(X_{i1}, \varepsilon_j) + \beta_1 Cov(\varepsilon_i, X_{j1}) + Cov(\varepsilon_i, \varepsilon_j) \tag{49}$$
$$= \beta_1^2 Cov(X_{i1}, X_{j1}) + 0 + 0 + 0 = \beta_1^2 Cov(X_{i1}, X_{j1}) \tag{50}$$

  Since X_{i1} and X_{j1} are not necessarily independent (unless we assume so).

- $Y_i \perp\!\!\!\perp Y_j | X_{i1}, X_{j1}$

$$Cov(Y_i \perp\!\!\!\perp Y_j | X_{i1} = x_{i1}, X_{j1} = x_{j1}) = Cov(\beta_1 X_{i1} + \varepsilon_i, \beta_1 X_{j1} + \varepsilon_j | X_{i1} = x_{i1}, X_{j1} = x_{j1}) \tag{51}$$
$$= Cov(\beta_1 x_{i1} + \varepsilon_i, \beta_1 x_{j1} + \varepsilon_j) \tag{52}$$
$$= Cov(c + \varepsilon_i, c' + \varepsilon_j) \tag{53}$$
$$= Cov(\varepsilon_i, \varepsilon_j) = 0 \tag{54}$$

10 Optimal Prediction for SLR

If we want to predict Y using X_1, we want the optimal prediction which minimizes MSE:

$$m^*(\cdot) = \arg\min_{m^*(\cdot)} \mathbb{E}_{X_1,Y}((Y - m(X_1))^2) \tag{55}$$

$$m^*(x_1) = \mathbb{E}(Y|X_1 = x_1) \tag{56}$$

In addition, if we assume (Y, X_1) follow the SLR model,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{57}$$

with $\mathbb{E}(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, $\varepsilon \perp\!\!\!\perp X_1$. Then the optimal prediction function has the linear form:

$$m^*(x_1) = \mathbb{E}(Y|X_1 = x_1) \tag{58}$$

$$= \mathbb{E}(\beta_0 + \beta_1 X_1 + \varepsilon|X_1 = x_1) \tag{59}$$

$$= \beta_0 + \beta_1 x_1 \tag{60}$$

11 Least Squares Estimators

Given n observations, $(x_{11}, y_1), \ldots, (x_{n1}, y_n)$, to estimate β_0, β_1 in SLR:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} MSE(\beta_0, \beta_1) \tag{61}$$

We estimate this by using the estimator $\widehat{MSE}(\beta_0, \beta_1)$.

- Residual: $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}$
- Residual Sum of Squares: $SS_{Res} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$

  An alternate formula is as follows: $SS_{Res} = \sum_{i=1}^n y_i^2 - n(\overline{y})^2 - \hat{\beta}_1 S_{XY} = SS_T - \hat{\beta}_1 S_{XY}$ where SS_T is the total sum of squares ($SS_T = \sum_{i=1}^n y_i^2 - n(\overline{y})^2 = \sum_{i=1}^n (y_i - \overline{y})^2$)
- (Empirical) MSE: $\widehat{MSE} = \frac{1}{n}\sum_{i=1}^n e_i^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2$

Least squares solve $\hat{\beta}_0, \hat{\beta}_1$ such that

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \widehat{MSE}(\beta_0, \beta_1) \tag{62}$$

$$= \arg\min_{\beta_0,\beta_1} \frac{1}{n}\sum_{i=1}^n e_i^2 \tag{63}$$

$$= \arg\min_{\beta_0,\beta_1} \frac{1}{n}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1})^2 \tag{64}$$

We can show that (ordinary least squares):

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_{i1} - \overline{x}_1)}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2} \tag{65}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}_1 \tag{66}$$

Note: least squares estimators are the same as the plugin estimators

Matrix Form: (MIGHT HAVE TO WATCH THIS PART OF THE LECTURE!)

In matrix format

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{67}$$

We must check to make sure that $\mathbf{X}^\top \mathbf{X}$ is invertible. If $rank(\mathbf{X}^\top \mathbf{X}) = colrank(X) = 2$, the it is invertible. This means, unless

$$\begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} \text{ is linearly dependent to } \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

then $\mathbf{X}^\top \mathbf{X}$ becomes not invertible (lol what).

To uniquely find a least square estimator, $n \geq 2$ in SLR (this makes a lot of sense. If it doesn't think what would happen if you had $n = 1$). Ideally:

$$(\beta_0^*, \beta_1^*) = \arg \min_{(\beta_0, \beta_1)} \mathbb{E}_{X,Y}[(Y - \beta_0 - \beta_1 X 1)^2] \approx (\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \tag{68}$$

Note: if the data points are all independent, the Law of Large Numbers tells us that:

$$\widehat{MSE}(\beta_0, \beta_1) \xrightarrow[n \to \infty]{P} MSE(\beta_0, \beta_1) \tag{69}$$

If the SLR assumptions are true, then

$$(\hat{\beta}_0, \hat{\beta}_1) \xrightarrow[n \to \infty]{P} (\beta_0^*, \beta_1^*) \tag{70}$$

For SLR, $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. The condition of expectation of Y given X_1 is $\mathbb{E}(Y|X_1) = \beta_0 + \beta_1 X_1$, which is also the true regression function. From these results, we see that not only β_0, β_1 is the parameters of the SLR, but also β_0, β_1 can minimize the expected MSE (i.e. $\beta_0 = \beta_0^*, \beta_1 = \beta_1^*$).

12 Statistical Properties of Least Squares Estimators

First, we must understand the difference between an estimate and an estimator:

- An *estimator* is a random variable. Take $X_1, \ldots, X_n \sim D(\mu)$, $\mu = \mathbb{E}[X]$, and X_i's are *iid*. $\hat{\theta}_n$ is an estimator of μ if X_i have not been observed:

$$\hat{\theta}_n = T(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{71}$$

- An *estimate* is the realized value of $\hat{\theta}_n$.

$$\hat{\theta}_n = T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{72}$$

Least Squares Estimators: Given that the SLR model assumptions are satisfied, we show that least square estimators are unbiased:

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \ \mathbb{E}(\hat{\beta}_0) = \beta_0 \tag{73}$$

and the variance of the estimators are:

$$Var(\hat{\beta}_1 | x_{11}, \ldots, x_{n1}) = \frac{\sigma^2}{S_{XX}}, \ Var(\hat{\beta}_0 | x_{11}, \ldots, x_{n1}) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}_1^2}{S_{XX}} \right) \tag{74}$$

Proof. See handout or do it yourself.

□

13 How to estimate σ^2

Though σ^2 is not used to estimate $\hat{\beta}_0, \hat{\beta}_1$, we can use it to understand:

- Randomness in Y.
- σ^2 is related to $Var(\hat{\beta}_0|\mathbf{X})$ and $Var(\hat{\beta}_1|\mathbf{X})$.

We can estimate σ^2 using the data. We can see

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \Rightarrow \varepsilon = Y - \beta_0 - \beta_1 X_1$$

We can show that this is an unbiased estimator:

$$\begin{aligned}
\mathbb{E}[(Y - \beta_0 - \beta_1 X_1)^2] &= \mathbb{E}(\varepsilon^2) && (75)\\
&= Var(\varepsilon) + (\mathbb{E}(\varepsilon))^2 && (76)\\
&= Var(\varepsilon) + 0 && (77)\\
&= \sigma^2 && (78)
\end{aligned}$$

Plug-in Principle: Replace X_1 with $(x_{11}, \ldots, x_{n1})$ and Y with $(y_1, \ldots, y_n)$, replace $\mathbb{E}(\cdot)$ with $\frac{1}{n}\sum_{i=1}^n \cdot$, β_1 with $\hat{\beta}_1$, and β_0 with $\hat{\beta}_0$.

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[(Y - \beta_0 - \beta_1 X_1)^2] && (79)\\
\text{(bias)} &\approx \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2}{n} && (80)\\
\text{(unbiased)} &\approx \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2}{n-2} && (81)\\
&= \frac{SS_{Res}}{n-2} = MS_{Res} := \hat{\sigma}^2 && (82)
\end{aligned}$$

The unbiased adjustment has little affect when n is large.

13.1 Using $\hat{\sigma}^2$ to estimate $Var(\hat{\beta}_0|\mathbf{X})$ and $Var(\hat{\beta}_1|\mathbf{X})$

From (74), we can plug in $\hat{\sigma^2}$:

$$\begin{aligned}
Var(\hat{\beta}_1|x_{11}, \ldots, x_{n1}) &= \frac{\hat{\sigma}^2}{S_{XX}} && (83)\\
Var(\hat{\beta}_0|x_{11}, \ldots, x_{n1}) &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}_1^2}{S_{XX}}\right) && (84)
\end{aligned}$$

where $x_{11}, \ldots, x_{n1} \equiv \mathbf{X}$. The *standard errors* (se) are the square roots of the variances:

$$\begin{aligned}
se(\hat{\beta}_1) &= \sqrt{\frac{\sigma^2}{S_{XX}}} && (85)\\
se(\hat{\beta}_0) &= \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}_1^2}{S_{XX}}\right)} && (86)
\end{aligned}$$

Using $\hat{\sigma}^2$ gives us the *estimated standard errors* (ese):

$$\begin{aligned}
ese(\hat{\beta}_1) &= \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} && (87)\\
ese(\hat{\beta}_0) &= \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}_1^2}{S_{XX}}\right)} && (88)
\end{aligned}$$

13.2 Sampling Distribution of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$

With an additional assumption (Gaussian-Noise Simple Linear Regression (GN-SLR)), we can show:

- $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) = N(\beta_1, se(\hat{\beta}_1)^2)$
- $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}_1^2}{S_{XX}}\right)\right) = N(\beta_0, se(\hat{\beta}_0)^2)$
- $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ (p=2, 1 predictor and 1 intercept)

GN-SLR Assumptions:

1. Same as SLR
2. Same as SLR
3. $\varepsilon \sim N(0, \sigma^2)$
4. Same as SLR

GN-SLR is a special case of SLR, which has much more general conditions, with an additional Gaussian assumption, thus GN-SLR $\xrightarrow{\text{implies}}$ SLR. Under GN-SLR, $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$, it suggests that $Y_i | X_{i1} \sim N(\beta_0 + \beta_1 X_{i1}, \sigma^2)$. (There's a checklist in this part of the notes (lec11), watch the associated lecture vid just know what's going on)

By standardizing $\hat{\beta}_1$ and $\hat{\beta}_0$, we can show that

- $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim N(0,1)$
- $\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim N(0,1)$

But, using *se* is not ideal since it is related to the unknown σ^2, thus we use *ese* instead:

- $T_1 = \frac{\hat{\beta}_1 - \beta_1}{ese(\hat{\beta}_1)} \sim t_{n-2}$
- $T_0 = \frac{\hat{\beta}_0 - \beta_0}{ese(\hat{\beta}_0)} \sim t_{n-2}$

14 Maximum Likelihood Estimation

Using the GN-SLR setting, assume $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$. We only observe $\{x_{i1}, y_i\}_{i=1}^n$. To estimate $\beta_0, \beta_1, \sigma^2$ using MLE:

$$Y_i | X_{i1} \sim N(\beta_0 + \beta_1 X_{i1}, \sigma^2)$$

where Y_i has the density function:

$$f_Y(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{i1})^2\right\} \tag{89}$$

The likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f_Y(y_i) \tag{90}$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{i1})^2\right\} \tag{91}$$

To estimate $\beta_0, \beta_1, \sigma^2$, we solve the following optimization problem:

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) = \arg \max_{\beta_0, \beta_1, \sigma^2} L(\beta_0, \beta_1, \sigma^2) \tag{92}$$

$$= \arg \max_{\beta_0, \beta_1, \sigma^2} \log L(\beta_0, \beta_1, \sigma^2) \tag{93}$$

$$= \arg \min_{\beta_0, \beta_1, \sigma^2} -\log L(\beta_0, \beta_1, \sigma^2) \tag{94}$$

The negative log-likelihood function

$$-\log L(\beta_0, \beta_1, \sigma^2) = \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 + c \tag{95}$$

$$\arg \min_{\beta_0, \beta_1, \sigma^2} -\log L(\beta_0, \beta_1, \sigma^2) = \min_{\sigma^2} \left\{ \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \right\} \tag{96}$$

which can be separated into

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \tag{97}$$

$$= \arg \min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \tag{98}$$

$$\min_{\sigma^2} \left\{ \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2 \right\} = \min_{\sigma^2} \left\{ \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} SS_{Res} \right\} \tag{99}$$

$$\Rightarrow \frac{\partial}{\partial \sigma^2} \left\{ \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} SS_{Res} \right\} \tag{100}$$

$$= \frac{n}{2\sigma^2} - \frac{SS_{Res}}{2\sigma^4} = 0 \tag{101}$$

So, we obtain:

- $\hat{\sigma}^2_{MLE} = \frac{SS_{Res}}{n}$ (biased)
- $\hat{\sigma}^2_{LS} = \frac{SS_{Res}}{n-2}$ (unbiased)

In summary:

- $\hat{\beta}_{plug} = \hat{\beta}_{LS} = \hat{\beta}_{MLE}$
- $\hat{\sigma}^2_{LS} = \frac{n}{n-2} \hat{\sigma}^2_{MLE}$