Lecture 7 Multivariate Linear Regression

Shiwei Lan[1]

[1]School of Mathematical and Statistical Sciences
Arizona State University

STP533 Multivariate Analysis
Spring 2025

Table of Contents

Regression Analysis

- Regression analysis is the statistical methodology for predicting values of one or more *response* (dependent) variables from a collection of *predictor* (independent) variables.
- It can also be used for assessing the effects of the predictor variables on the responses.
- The name *regresion*, dated back to 1885 by F. Galton.
- We first review the classical linear regression model with a single response. Then we generalize to linear model fir several dependent variables.

The Classical Linear Regression

- Suppose we have p predictor variables $X_1, \cdots, X_p$ and a response variable Y.
- For example, Y=current market value of a house, X_1=square feet, X_2=location, X_3=appraised value of last year, and X_4=quality of construction.
- A classical linear regression relates the average value of Y with a linear combination of X_i's.

$$Y_i = \beta_0 + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \epsilon_i, \quad i = 1, \cdots, n,$$

  where we assume $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$.
- If we denote $\mathbf{Y} = [Y_1, \cdots, Y_n]^T$, $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$, and $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_p]^T$, then we can rewrite

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

One-Way ANOVA as A Regression

Example 7.2 (The design matrix for one-way ANOVA as a regression model) Determine the design matrix if the linear regression model is applied to the one-way ANOVA situation in Example 6.6.

We create so-called *dummy* variables to handle the three population means: $\mu_1 = \mu + \tau_1$, $\mu_2 = \mu + \tau_2$, and $\mu_3 = \mu + \tau_3$. We set

$$z_1 = \begin{cases} 1 & \text{if the observation is from population 1} \\ 0 & \text{otherwise} \end{cases} \qquad z_2 = \begin{cases} 1 & \text{if the observation is from population 2} \\ 0 & \text{otherwise} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{if the observation is from population 3} \\ 0 & \text{otherwise} \end{cases}$$

and $\beta_0 = \mu$, $\beta_1 = \tau_1$, $\beta_2 = \tau_2$, $\beta_3 = \tau_3$. Then

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \beta_3 z_{j3} + \varepsilon_j, \qquad j = 1, 2, \ldots, 8$$

where we arrange the observations from the three populations in sequence. Thus, we obtain the observed response vector and design matrix

$$\underset{(8\times 1)}{\mathbf{Y}} = \begin{bmatrix} 9 \\ 6 \\ 9 \\ 0 \\ 2 \\ 3 \\ 1 \\ 2 \end{bmatrix}; \qquad \underset{(8\times 4)}{\mathbf{Z}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

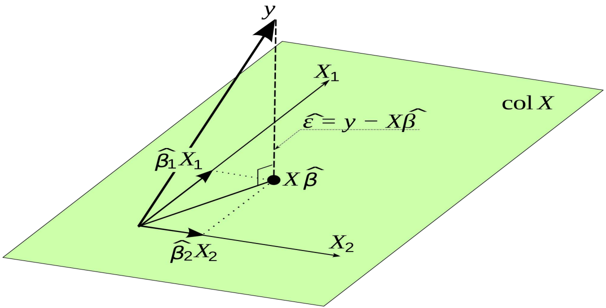
Least Square Estimation

- The least square estimation (LSE) minimizes the sum of square $S(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ with respect to β.
- Let $\mathbf{X}$ be full rank $p + 1 \leq n$. The LSE result of β is $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
- Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}$ be the *fitted values* of $\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called *hat matrix*.
- The residual vector can now be written

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

- The residual sum of squares becomes $S(\hat{\beta}) = \|\mathbf{e}\|_2^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}$.

Sum-of-Squares Decomposition



- Note $\mathbf{X} \perp \mathbf{e}$ and $\hat{\mathbf{y}} \perp \mathbf{e}$. Why?
- Then we have $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$.
- Further we have decomposition of the sum of squares about mean

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE}$$

Coefficient of Determination

- The decomposition of the sum of squares can also be written as $\mathbf{y}^T(\mathbf{I}-\mathbf{J})\mathbf{y} = \mathbf{y}^T(\mathbf{H}-\mathbf{J})\mathbf{y} + \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}$.

- We define the *coefficient of determination* as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- This quantity measure the proportion of the total variation in y's "explained" by the model with p predictors $\mathbf{X}$.
- If we plot $\hat{\mathbf{y}}$ against $\mathbf{y}$, what is the slope?

Sampling Properties of LSE

- We have the following property for LSE $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\mathrm{E}[\hat{\beta}] = \beta, \quad \mathrm{Cov}[\hat{\beta}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ has the following property

$$\mathrm{E}[\mathbf{e}] = \mathbf{0}, \quad \mathrm{Cov}[\mathbf{e}] = \sigma^2[\mathbf{I} - \mathbf{H}]$$

Inferences about the Regression Parameters

- Now we consider $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.
- Then the maximum likelihood estimator (MLE) of β is the same as LSE $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Moreover, we have

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

- The residual $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is independent of $\hat{\beta}$ and $SSE/n = \|\mathbf{e}\|^2/n$ is the MLE of σ^2. Moreover,

$$\frac{\|\mathbf{e}\|^2}{\sigma^2} \sim \chi^2(n-p-1).$$

- $MSE = \frac{SSE}{n-p-1} = \frac{\|\mathbf{e}\|^2}{n-p-1} =: s^2$ is an unbiased estimator of σ^2.

Inferences about the Regression Parameters

- $100(1-\alpha)\%$ CR for $\boldsymbol{\beta}$ is determined by

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq (p+1)s^2 F_{1-\alpha}(p+1, n-p-1).$$

- The $100(1-\alpha)\%$ SCI for β_j's are given by

$$\hat{\beta}_j \pm \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}\sqrt{(p+1)F_{1-\alpha}(p+1, n-p-1)}, \quad j = 0, 1, \cdots, p.$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ is the j-th diagonal element of $s^2(\mathbf{X}^T\mathbf{X})^{-1}$.

- For each β_j, the $100(1-\alpha)\%$ individual CI is

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-p-1)\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}, \quad j = 0, 1, \cdots, p.$$

Likelihood Ratio Tests for the Regression Parameters

- Suppose you hypothesize that only the first $q \leq p$ predictors are significant in explaining the response variable.
- We want to test $H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$. Denote $\boldsymbol{\beta}_2 = [\beta_{q+1}, \cdots, \beta_p]^T$.
- We devide $\mathbf{X} = [(\mathbf{X}_1)_{n\times(q+1)} | (\mathbf{X}_2)_{n\times(p-q)}]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T]^T$. Then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \epsilon$$

- The LRT rejects $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ if

$$\frac{(SSE(\mathbf{X}_1) - SSE(\mathbf{X}))/(p-q)}{s^2} > F_{1-\alpha}(p-q, n-p-1).$$

Likelihood Ratio Tests for the Regression Parameters

Example 7.5 (Testing the importance of additional predictors using the extra sum-of-squares approach) Male and female patrons rated the service in three establishments (locations) of a large restaurant chain. The service ratings were converted into an index. Table 7.2 contains the data for $n = 18$ customers. Each data point in the table is categorized according to location (1, 2, or 3) and gender (male $= 0$ and female $= 1$). This categorization has the format of a two-way table with unequal numbers of observations per cell. For instance, the combination of location 1 and male has 5 responses, while the combination of location 2 and female has 2 responses. Introducing three dummy variables to account for location and two dummy variables to account for gender, we can develop a regression model linking the service index Y to location, gender, and their "interaction" using the design matrix

Table 7.2 Restaurant-Service Data

Location	Gender	Service (Y)
1	0	15.2
1	0	21.2
1	0	27.3
1	0	21.2
1	0	21.2
1	1	36.4
1	1	92.4
2	0	27.3
2	0	15.2
2	0	9.1
2	0	18.2
2	0	50.0
2	1	44.0
2	1	63.6
3	0	15.2
3	0	30.3
3	1	36.4
3	1	40.9

Table of Contents

Multivariate Multiple Regression

- Now we consider the problem of modeling the relationship between m responses $Y_1, Y_2, \cdots, Y_m$ and p predictor variables $X_1, \cdots, X_p$.:

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \epsilon_k, \quad k = 1, \cdots, m$$

- Denote $\mathbf{Y} = [y_{ik}]_{n\times m}$, $\boldsymbol{\epsilon} = [\epsilon_{ik}]_{n\times m}$ and $\boldsymbol{\beta} = [\beta_{jk}]_{(p+1)\times m}$. The *multivariate linear regression model* is

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

where we assume $\mathrm{E}[\boldsymbol{\epsilon}_k] = \mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}_k, \boldsymbol{\epsilon}_{k'}] = \sigma_{kk'}\mathbf{I}_n$. Denote the inter-trial covariance as $\boldsymbol{\Sigma} = [\sigma_{kk'}]_{m\times m}$.

Least Square Estimation

- Viewing the multivariate linear regression as m parallel classical regression, we get LSE for each

$$\hat{\beta}_k = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}_k, \quad k = 1, \cdots, m$$

- Denote $\hat{\beta} = [\hat{\beta}_1, \cdots, \hat{\beta}_m]$. We have $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
- The predicted values and residuals now become

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{Y}, \quad \mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$$

- Then we have

$$\mathbf{X}^T\mathbf{E} = \mathbf{0}, \quad \hat{\mathbf{Y}}^T\mathbf{E} = \mathbf{0}$$

- Therefore the decomposition of sum of squares

$$\mathbf{Y}^T\mathbf{Y} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + \mathbf{E}^T\mathbf{E}$$

Fitting a Multivariate Linear Model

Example 7.8 (Fitting a multivariate straight-line regression model) To illustrate the calculations of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{Y}}$, and $\hat{\boldsymbol{\varepsilon}}$, we fit a straight-line regression model (see Panel 7.2),

$$Y_{j1} = \beta_{01} + \beta_{11} z_{j1} + \varepsilon_{j1}$$

$$Y_{j2} = \beta_{02} + \beta_{12} z_{j1} + \varepsilon_{j2}, \qquad j = 1, 2, \ldots, 5$$

to two responses Y_1 and Y_2 using the data in Example 7.3. These data, augmented by observations on an additional response, are as follows:

z_1	0	1	2	3	4
y_1	1	4	3	8	9
y_2	−1	−1	2	3	2

The design matrix $\mathbf{Z}$ remains unchanged from the single-response problem. We find that

$$\mathbf{Z}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \qquad (\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix}$$

Sampling Properties of LSE

- We have the following property for LSE $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \quad \mathrm{Cov}[\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\beta}}_{k'}] = \sigma_{kk'}(\mathbf{X}^T\mathbf{X})^{-1}$$

- The residual $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ has the following property

$$\mathrm{E}[\mathbf{E}] = \mathbf{0}, \quad \mathrm{Cov}[\mathbf{e}_k, \mathbf{e}_{k'}] = \sigma_{kk'}(n-p-1), \quad \mathrm{E}[\mathbf{E}^T\mathbf{E}]/(n-p-1) = \boldsymbol{\Sigma}.$$

- Moreover, $\mathrm{Cov}[\hat{\boldsymbol{\beta}}_k, \mathbf{e}_{k'}] = \mathbf{0}$.

Matrix Valued Normal Distribution

- So far we have not imposed any distribution assumption.
- For the following inference, we need the matrix valued normal distribution.

Definition

A random matrix $\mathbf{Y}_{n\times m}$ follows *matrix normal distribution*, $\mathcal{MN}(\boldsymbol{\mu}, \mathbf{C}_{n\times n}, \boldsymbol{\Sigma}_{m\times m})$, if $\mathrm{vec}(\mathbf{Y}) \sim N_{nm}(\mathrm{vec}(\mathbf{u}), \boldsymbol{\Sigma}\otimes\mathbf{C})$, where $\mathrm{vec}(\mathbf{Y}) = [\mathbf{Y}_1^T, \mathbf{Y}_2^T, \cdots, \mathbf{Y}_m^T]^T$.

- The density of $\mathbf{Y} \sim \mathcal{MN}(\boldsymbol{\mu}, \mathbf{C}_{n\times n}, \boldsymbol{\Sigma}_{m\times m})$ is

$$(2\pi)^{-mn/2}|\mathbf{C}|^{-m/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{Y}-\boldsymbol{\mu})]\right\}.$$

- In particular, we assume $\mathbf{C} = \mathbf{I}_n$. And $\boldsymbol{\Sigma}_{m\times m}$ is the inter-trial covariance.

Likelihood Ratio Tests for the Regression Parameters

- Now we consider the similar hypothesis test in the multivariate case $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}_{(p-q)\times m}$.
- The LRT involves the extra sum of squares and cross products

$$(\mathbf{Y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1)^T(\mathbf{Y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1) - (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$$

- The LRT statistic is defined as

$$\Lambda = \frac{\max_{\boldsymbol{\beta}_1,\boldsymbol{\Sigma}} L(\boldsymbol{\beta}_1,\boldsymbol{\Sigma})}{\max_{\boldsymbol{\beta},\boldsymbol{\Sigma}} L(\boldsymbol{\beta},\boldsymbol{\Sigma})} = \frac{L(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\Sigma}})}{L(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\Sigma}})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)^{n/2}$$

- The corresponding Wilk's lambda can be used in the following test statistic

$$-\left[n - p - 1 - \frac{1}{2}(m - p + q + 1)\right] \log\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|})\right) \dot{\sim} \chi^2(m(p-q)).$$

Testing for Additional Predictors

Example 7.9 (Testing the importance of additional predictors with a multivariate response) The service in three locations of a large restaurant chain was rated according to two measures of quality by male and female patrons. The first service-quality index was introduced in Example 7.5. Suppose we consider a regression model that allows for the effects of location, gender, and the location–gender interaction on both service-quality indices. The design matrix (see Example 7.5) remains the same for the two-response situation. We shall illustrate the test of no location-gender interaction in either response using Result 7.11. A computer program provides

$$\begin{pmatrix} \text{residual sum of squares} \\ \text{and cross products} \end{pmatrix} = n\hat{\Sigma} = \begin{bmatrix} 2977.39 & 1021.72 \\ 1021.72 & 2050.95 \end{bmatrix}$$

$$\begin{pmatrix} \text{extra sum of squares} \\ \text{and cross products} \end{pmatrix} = n(\hat{\Sigma}_1 - \hat{\Sigma}) = \begin{bmatrix} 441.76 & 246.16 \\ 246.16 & 366.12 \end{bmatrix}$$

Prediction from Multivariate Multiple Regressions

- Recall we have $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\epsilon \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Sigma})$.
- The task is to predict the mean response corresponding to $\mathbf{x}_0$. Note

$$\hat{\boldsymbol{\beta}}^T \mathbf{x}_0 \sim N_m(\boldsymbol{\beta}^T \mathbf{x}_0, \mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0 \mathbf{\Sigma}), \quad \perp \quad n\hat{\mathbf{\Sigma}} \sim W_{n-p-1}(\mathbf{\Sigma}).$$

- Therefore we have the T^2-statistic

$$T^2 = \left(\frac{\hat{\boldsymbol{\beta}}^T \mathbf{x}_0 - \boldsymbol{\beta}^T \mathbf{x}_0}{\sqrt{\mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \right)^T \left(\frac{n}{n-p-1} \hat{\mathbf{\Sigma}} \right)^{-1} \left(\frac{\hat{\boldsymbol{\beta}}^T \mathbf{x}_0 - \boldsymbol{\beta}^T \mathbf{x}_0}{\sqrt{\mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \right) \sim \frac{m(n-p-1)}{n-p-m} F(m, n-p-m)$$

- The $100(1-\alpha)\%$ CR is determined by $T^2 \le \frac{m(n-p-1)}{n-p-m} F_{1-\alpha}(m, n-p-m)$.
- The $100(1-\alpha)\%$ SCI for $\mathrm{E}[Y_j] = \mathbf{x}_0^T \beta_j$'s are

$$\mathbf{x}_0^T \hat{\beta}_j \pm \sqrt{\frac{m(n-p-1)}{n-p-m} F_{1-\alpha}(m, n-p-m)} \sqrt{\mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0 \left(\frac{n}{n-p-1} \hat{\sigma}_{jj} \right)}, \quad j = 1, \cdots, m.$$

Prediction from Multivariate Multiple Regressions

- To predict the new response corresponding to $\mathbf{x}_0$, i.e. $\mathbf{Y}_0 = \boldsymbol{\beta}^T\mathbf{x}_0 + \boldsymbol{\epsilon}_0$, we note

$$\mathbf{Y}_0 - \hat{\boldsymbol{\beta}}^T\mathbf{x}_0 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\mathbf{x}_0 + \boldsymbol{\epsilon}_0 \sim N_m(\mathbf{0}, (1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)\boldsymbol{\Sigma}).$$

- Therefore, the $100(1-\alpha)\%$ CR for $\mathbf{Y}_0$ becomes

$$\left(\mathbf{Y}_0 - \hat{\boldsymbol{\beta}}^T\mathbf{x}_0\right)^T \left(\frac{n}{n-p-1}\hat{\boldsymbol{\Sigma}}\right)^{-1} \left(\mathbf{Y}_0 - \hat{\boldsymbol{\beta}}^T\mathbf{x}_0\right) \le (1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)\frac{m(n-p-1)}{n-p-m}F_{1-\alpha}(m, n-p-m)$$

- The $100(1-\alpha)\%$ SCI for Y_{0j}'s are

$$\mathbf{x}_0^T\hat{\beta}_j \pm \sqrt{\frac{m(n-p-1)}{n-p-m}F_{1-\alpha}(m, n-p-m)}\sqrt{(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)\left(\frac{n}{n-p-1}\hat{\sigma}_{jj}\right)}, \quad j = 1, \cdots, m.$$

Prediction from Multivariate Multiple Regressions

Example 7.10 (Constructing a confidence ellipse and a prediction ellipse for bivariate responses) A second response variable was measured for the computer-requirement problem discussed in Example 7.6. Measurements on the response Y_2, disk input/output capacity, corresponding to the z_1 and z_2 values in that example were

$$\mathbf{y}_2' = [301.8, 396.1, 328.2, 307.4, 362.4, 369.5, 229.1]$$

Obtain the 95% confidence ellipse for $\boldsymbol{\beta}'\mathbf{z}_0$ and the 95% prediction ellipse for $\mathbf{Y}_0' = [Y_{01}, Y_{02}]$ for a site with the configuration $\mathbf{z}_0' = [1, 130, 7.5]$.

Computer calculations provide the fitted equation

$$\hat{y}_2 = 14.14 + 2.25z_1 + 5.67z_2$$

with $s = 1.812$. Thus, $\hat{\boldsymbol{\beta}}_{(2)}' = [14.14, 2.25, 5.67]$. From Example 7.6,

$$\hat{\boldsymbol{\beta}}_{(1)}' = [8.42, 1.08, 42], \quad \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(1)} = 151.97, \quad \text{and} \quad \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 = .34725$$

We find that

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(2)} = 14.14 + 2.25(130) + 5.67(7.5) = 349.17$$

and

$$n\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} (\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)}) & (\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)}) \\ (\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)})'(\mathbf{y}_{(1)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(1)}) & (\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)})'(\mathbf{y}_{(2)} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{(2)}) \end{bmatrix}$$

$$= \begin{bmatrix} 5.80 & 5.30 \\ 5.30 & 13.13 \end{bmatrix}$$

Since

$$\hat{\boldsymbol{\beta}}'\mathbf{z}_0 = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{(1)}' \\ \hline \hat{\boldsymbol{\beta}}_{(2)}' \end{bmatrix}\mathbf{z}_0 = \begin{bmatrix} \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(1)} \\ \hline \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(2)} \end{bmatrix} = \begin{bmatrix} 151.97 \\ 349.17 \end{bmatrix}$$

Linear Regression as Conditional Gaussians

- The multivariate linear regression can also be understood from the perspective of conditional Gaussian.
- Suppose $Y, \mathbf{X}$ jointly follow a multivariate normal $N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \boldsymbol{\mu}_{\mathbf{X}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{YY} & \boldsymbol{\sigma}_{Y\mathbf{X}} \\ \boldsymbol{\sigma}_{\mathbf{X}Y} & \boldsymbol{\Sigma}_{\mathbf{XX}} \end{bmatrix}$$

- Now we want to predict (explain) Y with linear function of $\mathbf{X}$, i.e. $\beta_0 + \boldsymbol{\beta}^T\mathbf{X}$.
- The solution with minimum square error is the $\mathrm{E}[Y|\mathbf{X}]$, which is by conditional Gaussian

$$\beta_0 + \boldsymbol{\beta}^T\mathbf{X} = \mathrm{E}[Y|\mathbf{X}] = \mu_Y + \boldsymbol{\sigma}_{Y\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$$

- Therefore we have

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\sigma}_{\mathbf{X}Y}, \; \beta_0 = \mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\mathbf{X}}, \quad \mathrm{E}[Y - (\beta_0 + \boldsymbol{\beta}^T\mathbf{X})]^2 = \sigma_{YY} - \boldsymbol{\sigma}_{Y\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\sigma}_{\mathbf{X}Y}.$$

Linear Regression as Conditional Gaussians

- Consider the predictor $\hat{Y} = \beta_0 + \boldsymbol{\beta}^T \mathbf{X} = \mu_Y + \boldsymbol{\sigma}_{Y\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} (\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$.
- What is $\mathrm{Corr}[Y, \hat{Y}]$?
- $\hat{Y}$ is the best linear unbiased estimator (BLUE). Why?
- But we do not know $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$? Maximum likelihood estimator (MLE)!

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \bar{Y} \\ \bar{\mathbf{X}} \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}} = \frac{n-1}{n} \mathbf{S}, \quad \mathbf{S} = \begin{bmatrix} s_{YY} & \mathbf{s}_{Y\mathbf{X}} \\ \mathbf{s}_{\mathbf{X}Y} & \mathbf{S}_{\mathbf{XX}} \end{bmatrix}$$

- Therefore we have the MLE of the regression coefficients

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{XX}}^{-1} \mathbf{s}_{\mathbf{X}Y}, \quad \hat{\beta}_0 = \bar{Y} - \mathbf{s}_{Y\mathbf{X}} \mathbf{S}_{\mathbf{XX}}^{-1} \bar{\mathbf{X}} = \bar{Y} - \hat{\boldsymbol{\beta}}^T \bar{\mathbf{X}},$$

$$\widehat{MSE} = \frac{n-1}{n} (s_{YY} - \mathbf{s}_{Y\mathbf{X}} \mathbf{S}_{\mathbf{XX}}^{-1} \mathbf{s}_{\mathbf{X}Y}).$$

Prediction of Multiple Response Variables

- The extension to multivariate multiple regression is straightforward.
- Suppose $\mathbf{Y}, \mathbf{X}$ jointly follow a multivariate normal $N_{m+p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{Y}} \\ \boldsymbol{\mu}_{\mathbf{X}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{YY}} & \boldsymbol{\Sigma}_{\mathbf{YX}} \\ \boldsymbol{\Sigma}_{\mathbf{XY}} & \boldsymbol{\Sigma}_{\mathbf{XX}} \end{bmatrix}$$

- The best (minimum square error) linear predictor of $\mathbf{Y}$, $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_{m\times p}\mathbf{X}$, is given by the conditional Gaussian $\mathrm{E}[\mathbf{Y}|\mathbf{X}]$

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{X} = \mathrm{E}[\mathbf{Y}|\mathbf{X}] = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$$

- Denote $\mathbf{e} = \mathbf{Y} - (\boldsymbol{\beta}_0 + \boldsymbol{\beta}\mathbf{X})$. Therefore we have

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}, \quad \boldsymbol{\beta}_0 = \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\beta}\boldsymbol{\mu}_{\mathbf{X}}, \quad \mathrm{E}[\mathbf{e}\mathbf{e}^T] = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}.$$

Prediction of Multiple Response Variables

- To compute the predictor $\hat{\mathbf{Y}} = \beta_0 + \beta \mathbf{X} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$, we substitute the parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with their MLEs

$$\hat{\mu} = \begin{bmatrix} \bar{\mathbf{Y}} \\ \bar{\mathbf{X}} \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S}, \quad \mathbf{S} = \begin{bmatrix} \mathbf{S}_{\mathbf{YY}} & \mathbf{S}_{\mathbf{YX}} \\ \mathbf{S}_{\mathbf{XY}} & \mathbf{S}_{\mathbf{XX}} \end{bmatrix}$$

- Therefore we have the MLE of the regression coefficients

$$\hat{\beta} = \mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{XY}}, \quad \hat{\beta}_0 = \bar{\mathbf{Y}} - \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\bar{\mathbf{X}} = \bar{\mathbf{Y}} - \hat{\beta}\bar{\mathbf{X}},$$

$$\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}} = \frac{n-1}{n}(\mathbf{S}_{YY} - \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{XY}}).$$

- The *partial correlation coefficient*, $\rho_{Y_1 Y_2|\mathbf{X}} = \frac{\sigma_{Y_1 Y_2|\mathbf{X}}}{\sqrt{\sigma_{Y_1 Y_1|\mathbf{X}}}\sqrt{\sigma_{Y_2 Y_2|\mathbf{X}}}}$, is used to measure the association between Y_1 and Y_2 after eliminating the effects of $\mathbf{X}$.