Spring 2014
Stat 8004: Statistical Methods II
Lecture 1

Jichun Xie

1 Review of Linear Models

1.1 Model and Notation

Linear Model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0} \text{ and } \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}. \tag{1}$$

Here suppose there are n samples and p covariates. Then $\mathbf{Y}$ is an $n \times 1$ vector and $\mathbf{X}$ is $n \times p$ matrix (the first column is the intercept).

Interpretation about $\boldsymbol{\beta}$: Fix other β_k's $(k \neq j)$, β_j is the increase of the expected value of Y caused by the unit increase of X_j.

1.2 Least Square Estimators

$$\begin{aligned} \text{Objective function:} \quad & \min_{\boldsymbol{\beta}\in\mathbb{R}^p} |\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|_2^2 \\ \Rightarrow \text{Normal equation:} \quad & \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \end{aligned}$$

When $\mathbf{X}$ is full rank, *i.e.* $\mathrm{Rank}(\mathbf{X}) = p$,

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

Geometric interpretation: A linear regression projects the outcome vector $\mathbf{Y}$ to the

linear space spanned by the columns of $\mathbf{X}$.

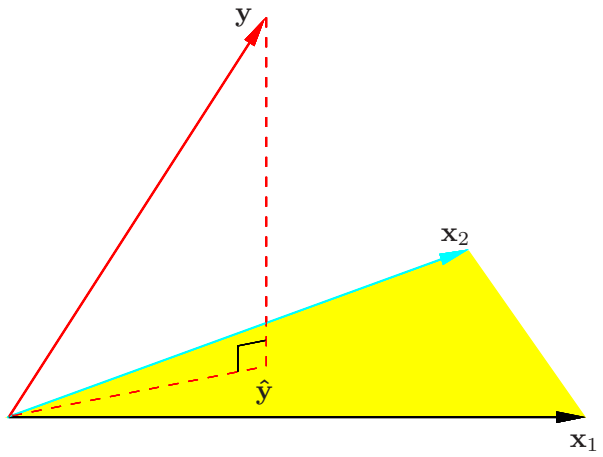

Figure 1: Linear regression and projection

Inference about $\hat{\boldsymbol{\beta}}$:

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

More inference results about $\hat{\boldsymbol{\beta}}$ (confidence interval, hypothesis testing) rely on the normal error assumption:

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

1.3 Hypothesis Testing and Confidence Interval of β_j

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

Here σ^2 is unknown, but we can estimate it with

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

Now the test statistic

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

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 $\boldsymbol{\beta}_j$:

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

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

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

Clinical examples: compare the effect of drug A and drug B on the disease, such as obesity and hypertension.

The Gauss-Markov Theorem: The best linear unbiased estimator (BLUE) of $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}$ is $\mathbf{q}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

What does the *best linear unbiased estimator* mean? How to prove the Gauss-Markov Theorem?

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

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

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

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

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

1.5 Prediction Intervals

Given covariate $\mathbf{X}_0$, how to estimate the confidence interval of $\mathbb{E}(Y_0)$ and Y_0?

First, we discuss how to estimate $\mathbb{E}(Y_0) = \mathbf{X}_0\boldsymbol{\beta}$.

Point Estimate: $\widehat{\mathbb{E}(Y_0)} = \mathbf{X}_0\hat{\boldsymbol{\beta}}$. The variance of $\widehat{\mathbb{E}(Y_0)}$: $\mathrm{Var}(\hat{Y}_0) = \sigma^2\mathbf{X}_0^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_0$.

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

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

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?

$$Y_0 = \mathbf{X}_0\boldsymbol{\beta} + \epsilon_0.$$

The estimator $\tilde{Y}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}} + \epsilon_0$. We don't know ϵ_0, but we know its distribution. We can estimate it with 0. Then the point estimator

$$\tilde{Y}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}}.$$

The variance is

$$\mathrm{Var}(\tilde{Y}_0) = \sigma^2\mathbf{X}_0(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}_0 + \sigma^2.$$

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

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

1.6 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 1: Consider the clinical trial example of bmi. The researchers collected the difference between the bmi before and after the patient take the treatment. Suppose they are not only interested in comparing Treatment A and Treatment B, but also testing whether the effect of Treatment A is equal to 0. How to do hypothesis testing?

Let

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

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 $\mathrm{Rank}(\mathbf{K}) = s$.

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

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

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:

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

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

$$F = \frac{SSR_m/(p-1)}{SSE/(n-p)} = \frac{SSR_m/(p-1)}{\hat{\sigma}^2},$$

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

Also $SSR_m = SSR_1 - SSR_2$, 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.