FALL 2013
Stat 8003: Statistical Methods I
Lecture 11

Jichun Xie

1 Multiple Linear Regression

- SLR: One independent variable (X)
- MLR: More than one independent variables ($X_1, \ldots, X_p$)

The data we observe are

$$(Y_i, X_{i1}, \ldots, X_{ip}), \quad i = 1, \ldots, n$$

And the model is

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

Now let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

The the model could be reformatted intro the matrix form:

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

where $\mathtt{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathtt{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

1.1 LSE for Multiple Linear Regression

Suppose $\mathbf{X}$ is full rank, *i.e.* $\texttt{Rank}(\mathbf{X}) = p + 1$. For the rest of the following notes, unless otherwise specified, we always assume this assumption is true.

In last lecture, we discussed SLR and the least-square estimator (LSE) for SLR. Today, we use matrix notations to derive LSE.

Define $Q(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. What is $Q(\boldsymbol{\beta})$ if you write it into the entry-wise form?

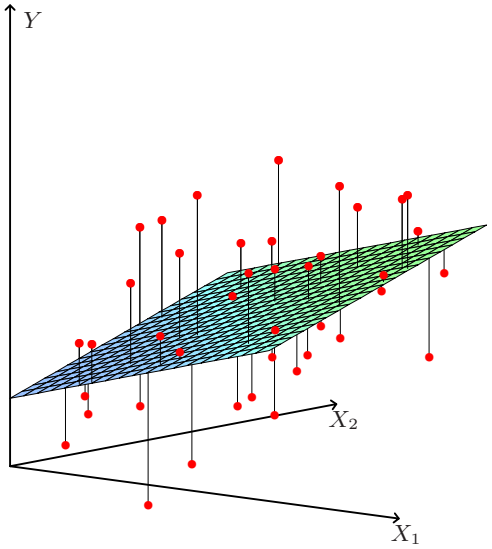

Figure 1: Linear least squares fitting with $\mathbf{X} \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y. ©Hastie, Tibishirani & Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

We want to minimize $Q(\boldsymbol{\beta})$. Take the derivative and set it to zero.

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}) \\ &= 0 - 2\mathbf{X}^{\mathrm{T}}\mathbf{Y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Set $\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$. We have

$$\text{Normal equation:} \quad \mathbf{X}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

And the LSE

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

Check if $\hat{\boldsymbol{\beta}}$ is the minimizer:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

is positive definite.

Geometric Interpretation for $\hat{\boldsymbol{\beta}}$

Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Geometrically, we want to choose $\hat{\boldsymbol{\beta}}$ so that the distance between $\mathbf{Y}$ and the column space $C(\mathbf{X})$ is minimized. Consequently, $\hat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$.

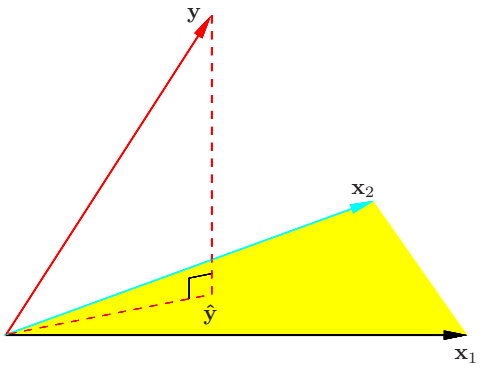

Figure 2: The n-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors x_1 and x_2. The projection $\hat{y}$ represents the vector of the least squares predictions. ©Hastie, Tibishirani & Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

The projection of $\mathbf{Y}$ is $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{P}_X\mathbf{Y}$. Here $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is called the projection matrix of $C(\mathbf{X})$, which can project any vector to the column space $C(\mathbf{X})$.

Geometrically, it is easy to see that

1. If $\mathbf{P}$ is a projection matrix, then $\mathbf{I} - \mathbf{P}$ is also a projection matrix.
2. $\mathbf{P}$ is idempotent, *i.e.* $\mathbf{P} = \mathbf{P}^2$.

Let's go back to the linear regression. The following statements are true.

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}_X\mathbf{Y} \in C(\mathbf{X})$.
- $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y} \perp C(\mathbf{X})$.
- If $\mathbf{Y} \in C(\mathbf{X})$, then $\hat{\mathbf{Y}} = \mathbf{Y}$.

1.2 Estimating σ^2

Residual sum of squares (RSS):

$$\begin{aligned} RSS =& \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ =& (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}}(\mathbf{Y} - \hat{\mathbf{Y}}) \\ =& (\mathbf{Y} - \mathbf{P}_X\mathbf{Y})^{\mathrm{T}}(\mathbf{Y} - \mathbf{P}_X\mathbf{Y}) \\ =& \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y} \\ =& \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}. \end{aligned}$$

The estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)} = \frac{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}}{n-(p+1)}.$$

Properties of $\hat{\boldsymbol{\beta}}$

1. $\hat{\boldsymbol{\beta}}$ is unbiased, $\mathtt{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
2. $\mathtt{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.
3. Gauss-Markov Theorem: Among all the unbiased linear estimators of $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$, the estimator $\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ has the minimum variance.
4. If $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I})$, the MLE is equal to the LSE $\hat{\boldsymbol{\beta}}$, the MLE of σ^2 is RSS/n. (Homework)

1. $\hat{\boldsymbol{\beta}}$ is unbiased.

$$\mathtt{E}(\hat{\boldsymbol{\beta}}) = \mathtt{E}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathtt{E}(\mathbf{Y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

2. $\mathtt{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

$$\begin{aligned} \mathtt{Var}(\hat{\boldsymbol{\beta}}) &= \mathtt{Var}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\} \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathtt{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathtt{Var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \end{aligned}$$

3. Proof of Gauss-Markov Theorem. Suppose $\mathbf{l}^{\mathrm{T}}\mathbf{Y}$ is a linear estimator of $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$. Then $\mathtt{E}(\mathbf{l}^{\mathrm{T}}\mathbf{Y}) = \mathbf{l}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$. The statement is true for all $\boldsymbol{\beta}$; and therefore, $\mathbf{l}^{\mathrm{T}}\mathbf{X} = \mathbf{q}^{\mathrm{T}}$. Now we

compare $\texttt{Var}(\mathbf{l}^{\mathrm{T}}\mathbf{Y})$ and $\texttt{Var}(\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$.

$$\begin{aligned}
&\texttt{Var}(\mathbf{l}^{\mathrm{T}}\mathbf{Y}) - \texttt{Var}(\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \\
=&\sigma^2\mathbf{l}^{\mathrm{T}}\mathbf{l} - \sigma^2\mathbf{q}^{\mathrm{T}}\texttt{Var}(\hat{\boldsymbol{\beta}})\mathbf{q} \\
=&\sigma^2\mathbf{l}^{\mathrm{T}}\mathbf{l} - \sigma^2\mathbf{l}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{l} \\
=&\sigma^2\mathbf{l}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{l} \\
=&\sigma^2\mathbf{l}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{l} \\
=&\sigma^2\{(\mathbf{I} - \mathbf{P}_X)\mathbf{l}\}^{\mathrm{T}}\{(\mathbf{I} - \mathbf{P}_X)\mathbf{l}\} \geq 0.
\end{aligned}$$

Therefore, $\texttt{Var}(\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \leq \texttt{Var}(\mathbf{l}^{\mathrm{T}}\mathbf{Y})$.

4. Multivariate normal distribution: $\mathbf{Y}_{n\times 1} \sim MVN(\boldsymbol{\mu}_{n\times 1}, \boldsymbol{\Sigma}_{n\times n})$. Then $\mathbf{Y}$ has the p.d.f:

$$f(\mathbf{y}_{n\times 1}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu} = \texttt{E}(\mathbf{Y})$ and $\boldsymbol{\Sigma} = \texttt{Cov}(\mathbf{Y})$.

In the linear regression model:

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

Then $\mathbf{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

What is the likelihood function of $\boldsymbol{\beta}$ and σ^2? What are the MLEs? (Homework)

1.3 The Link between MLR and SLR

What is LS in MLR doing? Define

$$\mathbf{X}_{-j} = (X_0, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$$

Regression by Successive Orthogonalization:

1. Regress $\mathbf{Y}$ on $\mathbf{X}_{-j}$ $\Rightarrow$ Save residuals

$$\texttt{Resid}(Y \mid \mathbf{X}_{-j}) = Y_i - \sum_{k\neq j}\hat{\beta}_k X_k$$

2. Regress X_j on X_{-j} $\Rightarrow$ Save residuals

$$\texttt{Resid}(X_j \mid \mathbf{X}_{-j}) = X_j - \sum_{k\neq j}\hat{\alpha}_k X_k$$

3. Regress $\texttt{Resid}(Y \mid \mathbf{X}_{-j})$ on $\texttt{Resid}(X_j \mid \mathbf{X}_{-j})$, then (Homework)

   - Slope from this regression = Slope for X_j in MLR.
   - RSS from this regression = RSS in MLR.

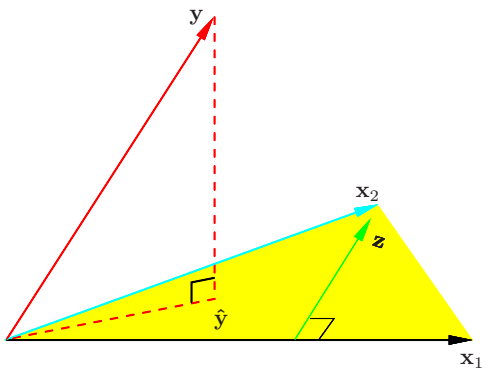

Figure 3: Least squares regression by orthogonalization of the inputs. The vector x_2 is regressed on the vector x_1, leaving the residual vector z. The regression of y on z gives the multiple regression coefficient of x_2. Adding together the projections of y on each of x_1 and z gives the least squares fit $\hat{y}$. ©Hastie, Tibishirani & Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

1.4 Hypothesis Testing and Confidence Interval of β_j

When $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}) \Rightarrow \hat{\beta}_j \sim N(\beta_j, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{jj})$.

Now suppose $\texttt{Rank}(X) = p$. Though σ^2 is unknown, we can estimate it with

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p} = \frac{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}}{n-p}.$$

Now the test statistic

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{jj}}} \sim T_{n-p}.$$

Why?

Consider the hypothesis testing:

$$\mathrm{H}_0 : \ \beta_j = 0 \quad \textit{vs.} \quad \mathrm{H}_1 : \ \beta_j \neq 0$$

It turns out the GLR test leads to rejecting H_0 when $\left|\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{jj}}}\right|$ is large.

Therefore, to control type I error at level α, we reject H_0 if

$$\left|\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{jj}}}\right| > t^{-1}_{n-p}(1-\alpha/2).$$

Also, based on the pivot method, we can construct a $(1-\alpha)$ CI of $\boldsymbol{\beta}_j$:

$$\hat{\beta}_j \pm t^{-1}_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{jj}}$$

1.5 Inference about $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$

Why are we interested in $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$?

Example: In a clinical trial, the patients with obesity are randomized. The bmi of the patients before and after the treatment are measured. Now the researchers are interested in whether the effect of Treatment B is the same as Treatment A.

We can use two sample t-test to solve the problem. However, let's see another solution.

Model:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i,\ i = 1, \ldots, n.$$

If the ith patient is in Group A, then $X_{i1} = 1$, $X_{i2} = 0$; otherwise, $X_{i1} = 0$ and $X_{i2} = 1$. Y_i is the difference of bmi before and after the treatment. There is no intercept term in the model.

To address the researchers' problem, we need to build the hypothesis:

$$\mathrm{H}_0 :\ \beta_1 = \beta_2 \quad vs. \quad \mathrm{H}_1 :\ \beta_1 \neq \beta_2.$$

It can be transformed to

$$\mathrm{H}_0 :\ \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} = 0 \quad vs. \quad \mathrm{H}_1 :\ \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} \neq 0,$$

where $\mathbf{q} = (1, -1)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$.

Gauss-Markov Theorem tells us the Best Linear Unbiased Estimator (BLUE) of $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$ is $\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. Now we further discuss the hypothesis testing and confidence interval of $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$.

Note that

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

Same as before, we can build up test statistic

$$\frac{\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}}{s\sqrt{\mathbf{q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{q}}} \sim T_{n-p}.$$

To test $\mathrm{H}_0: \ \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} = 0$, we reject H_0 when $\left|\frac{\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}}{s\sqrt{\mathbf{q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{q}}}\right| > t_{n-p}^{-1}(1-\alpha/2)$.

And the $(1-\alpha)$ CI of $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$ is

$$\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p}^{-1}(1-\alpha/2)s\sqrt{\mathbf{q}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{q}}.$$

1.6 Prediction Intervals

Suppose now comes a new subject with covariates x_0^{T}. How can we predict the corresponding Y_0? How can we estimate $\mathtt{E}(Y_0)$.

First, we discuss how to estimate $\mathtt{E}(Y_0) = x_0^{\mathrm{T}}\boldsymbol{\beta}$.

Point Estimate: $\widehat{\mathtt{E}(Y_0)} = x_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. The variance of $\widehat{\mathtt{E}(Y_0)}$: $\mathtt{Var}(\widehat{\mathtt{E}(Y_0)}) = \sigma^2 x_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}x_0$.

Therefore, $(1-\alpha)$ confidence interval of $\mathtt{E}(Y_0)$ is

$$\mathbf{X}_0\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)s\sqrt{x_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}x_0}.$$

Are the point estimate and CI the same as those of Y_0's?

Answer: The point estimate is the same, but the CI is different. Why?

$$Y_0 = x_0^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_0.$$

The estimator $\hat{Y}_0 = x_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{\epsilon}_0$. Consider $\hat{\epsilon}_0 = 0$. Then

$$\begin{aligned}
\mathtt{Var}(\hat{Y}_0) &= \mathtt{E}\{(\hat{Y}_0 - Y_0)^2\} \\
&= \mathtt{E}\{(x_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} - x_0^{\mathrm{T}}\boldsymbol{\beta})^2\} + \mathtt{E}\{(\hat{\epsilon}_0 - \epsilon_0)^2\} \\
&= \mathtt{Var}(x_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \mathtt{Var}(\epsilon_0) \\
&= \sigma^2 x_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}x_0 + \sigma^2
\end{aligned}$$

Therefore, the $(1-\alpha)$ confidence interval of Y_0 is

$$x_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)s\sqrt{1 + x_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}x_0}.$$

1.7 Testing Multiple Contrasts

$$\mathrm{H}_0: \ \mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m} \quad \textit{vs.} \quad \mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} \neq \mathbf{m}.$$

Why are we interested in such hypothesis testings?

Example: Consider the clinical trial example of bmi. Suppose the researchers are not only interested in comparing Treatment A and Treatment B. In addition, they are interested to test whether the effect of Treatment A is equal to 1. How to do the test simultaneously?

Let

$$\mathbf{K} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Then the hypothesis $\mathrm{H}_0 : \begin{cases} \beta_1 = \beta_2 \\ \beta_1 = 1 \end{cases}$ can be written as $\mathrm{H}_0: \ \mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$.

Example 2: Car price example. The investigator is interested in the relationship between the car prise and a bunch of variables, including engine size, horsepower, RPM, passenger capacity, rear seat room and whether the car is imported.

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_3 + \beta_4\mathbf{X}_4 + \beta_5\mathbf{X}_5 + \beta_6\mathbf{X}_6 + \boldsymbol{\epsilon}$$

Here $\mathbf{Y}$ stands for car price; $\mathbf{X}_i$ stands for the covariates.

The researchers would like to know whether the passenger capacity and the rear seat room would affect the car price. The hypothesis would be $\mathrm{H}_0: \ \beta_4 = \beta_5 = 0$. It can be written as the form $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$. How?

Now suppose $\mathbf{K}$ is an $(p) \times s$ matrix and $\boldsymbol{\beta}$ is an $(p) \times 1$ vector. Also suppose $\texttt{Rank}(\mathbf{K}) = s$.

If we conduct GLR test, in the end, we will reject H_0 if

$$F = \frac{(n-p)(\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{m})^{\mathrm{T}}(\mathbf{K}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{m})}{s\hat{\sigma}^2}$$

is large.

Under H_0, F follows F distribution $F_{s,n-p}$. Why?

Then we need to reject H_0 if $F > F^{-1}_{s,n-p}(1-\alpha)$.

Now let's think about some special cases:

Case 1: $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$.

What does it mean? (Think about Example 1)

In this case, $\mathbf{K} = \mathbf{I}_{p\times p}$ and $\mathbf{m} = \mathbf{0}_{p\times 1}$.

Then

$$F = \frac{\hat{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}}}{(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}})} \cdot \frac{n-p}{p} = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y}}{\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}} \cdot \frac{n-p}{p}$$

Reject H_0 if $F > F^{-1}_{p,n-p}(1-\alpha)$.

The numerator and denominator of F have interpretations. We call them

$$\text{Sum Square Regression} = SSR = \mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y}$$
$$\text{Sum Square Error} = SSE = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$$

Then, $F = \frac{SSR/p}{SSE/(n-p)}$.

Further,

$$\text{Total Sum Square} = SST = SSR + SSE = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}.$$

We therefore decompose SST into two independent part: SSR and SSE. In order to better display the results, we set up an Analysis of Variance (ANOVA) table for the test.

Table 1: The ANOVA table for $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$.

Source	SS	d.f.	MS	F-Statistic
Regression	$\mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y}$	p	$(\mathbf{Y}^{\mathrm{T}}\mathbf{P}_X\mathbf{Y})/p$	$F = \frac{SSR/p}{SSE/(n-p)}$
Error	$\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$	$n-p$	$(\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y})/(n-p)$	$\sim F_{p,n-p}$ under H_0
Total	$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$	n		

Case 2: $\boldsymbol{\beta}_1 = 0$.

Now consider the linear model with the intercept:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix}\mathbf{1}_{n\times 1} & \mathbf{X}_{1,n\times p}\end{pmatrix}\begin{pmatrix}\beta_0 \\ \boldsymbol{\beta}_1\end{pmatrix} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}^2)$. Here $\boldsymbol{\beta}_1$ is a $p \times 1$ vector.

We would like to test $\mathrm{H}_0 : \boldsymbol{\beta}_1 = 0$. What does it mean? (Think about Example 2).

If we would like the hypothesis into the form $\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{m}$, what would $\mathbf{K}$ and $\mathbf{m}$ be?

In this case,

$$F = \frac{SSR_m/p}{SSE/(n-p-1)},$$

where

$$SSR_m = \mathbf{Y}^{\mathrm{T}}(\mathbf{P}_X - \mathbf{J}/n)\mathbf{Y}$$
$$SSE = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y},$$

where $\mathbf{J}$ is the matrix of all ones. SSR_m is called the Sum Square Regression corrected for Means.

Similarly,

$$SST_m = SSR_m + SSE = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{J}/n)\mathbf{Y},$$

where SST_m is called Total Sum Square of Errors corrected for Means.

Table 2 displays the ANOVA analysis.

Table 2: The ANOVA table for $\mathrm{H}_0 : \boldsymbol{\beta}_1 = \mathbf{0}$.

Source	SS	d.f.	MS	F-Statistic
SSR_m	$\mathbf{Y}^{\mathrm{T}}(\mathbf{P}_X - \mathbf{J}/n)\mathbf{Y}$	p	$\{\mathbf{Y}^{\mathrm{T}}(\mathbf{P}_X - \mathbf{J}/n)\mathbf{Y}\}/p$	$F = \frac{SSR_m/p}{SSE/(n-p-1)}$
SSE	$\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$	$n-p-1$	$(\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_X)\mathbf{Y})/(n-p-1)$	$\sim F_{p,n-p-1}$ under H_0
SST_m	$\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{J}/n)\mathbf{Y}$	$n-1$		

Also $SSRm = SSR_1 - SSR_2$, where SSR_1 is the SSR of the full model, and SSR_2 is the SSR of the model with only the intercept. SSR_m can be viewed as the difference of SSR between a larger model and a smaller model. In hypothesis testing, the smaller model is also called the null model or the nested model; the larger model is also called the expanded or the alternative model.