Regression Analysis Notes

Week-2

Bivariate Normal & Relation to Simple Linear Regression

So far we have assumed that values of the x (the predictors) are fixed. THat may be realistic in certain situations, e.g., Demand curve, that may be realistic. How demand changes due to change in Price (maybe double), However, in most cases, we can't really fix X.

e.g., the effect of age on BP, people are walking into a hospital clinic & you record their age & BP. In such situations, does the linear model still hold if we condition on X?

The answer is yes, if $(X, Y) \sim N_2(\cdots)$. The conditional distribution is given by:

$$Y|X = x \sim N(\beta_0 + \beta_1 x + ..., \sigma^2)$$

This can be derived from the joint bivariate normal distribution. The conditional density is the ratio of the joint density to the marginal density:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Let the joint distribution be:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

Where $X \sim N(\mu_1, \sigma_1^2)$.

The joint PDF is:

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\frac{\rho(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

The marginal PDF of X is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\}$$

After dividing and simplifying the expression in the exponent, we get:

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2\sigma_2^2(1-\rho^2)} \left[y - \left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1) \right) \right]^2 \right\}$$

This shows that the conditional distribution of Y given $X = x$ is Normal:

$$Y|X = x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2) \right)$$

This is a simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$, where.

$$\epsilon \sim N(0, \sigma_2^2(1-\rho^2))$$

:

- Regression slope: $\beta_1 = \rho\frac{\sigma_2}{\sigma_1}$
- Regression intercept: $\beta_0 = \mu_2 - \beta_1\mu_1$

- Error variance: $\sigma^2 = \mathrm{Var}(\epsilon) = \sigma_2^2(1-\rho^2)$

β_0 and β_1 are the population versions of the least squares estimates ($\hat{\beta}_{0,LS} = \bar{y} - \hat{\beta}_{1,LS}\bar{x}$ and $\hat{\beta}_{1,LS} = \frac{S_{xy}}{S_{xx}}$).

$$\rho^2 = 1 - R^2 = 1 - \frac{\text{SSReg}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$
$$1 - \rho^2 = \frac{\sigma^2}{\sigma_2^2}, \qquad \sigma^2 = \mathrm{Var}(\epsilon), \quad \sigma_2^2 = \mathrm{Var}(Y)$$

Summary

When $(X, Y) \sim N_2$, the simple linear regression model holds. The conditional mean of Y given $X = x$ is linear in x. All parameters are population versions of the least squares estimates.

Goals

1. Estimate / Test for parameters (Inference).
2. Check whether model assumptions are valid (Diagnostics).

Diagnostics for Simple Linear Regression

1. **Linearity:** Plot residuals against x_i (or fitted values). If there's a pattern, the relationship may be non-linear.
2. **Independence of errors:** ϵ_i are independent of each other. This is difficult to check directly. Serial dependence is a particular for data from time series .
3. **Independence of errors and predictors:** ϵ_i are independent of x_i , x vs ϵ should have no pattern.
4. **Zero mean of errors:** $E(\epsilon_i) = 0$.
5. **Constant variance (Homoscedasticity):** $\mathrm{Var}(\epsilon_i) = \sigma^2$. Plot residuals $\hat{\epsilon}_i$ against fitted values. A funnel shape suggests non-constant variance.
6. **Normality of errors:** If you assume ϵ_i are normal for tests, check this with a QQ-plot of residuals of $\hat{\epsilon}_i$.
7. **Leverage:** Identify points that have a high influence on the regression line. Leverage measures how much the regression line changes by the addition of a point.

Matrix Notation for Regression

Simple Linear Regression

The model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. In matrix form, $\underline{y} = X\beta + \underline{\epsilon}$:

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1}, \quad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

Normal Equations

The least squares estimates are found by solving the normal equations, which are derived by minimizing the sum of squared errors.

$$\begin{aligned} \sum (y_i - \beta_0 - \beta_1 x_i) = 0 &\implies n\beta_0 + (\sum x_i)\beta_1 = \sum y_i \\ \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0 &\implies (\sum x_i)\beta_0 + (\sum x_i^2)\beta_1 = \sum x_i y_i \end{aligned}$$

In matrix form, this is written as:

$$X^T X \beta = X^T \underline{y}$$

The least squares estimates for β are given by:

$$\hat{\beta} = (X^T X)^{-1} X^T \underline{y}$$

Multiple Linear Regression

Single Continuous Response $\to y$
Multiple Predictors $\to x_1, \ldots, x_p$
Each observation $(y_i, x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n$

$i \to$ day

$y_i \to$ Total Rainfall on day i.

$x_{i1} \to$ Total Temp at 6 am on day i.

$x_{i2} \to$ Relative Humidity at 8AM on day i.

$\vdots$

$x_{ip} \to$ other factors.

Linear Model

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n$

Assumptions

1. Linear model holds
2. ϵ_i are independent of each other.
3. ϵ_i are independent of $x_{i1}, \ldots, x_{ip}$.
4. $E[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2$. ($\epsilon_i$ are continuous r.v.'s)

Matrix Notations

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \qquad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$$

The model can be written as

$$Y = X\beta + \epsilon \quad \to \quad \epsilon = Y - X\beta$$

Least Squares Criterion

Minimize $\sum_{i=1}^n \epsilon_i^2$ (minimize sum of squares of errors)

$$\begin{aligned} \sum \epsilon_i^2 &= \underline{\epsilon}^T \underline{\epsilon} = (Y - X\beta)^T (Y - X\beta) \\ &\text{where } \underline{\epsilon} = Y - X\beta \\ &= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T (X^T X)\beta \\ S(\beta) &= Y^T Y - 2Y^T X\beta + \beta^T (X^T X)\beta \quad 1 \end{aligned}$$

Need to take the derivative of ① w.r.t β,

$$\frac{\partial S(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial S}{\partial \beta_0} \\ \frac{\partial S}{\partial \beta_1} \\ \vdots \\ \frac{\partial S}{\partial \beta_p} \end{pmatrix} = 0 \quad \text{(Normal eq's)}$$

Let A be a $1 \times (p+1)$ matrix, $A = (a_0, a_1, \ldots, a_p)$.

$$A\beta = \sum_{j=0}^{p} a_j \beta_j$$

$$\frac{\partial}{\partial \beta}(A\beta) = A^T$$

In our case, $A = -2\underset{1\times n}{Y^T} \underset{n\times(p+1)}{X}$, so

$$\frac{\partial}{\partial \beta}(-2Y^T X\beta) = (-2Y^T X)^T = -2X^T Y$$

Now consider the quadratic term with a matrix $B_{(p+1)\times(p+1)}$.

$$B = \begin{pmatrix} b_{00} & b_{01} & \cdots \\ b_{10} & & \\ \vdots & & \\ b_{p0} & \ldots & b_{pp} \end{pmatrix} = ((b_{ij})) \qquad \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

B is symmetric.

Least Squares Estimation Notes

$$\frac{\partial(\beta^T B\beta)}{\partial \beta_i} = \sum_j b_{ij} B_j + \sum_i B_i b_{ij} = 2\sum_j b_{0j} B_j$$

$$\frac{\partial(\beta^T B\beta)}{\partial \beta_1} = 2b^T \beta$$

$$\begin{bmatrix} \frac{\partial(\beta^T B\beta)}{\partial \beta_0} \\ \frac{\partial(\beta^T B\beta)}{\partial \beta_1} \\ . \\ . \\ \frac{\partial(\beta^T B\beta)}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} 2b_0^T \beta \\ 2b_1^T \beta \\ . \\ . \\ 2b_p^T \beta \end{bmatrix} = 2B\beta$$

If $\beta = X^T X$ is symmetric, then

$$\frac{\partial(\beta^T X^T X\beta)}{\partial \beta} = 2X^T X\beta$$

$$[Y^T Y - 2Y^T X\beta + \beta^T (X^T X)\beta] = 0$$

Normal Equations
Thus,

$$-2X^T Y + 2X^T X\beta = 0 \Rightarrow X^T X\beta = X^T Y$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

This is the least squares estimate of β.

Condition: $n > p+1$.

For example, $p = 1, n = 2$.

In regression, there are more data points than predictor dimensions; otherwise, we can have a perfect fit with all $\hat{\varepsilon}_i = 0$, which will be the least squares solution, $\hat{\beta} = X^T Y$ if $n = p+1$. All the formulation of least squares or best line have to be done because we have more points than dimensions.

$$Y = X\beta + \varepsilon$$

Residuals are $y_i - \hat{y}_i = \hat{\varepsilon}_i$.

Fitted values are $\hat{y}_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_{ip}$.

Assumptions need to be verified for the errors.
We verify them for the residuals. - Plot residuals against fitted values of each predictor.(Linear model dependence on x_i)
- Plot squared residuals against fitted values of each predictor (Variance is same, check for homoscedasticity). - Normal Q–Q plot if normality is assumed.(outliers, influential observations , leverage)

Expectation, Variance , Covariance of matrix of $\hat{\beta}$

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad Y = X\beta + \varepsilon$$

$$E[X] = \mu, \quad Var(X) = \Sigma$$

$$E[AX + b] = AE[X] + b$$

$$Var(AX + b) = AVar(X)A^T$$

$$E[\hat{\beta}] = \beta + (X^T X)^{-1} X^T E[\varepsilon] = \beta$$

$$Var(\hat{\beta}) = (X^T X)^{-1} X^T Var(\varepsilon) X (X^T X)^{-1}$$

If $Var(\varepsilon) = \sigma^2 I$, then

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Variance Properties

If $Ax = \sum_{j=1}^{K} a_{ij} x_j$, then

$$Var(Ax + By) = a^2 Var(x) + b^2 Var(y) + 2ab\, Cov(x, y)$$

For a random vector X and a matrix A:

$$Var(AX) = A\, Var(X)\, A^T$$

Sum of Squares Decomposition

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= Y^T Y - 2Y^T X\hat{\beta} + \hat{\beta}^T (X^T X)\hat{\beta}$$

Substituting $\hat{\beta} = (X^T X)^{-1} X^T Y$:

$$SSR = Y^T (I - X(X^T X)^{-1} X^T) Y$$

Degrees of freedom:

Source	SS	df
Regression	SS_{Reg}	p
Residual	SS_{Res}	$n-(p+1)$
Total	SST	$n-1$

$$\hat{\sigma}^2 = \frac{SSR}{n-(p+1)}$$

$$R^2 = 1 - \frac{SSR}{SST}, \quad \text{Adjusted } R^2 = 1 - \frac{SSR/(n-(p+1))}{SST/(n-1)}$$

As more predictors are added, SSR decreases.