Unit 2: Linear models: Inference

Eugene Katsevich

September 15, 2021

We now understand the least squares estimator $\widehat{\boldsymbol{\beta}}$ from geometric and algebraic points of view. In Unit 2, we will switch to a probabilistic perspective to derive inferential statements for linear models, in the form of hypothesis tests and confidence intervals. In order to facilitate this, we will assume that the error terms are normally distributed:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n). \tag{1}$$

1 Building blocks for linear model inference

First we put in place some building blocks: The multivariate normal distribution (Section 1.1), the distributions of linear regression estimates and residuals (Section 1.2), and estimation of the noise variance σ^2 (Section 1.3).

1.1 The multivariate normal distribution

Recall that a random vector $\boldsymbol{w} \in \mathbb{R}^d$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariate matrix $\boldsymbol{\Sigma}$ if it has probability density

$$p(\boldsymbol{w}) = \frac{1}{\sqrt{(2\pi)^d \text{det}(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right).$$

These random vectors have lots of special properties, including:

- (Linear transformation) If $\boldsymbol{w} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{A}\boldsymbol{w} + \boldsymbol{b} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
- (Independence) If $\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \Sigma_{22} \end{pmatrix}\right)$, then $\boldsymbol{w}_1 \perp\!\!\!\perp \boldsymbol{w}_2$ if and only if $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

An important distribution related to the multivariate normal is the χ^2_d (chi-squared with d degrees of freedom) distribution, defined as

$$\chi^2_d \equiv \sum_{j=1}^d w_j^2 \quad \text{for} \quad w_1, \dots, w_d \overset{\text{i.i.d.}}{\sim} N(0,1).$$

1.2 The distributions of linear regression estimates and residuals

The most important distributional result in linear regression is that

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}). \tag{2}$$

Indeed, by the linear transformation property of the multivariate normal distribution,

$$\begin{aligned}\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \Longrightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} &\sim N((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\ &= N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).\end{aligned}$$

Next, let's consider the joint distribution of $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. We have

$$\begin{aligned}\begin{pmatrix}\widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{\epsilon}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{H}\boldsymbol{y} \\ (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{H} \\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\boldsymbol{y} &\sim N\left(\begin{pmatrix}\boldsymbol{H} \\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\boldsymbol{X}\boldsymbol{\beta}, \begin{pmatrix}\boldsymbol{H} \\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\cdot\sigma^2\boldsymbol{I}\begin{pmatrix}\boldsymbol{H} & \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\right) \\ &= N\left(\begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\sigma^2\boldsymbol{H} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2(\boldsymbol{I}-\boldsymbol{H})\end{pmatrix}\right).\end{aligned} \tag{3}$$

In other words,

$$\widehat{\boldsymbol{\mu}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}) \quad \text{and} \quad \widehat{\boldsymbol{\epsilon}} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{H})), \quad \text{with} \quad \widehat{\boldsymbol{\mu}} \perp\!\!\!\perp \widehat{\boldsymbol{\epsilon}}. \tag{4}$$

Since $\widehat{\boldsymbol{\beta}}$ is a deterministic function of $\widehat{\boldsymbol{\mu}}$ (in particular, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\mu}}$), it also follows that

$$\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\boldsymbol{\epsilon}}. \tag{5}$$

1.3 Estimation of the noise variance σ^2

We can't quite do inference for $\boldsymbol{\beta}$ based on the distributional result (2) because the noise variance σ^2 is unknown to us. Intuitively, since $\sigma^2 = \mathbb{E}[\epsilon_i^2]$, we can get an estimate of σ^2 by looking at the quantity $\|\widehat{\boldsymbol{\epsilon}}\|^2$. To get the distribution of this quantity, we need the following lemma:

Lemma 1.1. *Let $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{P})$ for some projection matrix $\boldsymbol{P}$. Then, $\|\boldsymbol{w}\|^2 \sim \chi^2_d$, where $d = \text{trace}(\boldsymbol{P})$ is the dimension of the subspace onto which $\boldsymbol{P}$ projects.*

Proof. Let $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ be an eigenvalue decomposition of $\boldsymbol{P}$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} \in \{0, 1\}$. We have $\boldsymbol{w} \overset{d}{=} \boldsymbol{U}\boldsymbol{D}\boldsymbol{z}$ for $\boldsymbol{z} \sim N(0, \boldsymbol{I}_n)$. Therefore,

$$\|\boldsymbol{w}\|^2 = \|\boldsymbol{D}\boldsymbol{z}\|^2 = \sum_{i:D_{ii}=1} z_i^2 \sim \chi^2_d, \quad \text{where } d = |\{i : D_{ii} = 1\}| = \text{trace}(D) = \text{trace}(\boldsymbol{P}).$$

□

Recall that $\boldsymbol{I} - \boldsymbol{H}$ is a projection onto the $(n-p)$-dimensional space $C(\boldsymbol{X})^\perp$, so by Lemma 1.1 and equation (4) we have

$$\|\widehat{\boldsymbol{\epsilon}}\|^2 \sim \sigma^2\chi^2_{n-p}. \tag{6}$$

From this result, it follows that $\mathbb{E}[\|\widehat{\boldsymbol{\epsilon}}\|^2] = n - p$, so

$$\widehat{\sigma}^2 \equiv \frac{1}{n-p}\|\widehat{\boldsymbol{\epsilon}}\|^2 \tag{7}$$

is an unbiased estimate for σ^2. Question to ponder: Why does the denominator need to be $n - p$ rather than n for the estimator above to be unbiased?

2 Hypothesis testing

Typically two types of null hypotheses are tested in a regression setting: Those involving one-dimensional parameters and those involving multi-dimensional parameters. For example, consider the null hypotheses $H_0 : \beta_j = 0$ and $H_0 : \beta_S = 0$ for $S \subseteq \{1, \dots, p\}$, respectively. We discuss tests of these two kinds of hypothesis in Sections 2.1 and 2.2, and then discuss the power of these tests in Section 2.3.

2.1 Testing a one-dimensional parameter

The most common question to ask in a linear regression context is: Is the jth predictor associated with the response, when controlling for the other predictors? In the language of hypothesis testing, this corresponds to the null hypothesis

$$H_0 : \beta_j = 0. \tag{8}$$

According to (2), we have $\widehat{\beta}_j \sim N(0, \sigma^2/s_j^2)$, where, as we learned in Unit 1,

$$s_j^2 \equiv [(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}]^{-1} = \|\boldsymbol{x}^{\perp}_{*j}\|^2. \tag{9}$$

Therefore,

$$z_j = \frac{\widehat{\beta}_j}{\sigma/s_j} \sim N(0,1), \tag{10}$$

and we could construct a level α two-sided test of the null hypothesis (8) via $\phi(\boldsymbol{X}, \boldsymbol{y}) = \mathbb{1}(|z_j| > z_{1-\alpha/2})$. Since we don't know σ^2, we can substitute in the unbiased estimate (7) derived in Section 1.3. This gives us the t-statistic

$$t_j \equiv \frac{\widehat{\beta}_j}{\widehat{\sigma}/s_j} = \frac{z_j}{\widehat{\sigma}/\sigma}. \tag{11}$$

2.2 Testing a multi-dimensional parameter

2.3 Power

3 Confidence and prediction intervals