

## Lecture 24 — Linear Regression, cont'd

Prof. Philippe Rigollet

Scribe: Anya Katsevich

Recall that we have observations  $(X_i, Y_i)$  such that for some true unknown  $\beta^*$ , we have

$$Y_i = X_i^T \beta^* + \epsilon_i, \quad i = 1, \dots, n,$$

where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ . These  $n$  equations can be written as a single matrix equation as follows:

$$\vec{Y} = \mathbb{X}\beta^* + \vec{\epsilon}, \quad (1)$$

where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \sim \mathcal{N}(0, I_n)$ . We found that

$$\hat{\beta}^{\text{MLE}} = \hat{\beta}^{\text{LS}} = \operatorname{argmin}_{\beta} \|Y - \mathbb{X}\beta\|^2 = \operatorname{argmin}_{\beta} \sum_{i=1}^n (Y_i - X_i^T \beta)^2,$$

and the explicit solution is

$$\hat{\beta}^{\text{LS}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \vec{Y}. \quad (2)$$

## 1 The distribution of $\hat{\beta}^{\text{LS}}$

In order to construct confidence intervals and test hypotheses about the ground truth coefficient vector  $\beta^*$ , we need to know the distribution of  $\hat{\beta}^{\text{LS}}$ , our estimator of  $\beta^*$ . To derive this distribution, we substitute the expression (1) for  $Y$  into the formula (2) for  $\hat{\beta}^{\text{LS}}$ , to get

$$\hat{\beta}^{\text{LS}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \vec{Y} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T (\mathbb{X}\beta^* + \vec{\epsilon}) = \beta^* + (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \vec{\epsilon}$$

Recall that  $\epsilon \sim \mathcal{N}(\sigma^2 I_n)$ , and using the formula for how the covariance transforms under matrix multiplication, we have  $A\epsilon \sim \mathcal{N}(0, A(\sigma^2 I_n)A^T)$ . Let's use this formula with  $A = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$  to find the distribution of  $(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \vec{\epsilon}$ . We have

$$A(\sigma^2 I_n)A^T = \sigma^2 AA^T = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} = \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}$$

Hence  $(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \vec{\epsilon} \sim \mathcal{N}_k(0, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1})$  and therefore

$$\hat{\beta}^{\text{LS}} = \beta^* + \mathcal{N}_k(0, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}) = \mathcal{N}_k(\beta^*, \sigma^2 (\mathbb{X}^T \mathbb{X})^{-1}). \quad (3)$$

In particular  $\hat{\beta}^{\text{LS}}$  is an *unbiased* estimator of  $\beta^*$ .

## 2 Confidence intervals and tests for the entries of $\beta^*$

**Example.** Suppose

$$Y = \text{blood pressure}, \quad X = \begin{pmatrix} \text{age} \\ \text{weight} \\ \text{shoe size} \end{pmatrix}.$$

Our model is then

$$Y = \underbrace{\beta_1^* \text{age} + \beta_2^* \text{weight} + \beta_3^* \text{shoe size}}_{X^\top \beta^*} + \epsilon.$$

The values of the individual coefficients tell us about the relationship between the associated feature and the response. For example,  $\beta_1^* > 0$  means  $Y$  is positively correlated with age. Or, if we find that  $\beta_3^* = 0$ , this means that shoe size is irrelevant to blood pressure, and we can throw shoe size out of our model.

*Testing whether or not  $\beta_j^*$  is zero is the most important kind of hypothesis test in linear regression.* Let's work out both the confidence interval for  $\beta_j^*$  and the hypothesis test for whether or not  $\beta_j^*$  is zero.

### 2.1 Confidence interval

We first use (3) to derive the distribution of  $\hat{\beta}_j^{\text{LS}}$ . We get

$$\hat{\beta}_j^{\text{LS}} \sim \mathcal{N}_1(\beta_j^*, \sigma^2 (\mathbb{X}^\top \mathbb{X})_{jj}^{-1}),$$

where the subscript 1 is a reminder that this a random variable in one dimension. Note that the variance of coordinate  $j$  depends on the full matrix  $\mathbb{X}^\top \mathbb{X}$ !

This gives rise to the following confidence interval for  $\beta_j^*$  with coverage  $1 - \alpha$ :

$$\beta_j^* \in \left[ \hat{\beta}_j \pm \sigma \sqrt{((\mathbb{X}^\top \mathbb{X})^{-1})_{jj} z_{\alpha/2}} \right]$$

This may seem different from the typical confidence interval we've constructed in the past, in that it has no  $1/\sqrt{n}$  term. However,  $\mathbb{X}^\top \mathbb{X}$  has size  $n$ , so the term  $((\mathbb{X}^\top \mathbb{X})^{-1})_{jj}$  plays the role of  $1/\sqrt{n}$ .

**Issue:** we typically don't know the true variance  $\sigma^2$ ! Remember that  $\sigma^2$  is the variance of each of the  $\epsilon_i$ 's. Of course, we don't know  $\epsilon_i$ , but we can use

$$\hat{\epsilon}_i = Y_i - X_i^\top \hat{\beta}^{\text{LS}}, \quad \text{"residual"}$$

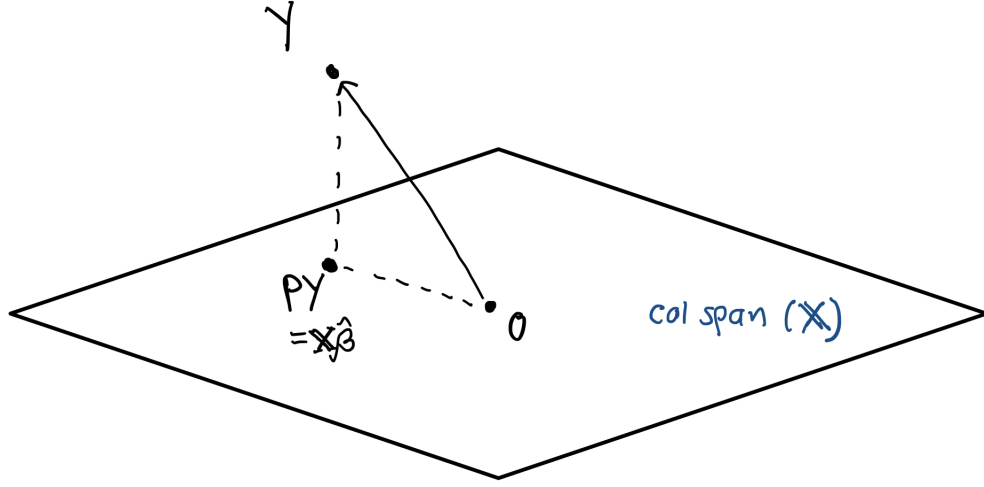


Figure 1: The vertical vector from  $PY = \mathbb{X}\hat{\beta}$  to  $Y$  is the *residual*.

as an approximation to  $\epsilon_i$ . Consider the vector

$$\vec{\hat{\epsilon}} = (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)^\top = \vec{Y} - \mathbb{X}\hat{\beta}.$$

This residual is the difference between  $\vec{Y}$  and its projection onto the column-span, i.e. it is the vertical component in Figure 1

Since  $\vec{Y}$  lives in dimension  $n$  and  $\mathbb{X}\hat{\beta}$  lives in dimension  $k$ , this means  $\hat{\epsilon} = \vec{Y} - \mathbb{X}\hat{\beta}$  lives in dimension  $n - k$ . In other words,  $\hat{\epsilon}$  has  $n - k$  degrees of freedom and hence

$$\|\hat{\epsilon}\|^2 \approx (n - k)\sigma^2.$$

We therefore use the following approximation to  $\sigma^2$ :

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|^2}{n - k} = \frac{1}{n - k} \sum_{i=1}^n \hat{\epsilon}_i^2.$$

## 2.2 Hypothesis test

Now, let's test

$$H_0 : \beta_j^* = 0 \quad \text{vs} \quad H_1 : \beta_j^* \neq 0.$$

at level  $\alpha$ . We reject the null if

$$|\hat{\beta}_j| \geq \hat{\sigma} \sqrt{((\mathbb{X}^\top \mathbb{X})^{-1})_{jj} z_{\alpha/2}},$$

which is equivalent to rejecting the null if the point 0 does not lie in the confidence interval  $\hat{\beta}_j \pm \sigma \sqrt{((\mathbb{X}^\top \mathbb{X})^{-1})_{jj}} z_{\alpha/2}$ . The p value is

$$\mathbb{P}(|Z| \geq |\hat{\beta}_j|/\hat{\sigma} \sqrt{((\mathbb{X}^\top \mathbb{X})^{-1})_{jj}})$$