STAT 500

Multiple Linear Regression Models

Notation

- $i = 1, \ldots, n$: number of observations.
- Y_i: quantitative response variable
- $x_{i1}, x_{i2}, \ldots, x_{ik}$: k explanatory variables
- Values of $x_{i1}, x_{i2}, \ldots, x_{ik}$ are treated as known and fixed

Research Questions

- Does the MLR model significantly explain the response variable Y_i and how well does it explain the variation in the response variable Y_i?
- Which explanatory variables are significant in the MLR model?
- Which set of explanatory variables are significant in the MLR model?
- What value of the conditional mean of Y_i would we predict for given values of $x_{i1}, x_{i2}, \ldots, x_{ik}$?
- What value of Y_i would we predict for given values of $x_{i1}, x_{i2}, \ldots, x_{ik}$?

Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Assumptions

- The values of the explanatory variables, $x_{i1},\ x_{i2}, \cdots x_{ik}$, are fixed
- $\mu_{Y|x_{i1},x_{i2},\ldots,x_{ik}} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ is the conditional mean of Y given the values of $x_{i1}, x_{i2}, \ldots, x_{ik}$
- additive random errors $\quad Y_i = \mu_{Y|x_{i1},x_{i2},\ldots,x_{ik}} + \epsilon_i$
- independent (uncorrelated) random errors
- homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- normally distributed random errors: $\epsilon_i \sim N(0, \sigma^2)$

Assumptions

- Conditional distribution of $\boldsymbol{Y_i}$ for a given set of values $\boldsymbol{x_{i1}, x_{i2}, \ldots, x_{ik}}$ is

$$\boldsymbol{N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik}, \sigma^2)}$$

- Equivalently, we have $\mathbf{Y} \sim \mathbf{MVN(X\beta, \boldsymbol{\sigma}^2 I_n)}$.

Parameters

- β_j = population coefficient (slope) for explanatory variable x_j
  - Change in the conditional mean of Y for a one unit increase in x_j, *holding all other explanatory variables constant*.
  - Linear effect of x_j on conditional mean of Y *after adjusting for linear effect of the other predictors on* Y *and linear effects of the other explanatory variables on* x_j.
- β_0 = population intercept – the conditional mean of Y when $x_1 = x_2 = \cdots = x_k = 0$.

Parameters

- Interpretation of parameters $\beta_0, \beta_1, \ldots, \beta_k$ depends on the presence or absence of other explanatory variables in the model.

- Example:

  – Model 1: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik} + \epsilon_i$

  – Model 2: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

- Interpretation of parameters β_0, β_1, and β_2 are NOT the same in the two models.

Parameters

- $\boldsymbol{\sigma^2}$ is the variation of responses about the conditional mean of $\boldsymbol{Y}$ for any specific values of $\boldsymbol{x_1, x_2, \ldots, x_k}$.

Least Squares Estimation

Find $\mathbf{b}$, the least squares estimator for $\boldsymbol{\beta}$, that minimizes

$$\begin{aligned} q(\mathrm{b}) &= \sum_{i=1}^{n} (Y_i - b_o - b_1 x_{i1} - \cdots - b_k x_{ik})^2 \\ &= (\mathrm{Y} - X\mathrm{b})^T (\mathrm{Y} - X\mathrm{b}) = \mathrm{e}^T \mathrm{e} \end{aligned}$$

where $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$ is the vector of residuals

- Solve the set of normal equations

$$(X^T X)\mathrm{b} = X^T \mathrm{Y}$$

- Solution: assuming X is of full column rank

$$\mathrm{b} = (X^T X)^{-1} X^T \mathrm{Y}$$

  is the unique solution to the normal equations.

Least Squares Estimation

- $\mathrm{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Y}$ is Best Linear Unbiased Estimator (blue) for $\boldsymbol{\beta}$

$$
\begin{aligned}
\boldsymbol{E}(\mathrm{b}) &= \boldsymbol{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{E}(\mathrm{Y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}
$$

Least Squares Estimation

$$\begin{aligned}
\text{Var}(\mathrm{b}) &= \text{Var}((\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\mathbf{Y}) \\
&= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\text{Var}(\mathbf{Y})\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{\sigma}^{2}\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} \\
&= \boldsymbol{\sigma}^{2}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}
\end{aligned}$$

- For any vector of constants $\boldsymbol{a}^{\boldsymbol{T}} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{\boldsymbol{k}+1})$,

$$\boldsymbol{Var}(\boldsymbol{a}^{\boldsymbol{T}}\mathrm{b}) = \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{Var}(\boldsymbol{b})\boldsymbol{a} = \boldsymbol{\sigma}^{2}\boldsymbol{a}^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{a}$$

  is no larger than $\boldsymbol{Var}(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{b}^*)$ for any other linear, unbiased estimator $\boldsymbol{b}^*$ for $\boldsymbol{\beta}$

Least Squares Estimation

- The derivation of $Var(\mathrm{b}) = \sigma^2(X^TX)^{-1}$
  - Required uncorrelated errors
  - Required homogeneous error variances
  - Did not require a normal distribution for the random errors (normality is needed for inference procedures)
- An unbiased estimator for σ^2 is

$$s_e^2 = MS_{error} = \frac{(\mathrm{Y} - X\mathrm{b})^T(\mathrm{Y} - X\mathrm{b})}{n - (k+1)} = \frac{e^Te}{df_{error}} = \frac{\Sigma\, e_i^2}{df_{error}}$$

- Estimate Var$(\mathrm{b}) = \sigma^2(X^TX)^{-1}$ as $MS_{error}(X^TX)^{-1}$

Least Squares Estimation

- $\hat{Y}_i = \mathrm{x}_i^T \mathrm{b} = b_o + b_1 x_{i1} + \cdots + b_k x_{ik}$ is the fitted value or predicted value

- Then

$$\hat{\mathrm{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \boldsymbol{X}\mathrm{b} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Y} = \boldsymbol{P_X}\mathrm{Y}$$

where $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the orthogonal projection matrix (the perpendicular projection operator) that projects $\mathbf{Y}$ onto the column space of matrix $\boldsymbol{X}$.

Least Squares Estimation

- Given $\hat{\mathbf{Y}} = \boldsymbol{X}\mathbf{b} = \boldsymbol{P_X}\mathbf{Y}$, $\boldsymbol{e_i} = \boldsymbol{Y_i} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ is the $\boldsymbol{i-th}$ residual
- Then $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{P_X Y} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y}$
- The matrix $\boldsymbol{I - P_X}$ projects $\boldsymbol{Y}$ onto the space orthogonal to the column space of $\boldsymbol{X}$ (the residual space) as $\boldsymbol{P_X(I - P_X) = 0}$

ANOVA

- Total variability in response variable

$$SS_{\text{Total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Total variability explained by the model

$$SS_{\text{model}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- Total variability not explained by the model

$$SS_{\text{error}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

ANOVA

- Partition the corrected total sum of squares as

$$\begin{aligned} SS_{\text{Total}} &= \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\ &= SS_{\text{error}} + SS_{\text{model}} \end{aligned}$$

This partitioning is also expressed as

$$Y^T(I - P_1)Y = Y^T(I - P_X)Y + Y^T(P_X - P_1)Y$$

where $P_1 = P_X$ with $X = [1\ 1\ 1\ \cdots\ 1]^T$

ANOVA Table

source of variation	degrees of freedom	sums of squares
model	k	$SS_{\text{model}} = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
error	$n-(k+1)$	$SS_{\text{error}} = \Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
Total	$n-1$	$SS_{\text{Total}} = \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2$

Estimated Error Variance

$$MS_{\text{error}} = \frac{SS_{\text{error}}}{n - (k + 1)}$$

- $E(MS_{\text{error}}) = \sigma^2$ (unbiased estimator)

- $s_e = \sqrt{MS_{\text{error}}}$

Estimated Model Variance

$$MS_{\text{model}} = \frac{SS_{\text{model}}}{k}$$

- $E(MS_{\text{model}}) = \sigma^2 + \frac{\beta^T X^T (P_X - P_1) X\beta}{k}$

- If at least one of the $\beta_j \neq 0, j = 1, \ldots, k$,

$$E(MS_{\text{model}}) > \sigma^2$$

F-test for Significance of Model

- $H_o : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- $H_a :$ at least one $\beta_j \neq 0, j = 1, \ldots, k$
- Test Statistic:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

- Reject H_0 if $F > F_{k,n-(k+1),1-\alpha}$

F-test for Significance of Model

- F-test from ANOVA Table is comparing two models:
  - Model under H_0

$$Y_i = \beta_0 + \epsilon_i$$

  - Model under H_a

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- We almost always reject H_0 in this test.

Coefficient of Determination

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Fraction of variation in the response variable that can be explained by the multiple linear regression model.
- Expressed as percentage: $0\% \leq R^2 \leq 100\%$
- Adding explanatory variables to the model will always increase the value of R^2.

Adjusted R^2

$$\text{adj } R^2 = 1 - \frac{MS_{\text{error}}}{SS_{\text{Total}}/(n-1)}$$

- Expressed as percentage: $0\% \le \text{adj } R^2 \le 100\%$
- Adjusts for the number of explanatory variables in model through degrees of freedom of $MS_{\text{error}} = n - (k+1)$
- Used primarily for model comparisons.

Inference for Population Coefficients

- Test for significance of x_j in model with other explanatory variables
- Two approaches
  - t-test for coefficient
  - Effect test (F-test)
- Results are equivalent

Inference for Population Coefficient

- Least squares estimate for β is $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$
- Any particular b_j is a linear combination of the elements of the vector $\mathbf{Y}$.
- Y_i are normal random variables, meaning that

$$b_j \text{ is } N(\beta_j, \sigma^2 (X^T X)^{-1}_{[j+1,j+1]})$$

where the variance is the $[j+1, j+1]$ element of the matrix $\sigma^2 (X^T X)^{-1}$

Hypothesis Test for Population Coefficient

- Null and Alternative Hypotheses
  $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$

- Test Statistic

$$T = \frac{b_j - 0}{s_e\sqrt{(X^TX)^{-1}_{[j+1,j+1]}}} = \frac{b_j - 0}{S_{b_j}}$$

- Reject H_0 if $|T| > t_{n-(k+1),1-\alpha/2}$

Hypothesis Test for Population Coefficients

- Model under $\boldsymbol{H_0}$

  $$\boldsymbol{Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i,j-1} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_k x_{ik} + \epsilon_i}$$

- Model under $\boldsymbol{H_a}$

  $$\boldsymbol{Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_k x_{ik} + \epsilon_i}$$

- Significance test for $\boldsymbol{x_j}$ depends on presence or absence of other explanatory variables in model.

Confidence Interval for Population Coefficient

- $100(1-\alpha)\%$ CI for β_j is

$$b_j \pm t_{n-(k+1),1-\alpha/2} S_{b_j}$$

Effect Test for Population Coefficient

- Fit two models
  - Model without $\boldsymbol{x_j}$
  - Model with $\boldsymbol{x_j}$
- Compare $\boldsymbol{SS}$error for both models
  - Reduced model without $\boldsymbol{x_j}$: $\boldsymbol{SSE}_{\text{r.model}}$
  - Full model with $\boldsymbol{x_j}$: $\boldsymbol{SSE}_{\text{f.model}}$

Effect Test for Population Coefficient

$$SSE_{\text{r.model}} - SSE_{\text{f.model}}$$

- Amount of error explained by adding x_j to the model.
- The only difference in these two models is the explanatory variable x_j
- Difference has 1 d.f.

Effect Test for Population Coefficient

- Compare amount of error explained to $MSE_{\text{f.model}}$

$$F = \frac{(SSE_{\text{r.model}} - SSE_{\text{f.model}})/1}{MSE_{\text{f.model}}}$$

- Large values of F indicate explanatory variable x_j should be included in the model.

Effect Test for Population Coefficient

- Null and Alternative Hypotheses

$$H_o : \beta_j = 0 \qquad H_a : \beta_j \neq 0$$

- Test Statistic

$$F = \frac{(SSE_{\text{r.model}} - SSE_{\text{f.model}})/1}{MSE_{\text{f.model}}}$$

- Decision - Reject H_o if $F > F_{1,n-(k+1),1-\alpha}$
- Conclusion about x_j is based on other explanatory variables in the model.

Partial F-Test

Effect test for significance of a group of m explanatory variables in the model

- Fit two models
  - Reduced Model without the m explanatory variables (only other $k - m$ explanatory variables)
  - Full Model with the m explanatory variables (plus other $k - m$ explanatory variables)

Partial F-Test

- Compare SSerror for both models
  - Reduced model without m explanatory variables: $SSE_{\text{r.model}}$
  - Full model with m explanatory variables: $SSE_{\text{f.model}}$

Partial F-Test

$$\boldsymbol{SSE}_{\text{r.model}} - \boldsymbol{SSE}_{\text{f.model}}$$

- Amount of error explained by adding the $\boldsymbol{m}$ explanatory variables to the model.
- The only difference in these two models is the $\boldsymbol{m}$ explanatory variables
- Difference has $\boldsymbol{m}$ d.f.

Partial F-Test

- Compare amount of error explained to $\boldsymbol{MSE}_{\text{f.model}}$

$$\boldsymbol{F} = \frac{(\boldsymbol{SSE}_{\text{r.model}} - \boldsymbol{SSE}_{\text{f.model}})/\boldsymbol{m}}{\boldsymbol{MSE}_{\text{f.model}}}$$

- Large values of $\boldsymbol{F}$ indicate group of $\boldsymbol{m}$ explanatory variables should be included in the model.

Partial F-Test

- $H_0 : \beta_j = 0$ for the m explanatory variables
- $H_a :$ at least one $\beta_j \neq 0$ for the m explanatory variables
- Test Statistic

$$F = \frac{(SSE_{\text{r.model}} - SSE_{\text{f.model}})/m}{MSE_{\text{f.model}}}$$

- Decision - Reject H_0 if $F > F_{m,n-(k+1),1-\alpha}$
- Conclusion about the significance of the m explanatory variables depends on the presence of the other $k-m$ explanatory variables in the model.

Inference for Conditional Means

Estimate the conditional mean response $\mu_{Y|\mathbf{x}}$ under specific values for vector $\mathbf{x} = (1, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)^{\mathbf{T}}$

- Point estimate is $\hat{\mu}_{Y|\mathbf{x}} = \mathbf{x}^{T}\hat{\beta}$

- Std error is $S_{\hat{\mu}_{Y|\mathbf{x}}} = \sqrt{MS_{\text{error}}\ \mathbf{x}^{\mathbf{T}}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{x}}$

- A $(1-\alpha) \times 100\%$ confidence interval for $\mu_{Y|\mathbf{x}}$ is

$$\hat{\mu}_{Y|\mathbf{x}} \pm t_{n-(k+1),1-\alpha/2}\ S_{\hat{\mu}_{Y|\mathbf{x}}}$$

- Simultaneous confidence region for an entire line segment (the Scheffe's method) is

$$\hat{Y} \pm \sqrt{(k+1)F_{k+1,n-k-1,1-\alpha}}\ S_{\hat{\mu}_{Y|\mathbf{x}}}$$

Prediction Intervals

Predict value of $Y_i = \mathrm{x}^T \beta + \epsilon_i$ that will be observed under specific values for vector $\mathbf{x} = (1, \mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_k)^{\mathrm{T}}$

- The predictor is $\hat{Y}_i = \mathrm{x}^{\mathrm{T}} \hat{\beta}$
- The standard error for the predictor is $S_{\hat{Y}} = \sqrt{MS_{\text{error}} + S^2_{\hat{\mu}_{Y|\mathrm{x}}}}$
- A $(1-\alpha) \times 100\%$ prediction interval is

$$\hat{Y}_i \pm t_{n-(k+1),1-\alpha/2} S_{\hat{Y}}$$