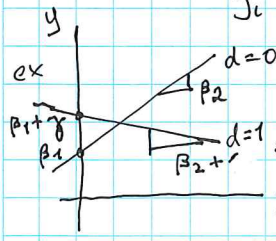
Computational Statistics Lecture 5: Confidence Intervals

08 March 2018

- Recall: Interaction dynamics (ex. btwn x_i, d_i), $x_i \in \mathbb{R}$, $d_i \in \{0, 1\}$.

$$y_i = \beta_1 + \beta_2 x_i + \gamma d_i + \delta d_i x_i + \epsilon_i$$

β_2: main effect of x_i; γ: main effect of d_i; δ: interaction btwn x_i and d_i



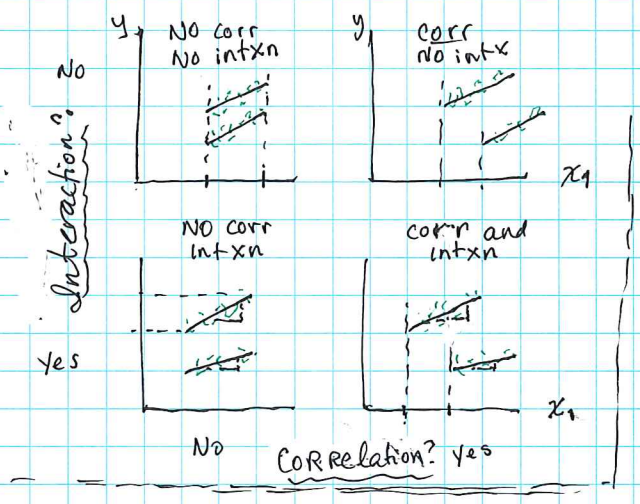
$$= (\beta_1 + \gamma d_i) + (\beta_2 + \delta d_i) x_i + \epsilon_i$$

y-int; slope; err

"effect" is wrt $\mathbb{E}$.

ex. Correlation vs. interaction —

- Interaction → change slope
- Correlation → x-shift



- Recall regression model:

$$Y_{n\times 1} = X_{n\times p}\,\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \text{with } \epsilon \sim N(0, \sigma^2 I) \text{ (err norm distr)}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{and}$$

- $\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ ($\equiv$ Cov of $\hat{\beta}$)
- From this, $\hat{\text{Var}}(\hat{\beta}_j) = \hat{\sigma}_2^2 [(X^T X)^{-1}]_{jj}$,

where R_j^2 is the R^2 from regressing x_j on the other variables.

- We can rewrite this as

$$\hat{\text{Var}}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{(n-1)\hat{\text{Var}}(x_j)} \cdot \frac{1}{1-R_j^2}$$

Behavior: If ... $n \uparrow$, then $\hat{\text{Var}}(\hat{\beta}_j) \downarrow$; $\sigma^2 \downarrow$, then $\hat{\text{Var}}(\hat{\beta}_j) \downarrow$; $\text{Var}(x_j) \uparrow$, " " $\downarrow$; $R_j^2 \uparrow$, " " $\uparrow$

take home message: it helps to have a good spread in X variables

- So $\hat{se}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(X^T X)^{-1}]_{jj}}$ (see R code output)

- Then, we have $\dfrac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)} \sim t_{n-p}$ $\Rightarrow$ $P\left(-t_{1-\frac{\alpha}{2},n-p} < \dfrac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)} < t_{1-\frac{\alpha}{2},n-p}\right) = 1-\alpha$

from which we derive the confidence interval which has prob $1-\alpha$.

Rewrite this as...

$$P\left(-\hat{se}\, t_{1-\frac{\alpha}{2},n-p} < \hat{\beta}_j - \beta_j < \hat{se}\, t_{1-\frac{\alpha}{2},n-p}\right) = 1-\alpha$$

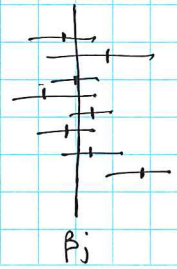
$$P\left(\hat{\beta}_j - \hat{se}\, t_{1-\frac{\alpha}{2},n-p} < \beta_j < \hat{\beta}_j + \hat{se}\, t_{1-\frac{\alpha}{2},n-p}\right) = 1-\alpha$$

R: center data

Thus, CI for β_j: $(1-\alpha)\cdot 100\%$ CI for β_j: $\hat{\beta}_j \pm \hat{se}(\hat{\beta}_j)\, t_{1-\frac{\alpha}{2},n-p}$ is a two-sided confidence interval which covers the true β_j with prob $1-\alpha$; here $t_{1-\frac{\alpha}{2},n-p}$ denotes the $1-\frac{\alpha}{2}$ quantile of a t_{n-p} distribution.

($\hat{\beta}_j$: point estimate; $\hat{se}$: estimated std dev.; $t_{1-\frac{\alpha}{2},n-p}$: quantile)

ex.



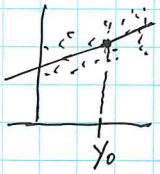
R: intervals vary in length ($\hat{se}$) and midpoint ($\hat{\beta}_j$)

in α% of cases, the midpoint $\hat{\beta}_j$ does not cover β_j.

Comp Stats Lecture 5: Confidence Intervals 08 Mar 2018

- Getting the CI for Y_0. Now consider a new point $X_0^T = (X_{01}, \ldots, X_{0p})$. We are interested in $E(Y_0)$ or Y_0. What is the difference?

→ $E[Y_0]$ will be the value on the regression line. Y_0 is a point and has more variability (since we have ϵ_0); won't be necessarily on regression line.



So,

$$Y_0 = X_0^T \beta + \epsilon \quad \text{and} \quad E(Y_0) = X_0^T \beta$$
$$\hat{Y}_0 = x_0^T \hat{\beta} + \epsilon \qquad = E(\hat{Y}_0)$$
$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$
$$\text{lm}(X_j \sim X_1 + \ldots + X_{j-1} + X_{j+1} + \ldots + X_p) \quad \text{(R code)}.$$

R_j^2 represents the proportion of variance in X_j that can be explained by a linear regression on the other predictors.

- Note that both quantities have expectation $E(Y_0) = X_0^T \beta$, so as a point estimate / best guess, we use $X_0^T \hat{\beta}$.
- Since $E(X_0^T \hat{\beta}) = X_0^T \beta$, this is an unbiased estimate.
- Note: $\text{Var}(X_0^T \hat{\beta}) = X_0^T \text{Var}(\hat{\beta}) X_0 = \sigma^2 X_0^T (X^T X)^{-1} X_0$.
- Hence, $$\frac{X_0^T \hat{\beta} - E(Y_0)}{\sigma \sqrt{X_0^T (X^T X)^{-1} X_0}} \sim t_{n-p}$$

→ So the $(1-\alpha)\cdot 100\%$ CI for $E(Y_0)$ is:

$$X_0^T \hat{\beta} \pm \sqrt{\hat{\sigma}^2 X_0^T (X^T X)^{-1} X_0}\; t_{1-\frac{\alpha}{2}; n-p}$$

- divide by * ⇒ Var to normalize: $$\frac{X_0^T \hat{\beta} - Y_0}{\text{Var}(X_0^T \hat{\beta} - Y_0)} \sim t_{n-p}$$

- Now for Y_0: $\text{Var}(X_0^T \hat{\beta} - Y_0) = \text{Var}(X_0^T \hat{\beta} - (X_0^T \beta + \epsilon)) = \text{Var}(X_0^T \hat{\beta} - \epsilon) =$

$$= \text{Var}(X_0^T \hat{\beta}) + \text{Var}(\epsilon) - 2\,\text{Cov}(X_0^T \hat{\beta}, \epsilon)$$
$$= [\sigma^2 X_0^T (X^T X)^{-1} X_0] + [\sigma^2] \qquad 0$$
$$= \sigma^2 (1 + X_0^T (X^T X)^{-1} X_0) = \text{Var}(X_0^T \hat{\beta} - Y_0).$$

accounts for widening of CI as X data spreads further from center

So, ① $(1-\alpha)$ CI for Y_0 is (aka "prediction" interval) $$X_0^T \hat{\beta} \pm \sigma \sqrt{1 + X_0^T (X^T X)^{-1} X_0}\; t_{1-\frac{\alpha}{2}; n-p}$$

CI of Y_0 is strictly larger than that of $E(Y_0)$

② $$\frac{X_0^T \hat{\beta} - Y_0}{\sigma \sqrt{1 + X_0^T (X^T X)^{-1} X_0}}$$ converges to zero as sample size increases

* - Bias-Variance Tradeoff (ISLR 2.2.1-2)

- Let $Y_i = f(x_i) + \epsilon_i$. Consider an estimator $\hat{f}(x_i)$. Then the mean squared error (MSE) is: $$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$ (More precisely, the training MSE as the points x_i are the data used to train the estimator.

How does $\hat{f}$ perform on new data? (ex. tomorrow's stock market values).

- (Now) From training data $(x_1, y_1), \ldots, (x_n, y_n)$ used to obtain estimator $\hat{f}$, we would like for $y_i \approx \hat{f}(x_i)$, i.e. small training MSE. Now, is $\hat{f}(x_0) \approx y_0$ for a new pair (x_0, y_0)? Test MSE at $x_0 = (y_0 - \hat{f}(x_0))^2$.

(From scratch, $MSE(x) = E[(\hat{f}(x) - f(x))^2] = \left(E(\hat{f}(x)) - f(x)\right)^2 + \text{Var}(\hat{f}(x))$.