L-5

1. **Linear Model**

$$Y = \beta_0 + \beta_1 \log x + \beta_2 x^2 + \epsilon \quad \text{is a Linear Model}$$

$$Y = \frac{\beta_1 x}{\beta_0 + x} + \epsilon \quad \text{is a non-linear model}$$

The 2nd one can't be expressed as a linear combination of known functions of x.

2. We are only dealing with models, which is not necessarily the truth.

"All models are wrong but some models are useful" – BOX (1976)

3. We do not claim a causal relation b/w X & Y. i.e. trying to predict Y based on X. Not claiming X causes Y or Y causes X.

4. Interpretation of the coefficient In univariate ($p = 1$) regression, the interpretation is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

$\beta_0 \rightarrow$ expected value of Y when $x = 0$.

$\beta_1 \rightarrow$ avg change in Y for a unit change in X.

Multiple Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \beta_1 > 0$$

If we disregard X_2, then the overall pattern of Y on X_1 is decreasing. $Y = \beta_0 + \beta_1 X_1 + \epsilon, \beta_1 < 0$.

In multiple regression, the interpretation of β_1 is the avg change in Y for a unit change in X_1, when X_2 is held constant. In general, the regression coefficient measures the change in response for a unit change in the corresponding predictor when all other predictors are held constant.

Collinearity A linear dependence b/w 2 or more predictors/columns of the design matrix. e.g. $\mathbf{x_1} = 2\mathbf{x_2}$ or $\mathbf{x_1} = \mathbf{x_2} + 2\mathbf{x_3}$. At the population level, this makes the coeff (β_1, β_2) ill-defined.

$y = x_1 + x_2 + 6$	$(\beta_1 = 1, \beta_2 = 1)$
$y = 3x_2 + 6$	$(0, 3)$
$y = -1.5x_2 + 6$	$(-1.5, 0)$

Infinitely many equations represent the same plane If $\mathbf{x_1} = 2\mathbf{x_2}$ then $(X^T X)$ can't be defined. At the estimate level:

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

As $(X^T X)$ is not full rank, $\det(X^T X) = 0$. Even if the det is not zero but very small, this $(X^T X)^{-1}$ becomes unstable conceptually. This is c/d near collinearity & we should try to avoid, we want to remove some of the columns, but which ones?

If the relation involves only 2 columns, then a pairwise plot e.g. (X_1, X_2), (X_1, X_3), (X_2, X_3) will reveal an exact straight line in one of the plots. Then drop any one of the variables in that plot. Even a very high correlation b/w X1,X3 dictates we drop one of them. What if $x_1 = x_2 + 2x_3$? This can't be detected in a particular pairwise plot.

- The columns not in the linear relation will have small numbers in the corresponding diagonal of $(X^T X)^{-1}$. Retain those.

- From the subset that is not retained, remove elements one by one & check if the overall determinant stabilizes.

L-6

- **Collinearity is**: Linear or approximate linear relationship b/w the predictor variables.
- **Why is it a problem?**: $\hat{\beta}$ is unstable.
- **How to detect**: $\det(X^TX) \approx 0$. i.e. $\sigma^2(X^TX)^{-1}$ is very large.

Another problem of Multiple Regression: Interaction Nature of relationship b/w Y & X_1 depends on the value of X_2. Our assumption is that every variable makes a distinct additive contribution to the response. Model with interaction term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

Start with a bigger (more interaction) model. Test the hypothesis whether the interaction term is zero. If the hypothesis ($H_0 : \beta_{12} = 0$) is not rejected, then we can go ahead with the no interaction simpler model.

Normality assumption and the geometry of least squares

$$Y = X\beta + \epsilon$$

Minimizing $\sum_{i=1}^{n} {\epsilon_i}^2 = ||\epsilon||^2$. $y \in \mathbb{R}^n$, $X\beta$ is a plane in the n-dimensional space. The distance is minimized when we project a perpendicular/normal from the point y to the plane. The normal equation is:

$$X^T\epsilon = 0$$
$$X^T(Y - X\beta) = 0$$
$$X^TY = X^TX\beta$$

Inference on Linear Regression

Assumptions: $E[\epsilon] = 0$, $Var(\epsilon) = \sigma^2 I_n$. Assume ϵ_i has a normal dist.

$$\epsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$$

Likelihood:

$$L(\beta, \sigma^2 | Y_1, ..., Y_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\epsilon^T\epsilon\right\}$$
$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right\}$$

Log-likelihood:

$$l(\beta, \sigma^2 | Y_1, ..., Y_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

For β, maximizing likelihood is equivalent to minimizing $(Y - X\beta)^T(Y - X\beta)$. This is also the least squares criterion. Under normality, $\hat{\beta}_{LS}$ is same as $\hat{\beta}_{MLE}$.

For σ^2:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n/2}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(Y - X\beta)^T(Y - X\beta) = 0$$
$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = \frac{1}{n}SSR$$

We were using $\frac{1}{n-(p+1)}SSR$ for σ^2 This is unbiased.

Distribution of Estimators and ANOVA in Linear Regression

1 Expectation of SSR

We begin with

$$\begin{aligned}
Y - X\hat{\beta} &= Y - X(X^TX)^{-1}X^TY \\
&= X\beta + \epsilon - X(X^TX)^{-1}X^T(X\beta + \epsilon) \\
&= X\beta + \epsilon - X(X^TX)^{-1}X^TX\beta - X(X^TX)^{-1}X^T\epsilon \\
&= \epsilon - X(X^TX)^{-1}X^T\epsilon.
\end{aligned}$$

Define

$$A = I - X(X^TX)^{-1}X^T.$$

Then

$$Y - X\hat{\beta} = A\epsilon.$$

Here, A is symmetric ($A^T = A$) and idempotent ($A^2 = A$).

If $\epsilon \sim N(0, \sigma^2 I)$, then

$$A\epsilon \sim N(0, \sigma^2 A).$$

[Quadratic Form Distribution] If $U \sim N(0, \sigma^2 I_k)$ and A is a symmetric, idempotent matrix of rank k, then

$$\frac{1}{\sigma^2}U^TAU \sim \chi^2_k.$$

Let $u = Y - X\hat{\beta}$. Then

$$\frac{1}{\sigma^2}u^Tu = \frac{1}{\sigma^2}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = \frac{SSR}{\sigma^2} \sim \chi^2_k.$$

The rank of A is

$$\begin{aligned}
k = (A) &= (I_n) - \big(X(X^TX)^{-1}X^T\big) \\
&= n - \big((X^TX)^{-1}X^TX\big) = n - (I_{p+1}) = n - (p+1).
\end{aligned}$$

Since $[\chi^2_k] = k$,

$$\left[\frac{SSR}{\sigma^2}\right] = n - (p+1) \quad \Rightarrow \quad [SSR] = \sigma^2\big(n - (p+1)\big).$$

An unbiased estimator of σ^2 is

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

2 Distribution of $\hat{\beta}$

$$\begin{aligned}
\hat{\beta} &= (X^TX)^{-1}X^TY \\
&= (X^TX)^{-1}X^T(X\beta + \epsilon) \\
&= \beta + (X^TX)^{-1}X^T\epsilon.
\end{aligned}$$

Let $A = (X^T X)^{-1} X^T$. Then $A\epsilon$ is normal with

$$[A\epsilon] = 0, \quad (A\epsilon) = \sigma^2 (X^T X)^{-1}.$$

Hence,

$$\hat{\beta} \sim N_{p+1}\big(\beta, \sigma^2 (X^T X)^{-1}\big).$$

In particular,

$$\hat{\beta}_i \sim N\big(\beta_i, \sigma^2 \big((X^T X)^{-1}\big)_{ii}\big).$$

[Independence] $\hat{\beta}$ and s^2 are independent.

3 t-distribution of the Coefficients

Using

$$\frac{N(0,1)}{\sqrt{\chi^2_k / k}} \sim t_k,$$

and under $H_0 : \beta_i = \beta_{i0}$,

$$\frac{\hat{\beta}_i - \beta_{i0}}{\sigma \sqrt{((X^T X)^{-1})/(n-(p+1))}} \sim N(0,1),$$

$$\frac{SSR}{\sigma^2} = \frac{(n-(p+1))s^2}{\sigma^2} \sim \chi^2_{n-(p+1)}.$$

Then

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{s \sqrt{((X^T X)^{-1})/(n-(p+1))}} \sim t_{n-(p+1)}.$$

4 Simple Linear Regression ($p = 1$)

Slope

$$T = \frac{\hat{\beta}_1}{s / \sqrt{\sum (x_i - \bar{x})^2}} \sim t_{n-2}.$$

$$s^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2.$$

Intercept

$$t = \frac{\hat{\beta}_0 - \beta_{0,a}}{s \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

$$\hat{\beta}_0 \pm t_{(n-2),\alpha/2} \cdot s \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

gives a $(1-\alpha)\%$ CI for β_0.

Confidence Interval for β_1

$$\hat{\beta}_1 \pm t_{(n-2),\alpha/2} \cdot \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

5 Proof of Independence of $\hat{\beta}$ and s^2

$$\hat{\beta} - \beta = A\epsilon,$$
$$U = B\epsilon, \quad B = I - X(X^TX)^{-1}X^T.$$

They are independent if

$$(A\epsilon, B\epsilon) = A\,(\epsilon)\,B^T = 0.$$

$$\begin{aligned} A(\sigma^2 I)B^T &= \sigma^2(X^TX)^{-1}X^T(I - X(X^TX)^{-1}X^T) \\ &= \sigma^2\left[(X^TX)^{-1}X^T - (X^TX)^{-1}X^TX(X^TX)^{-1}X^T\right] \\ &= 0. \end{aligned}$$

Since they are jointly normal, this implies independence.

6 ANOVA Table for Simple Linear Regression ($p = 1$)

Source	**SS**	**df**	**MS**	**F**
Regression	SS_{Reg}	p	$\frac{SS_{Reg}}{p}$	$\frac{MS_{Reg}}{MS_R}$
Residual (Error)	SSR	$n-(p+1)$	$\frac{SSR}{n-(p+1)}$	
Total	SST	$n-1$		

$$SST = \sum(y_i - \bar{y})^2, \quad SSR = \sum(y_i - \hat{y}_i)^2.$$

Under $H_0 : \beta_1 = \cdots = \beta_p = 0$:

$$\frac{SS_{Reg}}{\sigma^2} \sim \chi^2_p, \qquad \frac{SSR}{\sigma^2} \sim \chi^2_{n-(p+1)}.$$

These are independent, so

$$F = \frac{SS_{Reg}/p}{SSR/(n-(p+1))} \sim F_{p,n-(p+1)}.$$

7 Fisher–Cochran Theorem (Matrix Form)

$$\begin{aligned} \epsilon^T\epsilon = \epsilon^T\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\epsilon + \epsilon^T(I - X(X^TX)^{-1}X^T)\epsilon \\ + \epsilon^T\left(X(X^TX)^{-1}X^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\epsilon. \end{aligned}$$

$$SST = \epsilon^TB^{(1)}\epsilon + \epsilon^TB^{(2)}\epsilon + \epsilon^TB^{(3)}\epsilon,$$

where $B^{(1)}, B^{(2)}, B^{(3)}$ are symmetric and idempotent.

$$\operatorname{rank}(B^{(1)}) = 1, \quad \operatorname{rank}(B^{(2)}) = n - (p+1), \quad \operatorname{rank}(B^{(3)}) = p.$$

Sum of ranks:

$$1 + n - (p+1) + p = n.$$

By Cochran's theorem, the quadratic forms are independent χ^2 variables.

8 t-test vs F-test

For $p = 1$, testing $H_0 : \beta_1 = 0$ via t-test or ANOVA F-test gives the same result:

$$F = t^2.$$

For $p \geq 2$, the F-test checks joint significance:

$$H_0 : \beta_1 = \cdots = \beta_p = 0,$$

while the t-test checks individual effects.

9 Prediction

For a new point x_0, the point prediction is

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\beta} = \mathbf{x}_0^T (X^T X)^{-1} X^T Y.$$

$$\begin{aligned} [\hat{y}_0] &= \mathbf{x}_0^T (X^T X)^{-1} X^T X \beta = \mathbf{x}_0^T \beta, \\ (\hat{y}_0) &= \sigma^2 \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0. \end{aligned}$$

$$\hat{y}_0 \sim N\big(\mathbf{x}_0^T \beta, \sigma^2 \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0\big).$$

Figure 1: Scatter plot with fitted regression line and confidence/prediction intervals at x_0.