Multiple linear regression: Inference, Part III

Patrick Breheny

February 3

The Gauss-Markov Theorem

- Today we will start assuming distributional forms for the random errors, which in turn will allow us to develop confidence intervals and hypothesis tests
- Before we do that, we'll wrap up our least-squares results by proving one of the more famous theorems in statistics (which can be equivalently stated in one of two ways)
- **Gauss-Markov Theorem:** Suppose (1) holds and that we are interested in estimating $\boldsymbol{\lambda}^T\boldsymbol{\beta}$. Then $\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE).
- **Gauss-Markov Theorem:** Suppose (1) holds. Then $\hat{\boldsymbol{\beta}}$ is BLUE.

Comments

- Certainly, this is a very impressive result: regardless of the distribution of the outcome, the ordinary least squares estimate is the best linear unbiased estimator of any linear combination of the parameters $\{\beta_j\}$
- On the other hand, it is worth keeping in mind some caveats:
  - Once again, we are assuming that our model is correct; in particular, that there is not some additional variable out there which could help us explain y better
  - Why restrict ourselves to linear estimators?
  - Why restrict ourselves to unbiased estimators?

Linear combination of normals

- We will now state a number of facts about the normal distribution and related distributions, which we will use in the latter half of lecture when deriving distributional results for the least-squares estimators
- If X and Y are normally distributed, then X and Y are independent if and only if $\mathrm{Cov}(X, Y) = 0$
- Suppose the variables $\{Y_i\}$ are independent and normally distributed with means $\{\mu_i\}$ and variances $\{\sigma_i^2\}$; then

$$\sum_i a_i Y_i \sim \mathrm{N}\left(\sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2\right)$$

$$\mathbf{a}^T \mathbf{y} \sim \mathrm{N}\left(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}\right)$$

The multivariate normal distribution

- The normal distribution can be extended to describe vectors of random variables with what is called the *multivariate normal distribution*
- Suppose $\{Y_i\}$ are normally distributed random variables, with $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{y}) = \boldsymbol{\Sigma}$; then $\mathbf{y}$ is said to have a multivariate normal distribution, denoted

$$\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The χ^2 distribution

- Suppose $\{Z_i\}_{i=1}^n$ are independent random variables, each with a standard normal distribution; the sum of $\{Z_i^2\}$ is said to follow a χ^2 distribution with n degrees of freedom:

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

- From this definition, we can also see that if $\{X_i^2\}$ are independent random variables following χ^2 distributions with $\{n_i\}$ degrees of freedom, then $\sum_i X_i^2$ follows a χ^2 distributions with $\sum_i n_i$ degrees of freedom
- It is also straightforward to check that if X^2 follows a χ^2 distributions with n degrees of freedom, then $\mathrm{E}(X^2) = n$ and $\mathrm{Var}(X^2) = 2n$

The χ^2 distribution (cont'd)

- If the random variables are normal, but not necessarily independent or standard normal, they can still be combined to form a variable with a χ^2 distribution
- Suppose $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; then

$$(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2_n,$$

where n is the number of elements of $\mathbf{y}$

The χ^2 distribution (cont'd)

- Indeed, even if $\boldsymbol{\Sigma}$ is not invertible, a similar sort of result can be constructed
- Let $\boldsymbol{\Sigma}^-$ be a matrix that satisfies $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^-\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$ (such a matrix is called a *generalized inverse*)
- Suppose $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ not necessarily full-rank; then

$$(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^- (\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2_k,$$

where k is the rank of $\boldsymbol{\Sigma}$

The t distribution

Suppose that $Z \sim \mathrm{N}(0, 1)$, $X^2 \sim \chi^2_n$, and that Z and X^2 are independent; then

$$\frac{Z}{\sqrt{X^2/n}} \sim t_n,$$

the t-*distribution* with n degrees of freedom

The F distribution

- Suppose that $X_1^2 \sim \chi_n^2$, $X_2^2 \sim \chi_m^2$, and that X_1^2 and X_2^2 are independent; then

$$\frac{X_1^2/n}{X_2^2/m} \sim F_{n,m},$$

  the *F distribution* with n and m degrees of freedom
- Note that n and m are specifically ordered: $F_{n,m}$ is not the same thing as $F_{m,n}$

Relationship between the t and F distributions

- Finally, suppose that $T \sim t_n$; then note that

$$T^2 \sim \frac{Z^2}{X^2/n} \sim F_{1,n}$$

- In other words, the t distribution is a (transformed) special case of the F distribution

New assumptions

- The results we are about to derive will be based on the following set of assumptions: Suppose that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

  where $\mathbf{X}$ is a fixed $n \times p$ matrix of full column rank and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors $\{\epsilon_i\}$ which are independently distributed normal random variables with mean 0 and variance σ^2

- In other words,

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

- For the rest of this lecture, I will refer to the above set of assumptions by saying something along the lines of "Suppose that (2) holds"

The distribution of $\hat{\boldsymbol{\beta}}$

- Now, we are ready to derive the distribution of $\hat{\boldsymbol{\beta}}$
- **Theorem:** Suppose that (2) holds. Then

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

Two theorems about RSS

- Before we can prove the main distributional result which allows us to construct confidence intervals and carry out hypothesis tests, we need to prove two theorems about the residual sum of squares
- **Theorem:** Suppose that (2) holds. Then $\mathrm{RSS} \sim \sigma^2 \chi^2_{n-p}$.
- **Theorem:** Suppose that (2) holds. Then RSS and $\hat{\boldsymbol{\beta}}$ are independent.
- The second of these theorems relies on the fact that if $\mathbf{y}$ follows a normal distribution, then $\mathrm{Cov}(\mathbf{B}^T\mathbf{y}, \mathbf{y}^T\mathbf{A}\mathbf{y}) = 2\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}$

Distributional result

- We are now in a position to derive the following very important result:
- **Theorem:** Suppose that (2) holds. Then

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{\mathrm{SE}}} \sim t_{n-p},$$

where $\widehat{\mathrm{SE}}$ is the square root of $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$

- **Corollary:** Suppose that (2) holds. Then

$$\frac{\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} - \boldsymbol{\lambda}^T\boldsymbol{\beta}}{\widehat{\mathrm{SE}}} \sim t_{n-p},$$

where $\widehat{\mathrm{SE}}$ is the square root of $\hat{\sigma}^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\lambda}$

Hypothesis tests

- This result is very useful; for one thing, it allows us to carry out hypothesis tests
- Under $H_0 : \beta_j = 0$,

$$\frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}} \sim t_{n-p},$$

- Under $H_0 : \boldsymbol{\lambda}^T \boldsymbol{\beta} = 0$,

$$\frac{\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}}{\widehat{\mathrm{SE}}} \sim t_{n-p},$$

Confidence intervals

- It also allows us to construct confidence intervals
- Define $t_{\alpha,n}$ to be the upper point of a t distribution with n degrees of freedom; *i.e.*, suppose $T \sim t_n$; then $\Pr(T > t_{\alpha,n}) = \alpha$
- The following is a $(1-\alpha) \times 100\%$ confidence interval for β_j:

$$\left[\hat{\beta}_j - t_{\alpha/2,n-p}\widehat{\mathrm{SE}}, \hat{\beta}_j + t_{\alpha/2,n-p}\widehat{\mathrm{SE}}\right]$$

- The following is a $(1-\alpha) \times 100\%$ confidence interval for $\boldsymbol{\lambda}^T\boldsymbol{\beta}$:

$$\left[\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} - t_{\alpha/2,n-p}\widehat{\mathrm{SE}}, \boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}} + t_{\alpha/2,n-p}\widehat{\mathrm{SE}}\right]$$

Simultaneous hypothesis testing

- Finally, suppose we wanted to test several hypotheses at once (this is not merely of hypothetical interest; next time, we'll see some practical examples)
- For example, suppose we wanted to test whether $\boldsymbol{\lambda}_1^T\boldsymbol{\beta}, \boldsymbol{\lambda}_2^T\boldsymbol{\beta}, \ldots \boldsymbol{\lambda}_q^T\boldsymbol{\beta}$ were all equal to zero
- First, some notation: let's collect $\{\lambda_i\}$ into a $p \times q$ matrix $\boldsymbol{\Lambda}$, let $\hat{\boldsymbol{\tau}} = \boldsymbol{\Lambda}^T\hat{\boldsymbol{\beta}}$, and let $\mathbf{V} = \boldsymbol{\Lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\Lambda}$

The F-test

- **Theorem:** Suppose that (2) holds. Then

$$\frac{(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})^T \mathbf{V}^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})}{q\hat{\sigma}^2} \sim F_{q,n-p}$$

- We can then carry out a test of $H_0 : \boldsymbol{\Lambda}^T \boldsymbol{\beta} = \mathbf{0}$ based on

$$\frac{\hat{\boldsymbol{\tau}}^T \mathbf{V}^{-1} \hat{\boldsymbol{\tau}}}{q\hat{\sigma}^2} \sim F_{q,n-p}$$

The F-test (cont'd)

- In the special case where all the $\{\lambda_i\}$ test whether individual coefficients are equal to zero, the above can be equivalently stated in terms of the residual sums of squares coming from the "full" and "reduced" (leaving out those coefficients hypothesized to be zero) models
- **Corollary:** Suppose that that (2) holds and that we are testing whether some set of coefficients $\{\beta_j\}_{j=1}^{q}$ are all equal to zero. Let RSS_1 and RSS_0 denote the residual sums of squares for the full and reduced models, respectively. Then under $H_0 : \beta_j = 0$ for all j,

$$\frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/q}{\mathrm{RSS}_1/(n-p)} \sim F_{q,n-p},$$

where q is the difference in the number of parameters between the two models

Confidence intervals

- In theory, this F-distribution result can be used to construct confidence regions for $\boldsymbol{\tau}$ as well
- A $(1-\alpha)\times 100\%$ confidence set for $\boldsymbol{\tau}$ is given by the set of all $\boldsymbol{\tau}$ that satisfy

$$(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau})^T\mathbf{V}^{-1}(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}) \leq q\hat{\sigma}^2 F_{\alpha,q,n-p}$$

- Such confidence sets, however, are not common, as the resulting set is a q-dimensional ellipsoid, which is not easy to report and describe