Lecture 3. Properties of Least Squares Estimates; Tests and Confidence Regions. 01 Mar 2018

R code: Recall the advertising example. From fitting (lm()) we get the estimated coeffs ($\hat{\beta}$) so ex.

$$\widehat{sales} = 2.93 + 0.046 \cdot TV + 0.179 \cdot Radio + (-0.001) \cdot Newspaper$$

Recall
- residuals := $e_i = y_i - \hat{y}_i$; len$(y_i, \hat{y}_i) = n$.
- $\hat{\beta} = (X^TX)^{-1}X^TY$ with $\underset{(p\times p)}{Var(\hat{\beta})} = \sigma^2 (X^TX)^{-1}$

(need that variables are uncorrelated and have equal variance)

s.t. $Var(\hat{\beta}_j) = [\sigma^2(X^TX)^{-1}]_{jj}$ ← ie jth element in the diagonal of $Var(\hat{\beta})$

$$Var(\hat{\beta}) = \begin{bmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots \\ & Var(\hat{\beta}_2) & \\ & & \ddots \\ & & Var(\hat{\beta}_p) \end{bmatrix}$$

- $\sigma^2 = \frac{RSS}{n-p} = \frac{\sum_{i=1}^n e_i^2}{n-p}$, degrees of freedom := $(n-p)$ makes σ^2 unbiased.
  (p := rank of X)

* SEE Script 1.4 !!

- Residual standard error (RSS) $\equiv \hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-p}}$ helps measures the goodness of fit.

- Issue: residuals may not be ~~uniformly~~ normally (randomly) distributed ↔ they don't fluctuate around zero (randomly), ie skewed.

RSS: residuals wrt regression line ($\hat{y}_i$) $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

TSS: residuals wrt mean level/value ($\bar{y}_i$), ie $TSS = \sum_{i=1}^n (y_i - \bar{y}_i)^2$

Note $TSS \geq RSS$.

- Sample variance of y-values := $\frac{1}{n-1} \cdot TSS$
- Remaining variance after fitting the model := $\frac{1}{n-1} RSS$

- Proportion of variance that remains after fitting the model $\in [0,1]$:= $\frac{RSS}{TSS} = \frac{RSS/(n-1)}{TSS/(n-1)}$ (smaller value better)

- $R^2 = 1 - \frac{RSS}{TSS}$ =: proportion of variance that is explained by the model

$R^2 \in [0,1]$ → · larger values are better
· Not affected by scale of y
· R^2 is the same as $(\text{sample correlation})^2$, but R^2 is more general. (in simple linear regression)

- Adding more variables to the model can only increase R^2 (as RSS can only decrease)
→ means we can't fairly compare models of which one is a submodel/subset of the other.
(Account for this with R_{adj}!)

Computational Statistics — 07 Mar 2018

Lecture 3: Least Squares Estimates; Tests and Confidence Regions, cont'd

- **R^2 (cont'd)** To account for comparing subsets of models, we adjust R^2 according to the model's degrees of freedom $(n-p)$:

$$\text{i.e. } R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} \Rightarrow$$

adding unnecessary variables to the model will decrease R^2_{adj} (as $RSS/(n-p)$ can only increase)

$$\text{ie } R^2_{adj} < R^2$$

- **Standard error of $\hat{\beta}_j$.** We generally can't calculate $Var(\hat{\beta}_j) = [\sigma^2 (X^T X)^{-1}]_{jj}$ (σ is unknown) but we can calculate $\hat{Var}(\hat{\beta}_j) = [\hat{\sigma}^2 (X^T X)^{-1}]_{jj}$

$$\rightarrow \text{Standard of error of } \hat{\beta}_j = \sqrt{\hat{Var}(\hat{\beta}_j)} = \sqrt{[\hat{\sigma}^2 (X^T X)^{-1}]_{jj}}$$

- **Distribution of least squares estimates assuming Gaussian errors.**

Since we assume a linear model and require $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim N(0, \sigma^2)$, we can show that

(i)
$$\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1}) \equiv N_p(\beta, Var(\hat{\beta}))$$

$$\hat{\beta}_j \sim N(\beta_j, [\sigma^2 (X^T X)^{-1}]_{jj}) \equiv N(\beta_j, Var(\hat{\beta}_j))$$

- **p-Values**: prob (observing a test statistic that is at least as extreme than the one we saw when the null hypothesis is true)

ie for $Y = X\beta + \varepsilon$, with $\begin{cases} \beta_j = 0 \text{ if } H_0 \text{ (null hypothesis)} \\ \beta_j \neq 0 \text{ if } H_a \text{ (alternative hypothesis)} \end{cases}$

- From the normal distribution of $\hat{\beta}_j$ we derive

Under the null-hypothesis, $\hat{\beta}_j \sim N(0, [\sigma^2 (X^T X)^{-1}]_{jj} \equiv Var(\hat{\beta}_j))$

$\rightarrow$ we can get $\dfrac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta})}} \sim N(0,1)$ (standard normal)

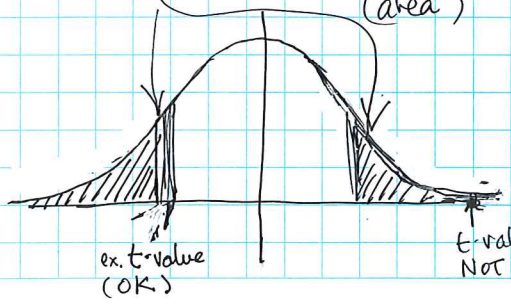
Substituting the unknown quantity σ^2 with the estimate $\hat{\sigma}^2$ we get the t-test statistic

$$\text{ie } T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}} \sim t_{n-p} \text{ under } H_0$$

(slightly different distr from $N(0,1)$; thicker tails)
As $n-p \to \infty$, $t_{n-p} \approx N(0,1)$

the p-value in the t-test refers to / is defined as:

(and or H_0) $p = Prob(|T\text{-statistic}| > |t\text{-value}| \mid H_0 \text{ is true})$ (area)



"An individual t-test for H_0: should be interpreted as quantifying the effect of the j^{th} predictor, after having subtracted the linear effects of all other predictor variables on Y."

- p values can vary substantially depending on which/how many variables are included in the model ⇒ need to consider which ones to include.

Lecture 3: Least Squares Estimates (Properties); tests, confidence regions, cont'd 07 Mar 2018

p Values, cont'd

ex. If we regress sales ~ TV + Radio + Newspaper, then the p-value for Newspaper: H_0: sales ~ TV + Radio
H_a: sales ~ TV + Radio + Newspaper

- An insignificant p-value says that there is no / little evidence in the data that shows that H_a is better than H_0. ie. so if you're already using radio + TV, it doesn't help to add Newspaper

- ex. Sales ~ Newspaper : H_0 : sales = $\beta_1 + \varepsilon$

F-test. Question: For a given variable after adding all the other variables, does it help to add this one?

- Partial F-test (anova) can compare subset and the whole model.
  ex. Model 1: sales ~ TV + radio
  Model 2: sales ~ TV + radio + newspaper
  - If we want to look at only newspaper, then
    Model 1: sales ~ 1 , Model 2: sales ~ newspaper

F-statistic example: checking full vs. empty model (football data)
↳ It is important for checking / looking at multiple testing problems.

- If F-stat is bad, it means that the model is not good (start over). If good / significant, continue analysis, looking at specific variables.

- Internally creates dummy variables.
- Ex. $y_i^{\text{balance}} = \beta_1 + \beta_2 \cdot \text{income} + \beta_3 x_i + \varepsilon_i$, $x_i = \begin{cases} 1 & \text{if } i^{\text{person}} \text{ is student} \\ 0 & \text{otherwise} \end{cases}$.

$$= \begin{cases} \beta_1 + \beta_2 \cdot \text{income} + \beta_3 + \varepsilon_i & \text{if student} \\ \beta_1 + \beta_2 \cdot \text{income} + \varepsilon_i & \text{otherwise} \end{cases}$$

↳ slope is the same (β_2)
yint differs: student: $\beta_1 + \beta_3$
other: β_1