Outline

Multiple Linear Regression Model

The formal multiple linear regression (MLR) model for the data $(x_{i1}, x_{i2}, \ldots, x_{i,p-1}, y_i)$ is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

for $i = 1, 2, \ldots, n$, where

- Y_i is the *i*th observation of the **response variable.**
- X_{ik} is the *i*th observation of the *k*th **explanatory variable** for $k = 1, \ldots, p-1$.
- ε_i is the *i*th **random error** term.
- The random errors follow a normal distribution with mean zero and variance σ^2 and are independent of each other.
- That is, $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$.

Model Parameters

- The model parameters are $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$, and σ^2 (population parameters).
- β_0 and $\beta_1, \beta_2, \ldots, \beta_{p-1}$: **regression coefficients.**
- β_0: **intercept.**
  β_0 interpreted as ____________________
- β_k: **slope** for $k = 1, \ldots, p-1$.
  β_k interpreted as____________________
- σ^2: **error variance**, sometimes written as σ_ε^2.

Features of Multiple Linear Regression Model

Under the MLR model for the data $(x_{i1}, x_{i2}, \ldots, x_{i,p-1} y_i)$:

- Multiple ____________________
- Linear ____________________
- Regression regression toward the mean [Galton, 1886]
- Randomness Q: What kind of distribution does Y_i have?

  ____
- Independence. ____________________
- The model parameters: $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}, \sigma^2$.

Notation

- Response variable: $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)'$.
- Design matrix:

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

- Random error: $\boldsymbol{\varepsilon}_{n\times 1} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$.
- Regression coefficients: $\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$.
- The multiple linear regression model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}_{n\times 1}, \sigma^2 \boldsymbol{I}_{n\times n}).$$

Example: $p = 3$

- Example: # of explanatory variables = 2.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2),$$

  for $i = 1, \ldots, n$.
- Mean response:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}.$$

- Interpretation:
  - β_0: Intercept. The mean response $\mathbb{E}(Y)$ at $X_1 = X_2 = 0$.
  - β_1: Slope. The change in the mean response $\mathbb{E}(Y)$ per unit increase in X_1, when X_2 is held constant.
  - β_2: Slope. The change in the mean response $\mathbb{E}(Y)$ per unit increase in X_2, when X_1 is held constant.

Dummy variable

- The predictors in the linear model can be either continuous (e.g., age, height) or categorical (e.g., gender, group)
- For a categorical predictor that has p categories, define $p-1$ dummy variables:

$$X_{ik} = \begin{cases} 1 & \text{observation } i \text{ is in category } k \\ 0 & \text{otherwise} \end{cases}$$

  where $k = 1, \ldots, p-1$.
- Include dummy variables as predictors in the linear model;
- Example. Consider n i.i.d. observations from the following model:

$$Y = \beta_0 + \beta_1 \text{Age} + \beta_2 X + \varepsilon, \quad \text{where } \varepsilon \sim i.i.d.\ N(0, \sigma^2),$$

  with $X = 1$ if male, $X = 0$ if female.
- What is the interpretation for β_0, β_1, and β_2?

Example with categorical variables

Consider the effect of education on hourly wages (Y). The education is classified into three categories:

Option in Survey (O)	Meaning (M)
1	College dropout
2	College
3	MS and above

Which model makes more sense?

- $Y = \beta_0 + \beta_1 O + \varepsilon$?
- $Y = \beta_0 + \beta_1 1_{\text{college}} + \beta_2 1_{\text{MS and above}} + \varepsilon$?
- $Y = \beta_0 + \beta_1 1_{\text{college dropout}} + \beta_2 1_{\text{college}} + \varepsilon$?

(In all cases, assume $\varepsilon \sim i.i.d. N(0, \sigma^2)$)

Example (Cont.)

- To include the eduction as predictor in a regression model, define 2 dummy variables X_1 and X_2:

Option in Survey (O)	Meaning (M)	X_1	X_2
1	College dropout	0	0
2	College	1	0
3	MS and above	0	1

- Baseline (all dummies 0): college dropout;
- $X_1 = 1$, if the highest degree is college, 0 otherwise;
- $X_2 = 1$, if degree with MS and above, 0 otherwise.

Include X_1 and X_2 as dummy variables in a regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_3 + \ldots + \beta_3 X_p}_{\text{other predictors, e.g., age}} + \varepsilon, \quad \varepsilon \sim i.i.d.\ N(0, \sigma^2).$$

Outline

Least Squares Estimation

- Consider the criterion:

$$
\begin{aligned}
Q(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_{p-1} X_{i,p-1})^2 \\
&= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).
\end{aligned}
$$

- Following the arguments for SLR in matrix terms, we can show that the least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

assuming that the $p \times p$ matrix $\boldsymbol{X}'\boldsymbol{X}$ is invertible.

Fitted Values and Residuals

- Fitted values: $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)'$.
- Following the arguments for SLR in matrix terms, we have

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{H}\boldsymbol{Y},$$

where the "hat matrix" is defined as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

- Residuals: $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)'$.
- Following the arguments for SLR in matrix terms, we have

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

Properties of the hat matrix $\boldsymbol{H}$

- $\boldsymbol{H}$ is symmetric and idempotent:
  $\boldsymbol{H}^2 = \boldsymbol{H}$, and Rank($\boldsymbol{H}$) = Tr($\boldsymbol{H}$) = p.
- $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent:
  $(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$, and Rank($\boldsymbol{I} - \boldsymbol{H}$) = Tr($\boldsymbol{I} - \boldsymbol{H}$) = $n - p$.

Estimation of Regression Coefficients

- The LS estimate $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$. That is,

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

- The variance-covariance matrix is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \in \mathbb{R}^{p \times p}$$

where

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{Var}(\hat{\beta}_1) & \cdots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_{p-1}) \\ \vdots & \vdots & \vdots & \\ \text{Cov}(\hat{\beta}_{p-1}, \hat{\beta}_0) & \text{Cov}(\hat{\beta}_{p-1}, \hat{\beta}_1) & \cdots & \text{Var}(\hat{\beta}_{p-1}) \end{bmatrix}$$

Distribution of regression coefficients estimates

$$\hat{\boldsymbol{\beta}} \sim \mathcal{MVN}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

Inference of Regression Coefficients

- The estimated variance-covariance matrix.

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Marginally, we have

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\mathrm{Var}(\hat{\beta}_k)}}} \sim T_{n-p}, \quad \text{for all } k = 0, 1, \ldots, p-1.$$

Inference of Regression Coefficients

- Thus the $(1-\alpha)$ confidence interval for β_k is

$$\hat{\beta}_k \pm t_{n-p,\alpha/2}\sqrt{\widehat{\text{Var}(\hat{\beta}_k)}}.$$

- Hypothesis testing:

$$H_0 : \beta_k = \beta_k^0 \text{ versus } H_A : \beta_k \neq \beta_k^0.$$

- Under the H_0, we have

$$T^* = \frac{\hat{\beta}_k - \beta_k^0}{\sqrt{\widehat{\text{Var}(\hat{\beta}_k)}}} \sim T_{n-p}, \quad \text{Why } n-p?$$

Inference on the linear contrast

Recall the study that investigates the effect of education on hourly salary (Y):

Education	X_1	X_2
College dropout	0	0
College	1	0
MS and above	0	1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \text{where} \quad \varepsilon \sim i.i.d.\ N(0, \sigma^2).$$

Suppose we are interested in testing:

- The mean salary for "MS and above" is the same as for "College": ________________
- The mean salary for "College" is the same as for "College dropout": ________________
- The mean salary for "MS and above" is twice as that or "College": ________________

Inference on the linear contrast

- All these hypothesis tests could be expressed as a linear contrast:

$$H_0 : c_0\beta_0+c_1\beta_1+c_2\beta_2 = 0 \quad \text{v.s.} \quad H_\alpha : c_0\beta_0+c_1\beta_1+c_2\beta_2 \neq 0,$$

  for a given vector $\boldsymbol{c} = (c_0, c_1, c_2)$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$.
- What is the distribution of $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ under the null? Multivariate normal with

$$\mathbb{E}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \boldsymbol{c}'\boldsymbol{\beta}, \quad \text{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \underline{\hspace{3cm}} = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$$

- In case σ^2 is unknown, plug in the estimator $\hat{\sigma}^2$. (what is the form of $\hat{\sigma}^2$?)

$$\frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\widehat{\text{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})}}} \sim T_{n-3}$$

Outline

Estimation of Mean Response

- Define a new observation with predictor $\boldsymbol{X}_h = (1, X_{h1}, \ldots, X_{h,p-1})'$. Estimate $\mu_h = \mathbb{E}(\boldsymbol{X}_h'\beta)$?
- The estimated mean response corresponding to $\boldsymbol{X}_h$:

$$\hat{\mu}_h = \boldsymbol{X}_h'\hat{\beta}.$$

- Distribution of $\hat{\mu}_h$:

$$\hat{\mu}_h \sim N\left(\boldsymbol{X}_h'\beta, \sigma\sqrt{\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h}\right).$$

- Mean. ____________________
- Variance. ____________________

Confidence Intervals for Mean Response

- Estimated variance.

$$\widehat{\mathrm{SD}(\hat{\mu}_h)} = \hat{\sigma}\sqrt{\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h}.$$

- The $(1-\alpha)$ confidence interval for $\hat{\mu}_h$ is

$$\hat{\mu}_h \pm t_{n-p,\alpha/2}\widehat{\mathrm{SD}(\hat{\mu}_h)}$$

- Hypothesis tests on μ_h can be carried out similarly.

Prediction of New Observation

- The predicted new observation corresponding to $\boldsymbol{X}_h$:

$$\widehat{Y}_h = \boldsymbol{X}_h' \hat{\beta}.$$

- What is the MSE of $\widehat{Y}_h$ for predicting $Y_{h(\mathrm{new})}$?
- Prediction error variance:

$$\mathrm{Var}(\widehat{Y}_h - Y_{h(\mathrm{new})}) = \sigma^2 \left(1 + \boldsymbol{X}_h' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}_h \right).$$

- Distribution of $\widehat{Y}_h - Y_{h(\mathrm{new})}$:

$$\widehat{Y}_h - Y_{h(\mathrm{new})} \sim N\left(0, \sigma \sqrt{1 + \boldsymbol{X}_h' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}_h}\right).$$

Prediction Intervals for New Observation

- The estimated prediction error variance is

$$\widehat{\sigma}_{\text{pred}} = \hat{\sigma}\sqrt{1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h}.$$

- The $(1-\alpha)$ prediction interval for $Y_{h(\text{new})}$ is

$$\widehat{Y}_h \pm t_{n-p,\alpha/2}\widehat{\sigma}_{\text{pred}}$$

- Note that

$$\frac{Y_h - Y_{h(\text{new})}}{\widehat{\sigma}_{\text{pred}}} \sim T_{n-p}$$

Outline

Model Assumptions

- The relationship between the response variable Y and the explanatory variables $X_1, X_2, \ldots, X_{p-1}$ is

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} \qquad E(\varepsilon_i) = 0$$

- Equal variance:

$$Var(Y_i) = Var(\varepsilon_i) = \sigma^2.$$

- Independence:

$$Cov(Y_i, Y_{i'}) = Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{for} \quad i \neq i'.$$

- Normal distribution:

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}, \sigma^2) \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Model Diagnostics and Remedial Measures

- EDA.
  - Scatter plot matrix.
  - Sampling correlation matrix.
- Residuals: raw, studentized. See L09.pdf for more details.
- Graphical techniques:
  - Plot residuals against $X_{i1}, \ldots, X_{i,p-1}$.
  - Plot residuals against $\hat{Y}_i$.
  - Box plot of residuals.
  - Normal QQ plot of residuals.
- Remedial measures:
  - Transformation.
  - Box-Cox method. See L10.pdf for more details.
  - Weighted least-square.

Outline

A geometric interpretation

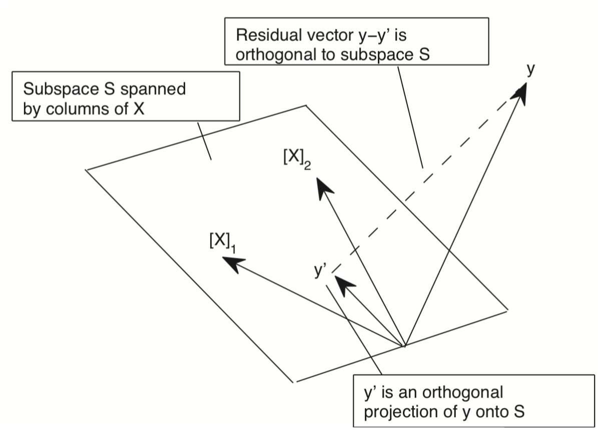
- Recall Least-square cost for linear regression:

$$Q(\beta) = (\boldsymbol{Y} - \beta \boldsymbol{X})'(\boldsymbol{Y} - \beta \boldsymbol{X})$$

- Normal equation (i.e. gradient):

$$\frac{\partial Q(\beta)}{\partial \beta} = 0 \rightarrow \boldsymbol{X}'(\boldsymbol{Y} - \beta \boldsymbol{X}) = 0$$

- Residual $\boldsymbol{e} = \boldsymbol{Y} - \hat{\beta}\boldsymbol{X}$ are orthogonal to columns of $\boldsymbol{X}$
- $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}$ gives the “best” reconstruction of $\boldsymbol{Y}$ in the range of $\boldsymbol{X}$.
- Recall the range of a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the linear space $\subset \mathbb{R}^{p}$ spanned by the columns of $\boldsymbol{X}$.



- Recall "hat matrix": $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, and $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{H}\boldsymbol{Y}$
- $\boldsymbol{H}$ projects $\boldsymbol{Y}$ onto the span of $\boldsymbol{X}$.
- $\boldsymbol{I} - \boldsymbol{H}$ projects $\boldsymbol{Y}$ onto the space orthogonal to $\boldsymbol{X}$.

Outline

ANOVA Approach to Regression Analysis

- The idea is to partition the variation into

$$\text{SS Total} = \text{SS Model} + \text{SS Error}$$

- Why partition the variation?
    -
- In the linear regression, consider three types of partitions.
    - Deviation for each observation.
    - Total sum of squares.
    - Degrees of freedom.

Partitioning Deviation of Each Observation

$$\underbrace{Y_i - \bar{Y}}_{\text{total dev}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{dev of fitted from mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{dev of obs from fitted}} .$$

- If $\{\hat{Y}_i - \bar{Y}\}$ are large in relation to $\{Y_i - \hat{Y}_i\}$: then the regression relation explains (or accounts for) a large proportion of the total variation in $\{Y_i\}$.
- If $\{\hat{Y}_i - \bar{Y}\}$ are small in relation to $\{Y_i - \hat{Y}_i\}$: then the regression relation explains (or accounts for) a small proportion of the total variation in $\{Y_i\}$.

Partitioning Total Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SSTO}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

- The **total sum of squares (SSTO)** is

$$\text{SSTO} \;=\; \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2$$

A measure of total variation in the data (compare to variance).

Partitioning Total Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SSTO}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

- The **regression sum of squares (SSR)** is

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \quad \text{where } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

Partitioning Total Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SSTO}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

- The **error sum of squares (SSE)** is

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \text{SSTO} - \text{SSR}$$

Partitioning Degrees of Freedom

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{df}=n-1} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{df}=1} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{df}=n-2}.$$

- SSTO df $= n - 1$:
  μ_Y is estimated by $\bar{Y}$.
- SSE df $= n - 2$:
  $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are estimated by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$.