Lecture 4 — 22/03/2022

The random linear model:

equivalent:

1) $y_i = z_i^T\beta + \varepsilon_i$ $\quad \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$, $i=1,\dots,n$, $\quad z_i \in \mathbb{R}^p$

2) $y = Z\beta + \varepsilon$ $\quad \varepsilon \sim N(0, \sigma^2 I_n)$ $\quad \beta \in \mathbb{R}^p, Z \in \mathbb{R}^{n\times p}$

3) $y \sim N(Z\beta, \sigma^2 I)$

Thm. (distributional properties of LS)

$$\begin{cases} \hat\beta \sim N(\beta, \sigma^2 (Z^T Z)^{-1}) \\ \hat y \sim N(Z\beta, \sigma^2 H) \\ \hat\varepsilon \sim N(0, \sigma^2 (I - H)) \end{cases}$$

Furthermore, $\hat\varepsilon$ ind. of $\hat\beta$ & $\hat y$

Thm.

$$\frac{1}{\sigma^2}\sum_{i=1}^n \hat\varepsilon^2 = \frac{\|\hat\varepsilon\|^2}{\sigma^2} \sim \chi^2_{n-p}$$

Prf. We need to show that $\frac{\|\hat\varepsilon\|^2}{\sigma^2}$ can be written as the sum of squares of $n-p$ ind, standard normal RVs.

We have: $\quad y = \hat y + \hat\varepsilon \quad$ and $\quad \hat y = Hy$

Therefore:

$$\underline{(I-H)y} = y - \hat y = \hat\varepsilon$$

$$\underline{(I-H)y} = (I-H)(Z\beta + \varepsilon) = (I-H)Z\beta + (I-H)\varepsilon = (I-H)\varepsilon$$

bc. $(I-H)Z\beta = Z\beta - \underbrace{Z(Z^TZ)^{-1}Z^T}_{H} Z\beta = 0$

Consequently

$$\|\hat{\varepsilon}\|^2 = \hat{\varepsilon}^T\varepsilon = \left((I-H)y\right)^T(I-H)y$$

$$= \left((I-H)\varepsilon\right)^T(I-H)\varepsilon = \varepsilon^T(I-H)^T(I-H)\varepsilon$$

$$= \varepsilon^T(I-H)^2\varepsilon = \varepsilon^T(I-H)\varepsilon$$

write: $\underline{I-H = P^T\Lambda P}$ $\quad \boxed{P^TP=I}$ $\qquad \Lambda = diag(\lambda_1,\dots,\lambda_n)$ $\quad \lambda_i\in\{0,1\}$

$\underline{\sum_{i=1}^n \lambda_i = n-p}$ bc. $I-H$ is PPM of rank $n-p$

$$\eta = P\varepsilon \sim N(0,\sigma^2 I)$$

$$Var(P\varepsilon) = P^T \overbrace{Var(\varepsilon)}^{\sigma^2 I} P = \sigma^2 P^T I P = \sigma^2 I$$

$$\eta_1,\dots,\eta_n \overset{iid}{\sim} N(0,\sigma^2)$$

$$\|\hat{\varepsilon}\|^2 = \varepsilon^T(I-H)\varepsilon = \varepsilon^T P^T\Lambda P\varepsilon = \eta^T\Lambda\eta$$

$$= \sum_{i=1}^n \eta_i^2\lambda_i = \underline{\sum_{i:\lambda_i=1}\eta_i^2} \sim \underline{\chi^2_{n-p}\cdot\sigma^2}$$

$$\Rightarrow \frac{\|\hat{\varepsilon}\|^2}{\sigma^2} \sim \chi^2_{n-p}$$

x is eigenvector of $I-H$ with eigenvalue λ iff

$$(I-H)x = \lambda x$$

$$x - Hx = \lambda x$$

$$Hx = x - \lambda x$$

$$(1-\lambda)x$$

iff x is an eigenvector of H with eigenvalue $1-\lambda$

$\square$

Application: t-test

From: $\hat{\beta}-\beta \sim N(0, \sigma^2(Z^TZ)^{-1})$

$$\frac{\hat{\beta}-\beta}{\sigma} \sim N(0, (Z^TZ)^{-1})$$

$$\frac{c^T(\hat{\beta}-\beta)}{\sigma} \sim N(0, c^T(Z^TZ)^{-1}c) \quad c\in\mathbb{R}$$

$\in \mathbb{R}^1$

$$v = \frac{c^T(\hat{\beta}-\beta)}{\sigma\sqrt{c^T(Z^TZ)^{-1}c}} \sim \underline{N(0,1)}$$

Define:

$$S^2 := \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p}$$

we have $\|\hat{\varepsilon}\|^2/\sigma^2 \sim \chi^2_{n-p}$

hence

$$\frac{S^2}{\sigma^2}(n-p) \sim \chi^2_{n-p}$$

because $\hat{\varepsilon}$ and $\beta-\hat{\beta}$ are ind. therefore

$$t := \frac{c^T(\hat{\beta}-\beta)}{S\sqrt{c^T(Z^TZ)^{-1}c}} = \frac{v}{\sqrt{S^2/\sigma^2}} = \frac{v}{\sqrt{\|\hat{\varepsilon}\|^2/((n-p)\sigma^2)}} \sim t_{n-p}$$

Suppose that

$$c = [0,\dots,0,\underset{\uparrow}{1},0\dots]^T \in \mathbb{R}^p$$

1 in the j-th entry

If we hypothesize that $\beta_j = 0$, we would have

$$t = \frac{\hat{\beta}_j - 0}{S\sqrt{c^T(Z^TZ)^{-1}c}} \sim t_{n-p}$$

very large or small values of t are evidence against our hypothesis

Application: F-test for extra sum-of-squares

- Suppose a full model $y = Z\beta + \varepsilon \quad \beta \in \mathbb{R}^p$
  and a small model $y = \tilde{Z}\gamma + \varepsilon \quad \gamma \in \mathbb{R}^q$

  $\tilde{Z}$ is obtained by removing columns from Z
  or equivalently, set $\beta_j = 0$ for some j-s

- We want to test wheather the small model "is a valid representation of the data"

- We fit $\hat{\beta}$ and $\hat{\gamma}$, and write:

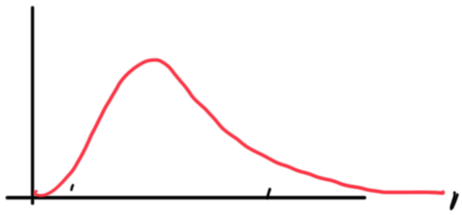
$$SS_{Full} = \sum_{i=1}^{n} (y_i - Z_i^T \hat{\beta})^2$$

$$SS_{sub} = \sum_{i=1}^{n} (y_i - \tilde{Z}_i^T \hat{\gamma})^2$$

- We know that $SS_{Full} < SS_{sub}$

- We can use:

$$F = \frac{\frac{1}{p-q}(SS_{sub} - SS_{full})}{\frac{1}{n-p} SS_{full}} \sim F_{p-q,\, n-p}$$



Gauss-Markov Theorem

Let $Y = Z\beta + \varepsilon$ where Z is a non-random $n \times p$ matrix, β is an unknown point in $\mathbb{R}^p$, and ε is a random vector with mean 0 and variance $\sigma^2 I$.

Let $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$ and fix $c \in \mathbb{R}^p$. If $\lambda \in \mathbb{R}^n$ satisfies

$\mathbb{E}[\lambda^T Y] = c^T \beta$, then

$$\mathrm{Var}(\lambda^T Y) \geq \mathrm{Var}(c^T \hat{\beta})$$

Conclusions:

- The theorem states that the least squares estimate $\hat{\beta} = (Z^T Z)^{-1} Z y$ (which is linear in y) has minimal variance over all linear, unbiased estimators of β
- The theorem does not require normality
- Takeaway: to beat LS, you need bias or non-normality

Introduction to Statistical Inference

Mean & Variances

- Suppose that we have no x's and $Z = [1, \ldots, 1]^T$ so that

$$y_i = \mu + \varepsilon_i \qquad (\mu = \beta_0)$$

Example: Say that we obtained data on the ages of 17 of our users

$$y = (15, 7, 9, 10, 7, 8, 20, 16, 9, 19, 14, 10, 11, 10, 10$$

- The average of the sample is

$$\hat{\mu} = \frac{1}{n} \sum^{17} y_i \approx 11.38$$

$i=1$

is this a good estimate of the "true" μ?

- If (Y_i) is iid and has variance σ^2,

then $Var(\bar{Y}) = Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2} Var(\sum Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \overbrace{Var(Y_i)}^{\sigma^2}$

$= \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$

- How to get σ^2:

one option is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

However, $\hat{\sigma}^2$ is biased downwards since

$$\hat{\sigma}^2 \le \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

($\hat{\sigma}^2$ minimizes sum of squares by design)

we typically use

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

(Indeed $E[s^2] = \sigma^2$ while $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$)

In our example, $s^2 \doteq 16.38$, so $\widehat{var\,\bar{Y}} = \frac{s^2}{n} \approx 0.964$

$$\bar{y} \pm 2\sqrt{\widehat{var\,\bar{Y}}} = 11.4 \pm 2\sqrt{0.964} \approx [9.5, 13.5]$$

- The logic: if $Z \sim N(\mu, \sigma^2)$ then

$$Pr(|Z - \mu| \le 2\sigma) = Pr(Z \in [\mu - 2\sigma, \mu + 2\sigma]) \ge 0.9$$

If $\bar{Y} \sim N(\mu, s^2/n)$, then $Pr(\bar{Y} \in [9.5, 13.5]) \ge 0.9$

- but the quality of our variance estimate depends on $\text{var}(\widehat{\text{var}(\bar{Y})})$

We have

$$\text{Var}[\widehat{\text{var}(\bar{Y})}] = \text{Var}[s^2] = \sigma^4\left(\frac{2}{n-1} + \frac{\kappa}{n}\right)$$

Where κ is the kurtosis.

- We don't know κ, so we can plug-in its estimate and obtain $\widehat{\text{Var}[\widehat{\text{var}(\bar{Y})}]}$

- In general, we estimate $\text{Var}(\widehat{\text{var}}^{(k)}(\bar{Y}))$ using $\widehat{\text{var}}^{(k+1)}(\bar{Y})$

- This is what Tukey called "the staircase of inference". It tells you that we cannot eliminate all doubt in any of our findings
- Most people stop at the mean and var

Testing

Suppose we want to know whether the average age of our users is less than 10

- Set $\mu = E(Y_i)$ and $\mu_0 = 10$

$$H_0: \mu = \mu_0$$

Our alternative hypothesis:

$$H_1: \mu \neq \mu_0$$

(other options are: $H_1: \mu < \mu_0$)

(or $H_1 : \mu > \mu_0$)

we reject if observed data is unlikely under H_0. If not, we fail reject.

One sample t-test

- Assume $Y_1, \dots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 are unknown
- We test $H_0 : \mu = \mu_0$ using

$$t = \frac{\bar{Y} - \mu_0}{s} \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- If $\mu = \mu_0$, then $t \sim t_{n-1}$
- If H_0 is true, our t-statistic t is a sample from a common part of t_{n-1}
- If we get an extreme value of t, it is unlikely that $\mu = \mu_0$, in which case we reject H_0.
- If $H_1 : \mu \neq \mu_0$, reject if $\quad t_{obs} = \dfrac{\bar{y} - \mu_0}{s}$

$$p = Pr(|T_{n-1}| \geq t_{obs}) \quad T \sim t_{n-1}$$

is small

- If H_1 $\mu > \mu_0$, reject if

$$p_> = Pr(T_{n-1} \geq t_{obs}) \quad T \sim t_{n-1}$$

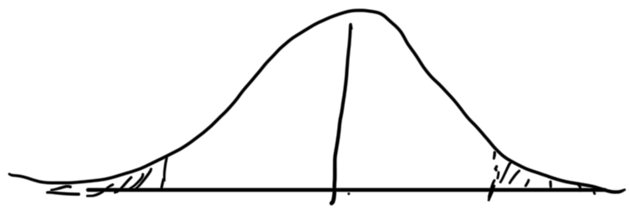
is small

- The probabilities p and $p_>$ are called <u>p-values</u>

- In words: "a p-value is the probability of observing what we got or a more extreme value under the null H_0"

- If the p-value is small, either H_0 is false or a very rare event occured

- in our example, $p = Pr(|T_{16}| \geq 1.438) = 2 \cdot 0.085 = .1$



-1.438 $\qquad$ 1.438

we cannot reject at level $\alpha = 0.05$
(or $\alpha = 0.01$, or $\alpha = 0.001$)

- One tailed test warning:

$$p_{\geq} = Pr(T_{n-1} \geq t_{obs}) = \frac{1}{2} Pr(|T_{n-1}| \geq t_{obs})$$

should rarely be used. $\qquad t_{obs} = \frac{\bar{y} - \mu_0}{s}$

- The strength of evidence against H_0 depend on the effect size (e.g. $|\mu - \mu_0|$) and the sample size n. For small sample sizes, it may simply be impossible to obtain small enough p-value that to convince us to reject H_0.
"p measures the sample size" (R. Olshen)