STAT 151A: Lab 3

Billy Fang

15 September 2017

Feedback form is at the same place: https://goo.gl/forms/fKjLeKItix2Djg5l2. Please leave comments and suggestions for lab, office hours, etc.

1 Review/clarification of the characters in our story

1.1 Look Ma, no models!

All you have is your data:

$$(y_i, x_{i,1}, \dots, x_{i,p}), \qquad i = 1, \dots, n.$$

No assumption about randomness or data-generating process so far.

Goal: you want to find coefficients $\beta_0, \dots, \beta_p$ such that for each i, $\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$ is close to y_i.

In vector form, with

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

this is the same as **wanting $X\beta$ to be close to y element-wise.**

We call these differences $e_i := y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})$ the **residuals**; they measure the error of this particular choice of coefficients β in measuring the y_i. In vector form, $e = y - X\beta$.

The "best" β will be such that $X\beta$ is "closest" to y. How do we define closeness? Intuitively, we want all the residuals to be small, but how do we state this quantitatively? So far, the only notion we have studied is sum of squares of residuals. Many ways to write the same quantity:

$$S(\beta) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})]^2 = e^\top e = \|e\|^2 = \|X\beta - y\|^2.$$

So, let us define $\widehat{\beta}$ as the vector that make this quantity the smallest, and call it the **least squares coefficients**. The resulting **fitted values** for y are obtained by using these coefficients in the model.

$$\widehat{y}_i := \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_p x_{i,p},$$

or in vector form,

$$\widehat{y} := X\widehat{\beta}.$$

What do we know about $\widehat{\beta}$. By taking [partial] derivatives of $S(\beta)$, we see that $\widehat{\beta}$ must satisfy the normal equations

$$X^\top X \widehat{\beta} = X^\top y.$$

Does such a $\widehat{\beta}$ exist? When is it unique?

If $X^\top X$ is invertible, then we can write the unique solution as

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y.$$

We can also talk about our favorite sum of squares quantities. Below, $\overline{y} := \frac{1}{n}\sum_{i=1}^n y_i$.

$$\begin{aligned}
\text{RSS} &= S(\widehat{\beta}) = \|y - \widehat{y}\|^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2 && \text{this is the thing we were minimizing!} \\
\text{RegSS} &= \|\widehat{y} - \overline{y}\|^2 = \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2, \\
\text{TSS} &= \|y - \overline{y}\|^2.
\end{aligned}$$

By orthogonality / Pythagorean theorem, we have

$$\text{RSS} + \text{RegSS} = \text{TSS}.$$

Finally, the multiple correlation R^2 is defined as

$$R^2 = \frac{\text{RegSS}}{\text{TSS}}.$$

So far we do not have any notion of randomness, no statistical model. With just our data y and X, we can always perform least squares (regardless of whether it is a good idea or not), e.g. throwing the data into `lm()`.

1.2 Statistics's Next Top Model

Without further assumptions on y and X we cannot talk about the performance of the least squares estimator. For instance, if the relationship between the response variable and the explanatory variables are far from linear, then this estimator will not be good.

Thus, we often consider the following model.

$$y = X\beta^* + \epsilon,$$

where ϵ is a random vector representing the noise. Here β^* is some "true" coefficient vector. We typically think of X as fixed/known data (sometimes denoted as conditioning on X rather than fixing X), and view our observed response variable y as a noisy measurement of $X\beta^*$. The only source randomness in this model is in ϵ, so y, and any function of y (such as the least squares coefficient $\widehat{\beta}$) is random.

Thus this model imposes some underlying linear relationship between y and X, but it is masked by noise. Nonetheless, with additional assumptions about the noise ϵ, we can make concrete mathematical/probabilistic statements about the performance of estimator $\widehat{\beta}$ **under this model**, and do fun statistical things like tests.

Chapters 6 and 9 of Fox (as well as the recent lectures) assume

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

are i.i.d. (independent and identically distributed). This can be rewritten using the multivariate Gaussian distribution.

$$\epsilon \sim N_n(0, \sigma^2 I_n).$$

This very strong assumption is the "nicest" setting possible to study the least squares estimate. For instance, in this setting $\widehat{\beta}$ is the MLE of β^*. Moreover, the $\widehat{\beta}$ is itself Gaussian. In weaker settings, the Gauss-Markov theorem also gives a positive result about the performance of least squares: it is the Best Linear Unbiased Estimator (BLUE).

Whether or not such assumptions are reasonable is a deeper question, and as you drop assumptions, you are able to say less and less about See 6.1.1 in Fox for some discussion about common assumptions about the noise.

When studying properties of the least squares estimator, it is important to distinguish between intrinsic properties of the estimator that **do not depend on a model** (e.g., sum of residuals is zero, normal equations, Homework 2 Question 3, etc.), and properties of the estimator **under a specific model** (e.g., unbiasedness).

2 Exercises

Exercise 2.1. *Show the following.*

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$
$$\frac{\text{RegSS}}{\text{RSS}} = \frac{R^2}{1 - R^2}.$$

The second equality may be useful for your homework. ■

Exercise 2.2 (Fox Exercise 5.6)**.** *Why is it the case that the multiple-correlation coefficient R^2 can never get smaller when an explanatory variable is added to the regression equation?* ■

In the following two problems, we assume the model $y = X\beta^* + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I_n)$ and X is $n \times (p+1)$, and assume $X^\top X$ is invertible. Let $\widehat{\beta}$ be the least squares coefficient, and let $\widehat{y} = X\widehat{\beta}$ be the fitted values with residuals $e = y - \widehat{y}$. Let $H := X(X^\top X)^{-1}X^\top$.

Exercise 2.3.

(a) What is the distribution of y?

(b) What is the distribution of $\widehat{\beta}$?

Remember that a linear transformation of a Gaussian vector is also a Gaussian vector. ■

Exercise 2.4.

(a) Show $e = (I - H)y$.

(b) Show $(I - H)y = (I - H)\epsilon$.

(c) In your homework you prove directly that H is symmetric ($H^\top = H$) and idempotent ($H^2 = H$). Use this to prove that $I - H$ is also symmetric and idempotent.

(d) Show that consequently,
$$RSS = e^\top e = \epsilon^\top (I - H)\epsilon.$$

(e) Show that for any square matrix A,
$$\mathbb{E}[\epsilon^\top A\epsilon] = \sigma^2 \operatorname{tr}(A),$$
where $\operatorname{tr}(A) = A_{11} + A_{22} + \cdots + A_{nn}$ is the sum of the diagonal entries of A.

(f) Show that $\operatorname{tr}(H) = \operatorname{tr}(I_{p+1})$. (Hint: for any matrices A and B, we have $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, provided A and B can be multiplied together.)

(g) Use part (e) to show that
$$\mathbb{E}[\text{RSS}] = \sigma^2(n - (p + 1))$$

(h) What is therefore an unbiased estimator of σ^2?

In your lecture notes, it is also mentioned that RSS *is actually a χ^2 random variable with $n - p - 1$ degrees of freedom.* ■

3 Test time

Suggested reading: Sections 6.2.2 and 9.4.1-9.4.3 of Fox. The last section may or may not be helpful for your homework.

See `STAT151A_lab03_demos.html` on bCourses.