Bootstrap for regression

$$y = X\beta + \varepsilon$$

Bootstrapping pairs

Given data $z_1 = (x_1, y_1), \ldots, (x_n, y_n)$,

1. Get $z^{*1}, \ldots, z^{*n}$ by sampling w/ replacement from $\{z_1, \ldots, z_n\}$
2. Compute OLS $\hat{\beta}$ from $z^{*1}, \ldots, z^{*n}$, call it $\hat{\beta}^*(b)$

Do 1-2 for $b = 1, \ldots, B$

Usual theory (A1-A3) says

$$\operatorname{Var} \hat{\beta}_{OLS} = \sigma^2 (X^T X)^{-1}$$

Can use bootstrap to approximate

$$\widehat{\mathrm{Var}}\,\hat{\beta}_{OLS} \approx \frac{1}{B-p} \sum_{b=1}^{B} \left(\hat{\beta}^*(b) - \hat{\beta}^*(\cdot)\right)\left(\hat{\beta}^*(b) - \hat{\beta}^*(\cdot)\right)^T$$

where $\hat{\beta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^*(b)$.

Approx 95% CI for β_j is

$$\hat{\beta}_j \pm 2\sqrt{\left(\widehat{\mathrm{Var}}\,\hat{\beta}\right)_{jj}}$$

instead of

$$\hat{\beta}_j \pm 2\sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{jj}}$$

Bootstrapping residuals

Start with the original model $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, do OLS, and form <u>estimated residuals</u>

$$\hat{\underline{\varepsilon}} = \underline{Y} - X\hat{\underline{\beta}}_{OLS}$$

with ith estimated resid $\hat{\varepsilon}_i$.

<u>Algorithm</u>

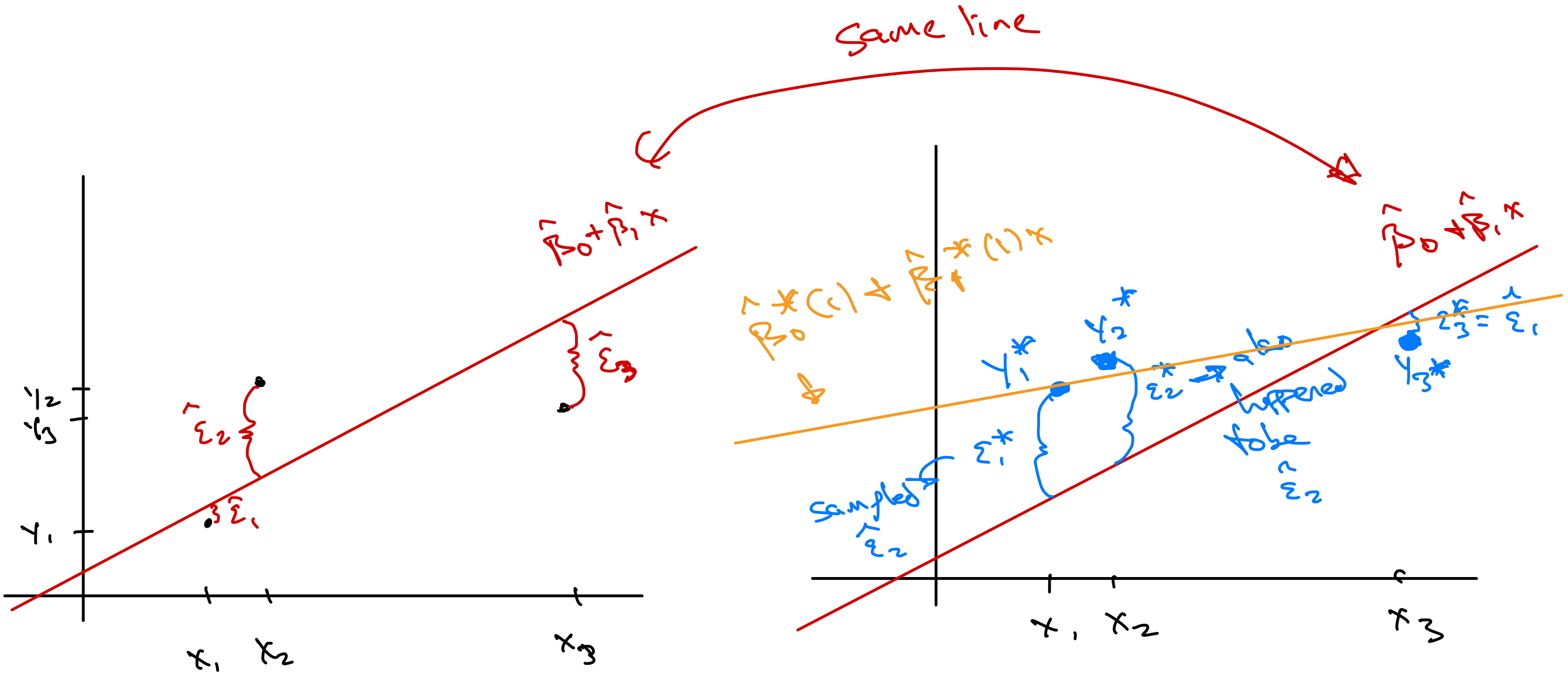
① Generate bootstrap sampled residuals $\varepsilon_1^*, \ldots, \varepsilon_n^*$ with replacement from $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$

② Generate

$$Y_i^* = \underline{x}_i^T \hat{\underline{\beta}}_{OLS} + \varepsilon_i^* \quad , i = 1, \ldots, n$$

③ Calculate OLS estimate of $\underline{\beta}$ based on $\{(\underline{x}_i, Y_i^*)\}$, $\hat{\underline{\beta}}^*(b)$

for $b = 1, \dots, B$.


same line
$\hat\beta_0 + \hat\beta_1 x$
Y_1
Y_2
Y_3
$\hat\varepsilon_1$
$\hat\varepsilon_2$
$\hat\varepsilon_3$
x_1
x_2
x_3
$\hat\beta_0^*(1) + \hat\beta_1^*(1)x$
Y_1^*
Y_2^*
Y_3^*
ε_1^*
ε_2^*
$\varepsilon_3^* = \hat\varepsilon_1$
sampled $\hat\varepsilon_2$
also happened to be $\hat\varepsilon_2$

Which to choose? Pairs vs. resid

Depends on how much you believe assumptions A1-A3!

If ε_i doesn't depend on i, either is ok. If ε_i <u>does</u> depend on i, e.g. errors are heteroskedastic, then residual bootstrap is a bad idea as it breaks relationship between ε_i + x_i.

Parametrically

Assume (+ check!) $\underline{\varepsilon}$ come from some distribution, e.g. $N(0, \sigma^2)$, then follow residual bootstrap but in step 1 generate $\varepsilon_i^* \sim N(0, \hat{\sigma}^2)$

[instead of $\varepsilon_i^* \sim \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$]

Bootstrap for the lasso

Tibshirani (1996) suggested the bootstrap could be used to estimate SEs in the lasso. Knight and Fu (2000) showed it may not work when some β_is $= 0$ or are close to 0.

Chatterjee + Lahiri (2011) found a fix by thresholding small $\hat{\beta}_i$s to zero.