**Spring 2023**

# Bootstrap for Regression

Now, we will consider the bootstrap in the regression problem.

For simplicity, we consider the case where we only have one response variable and one covariate and we will first focus on linear regression. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be the observed data. $Y_i$'s are the response values and $X_i$'s are the corresponding covariate.

The linear regression fits the model

$$\mathbb{E}(Y_i|X_i = x) = \beta_0 + \beta_1 \cdot x$$

and we use the observed data to find the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Or sometimes people write

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i,$$

where each $\epsilon_i$ is a mean 0 noise.

The fitted coefficients have a close formed solution:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad \widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{X}_n.$$

After obtaining the regression coefficients, the *residuals* are

$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i.$$

The quantity $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ is the predicted value of the response given the covariate being $X_i$ based on the fitted linear regression model (sometimes we just call it linear model). In a sense, the residuals represent the random errors that cannot be explained by our linear model.

In what follows, we will introduce several approaches to study the uncertainty (e.g., variance, MSE, or CI) of the fitted parameter $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Namely, we want to estimate quantities like

$$\mathsf{Var}(\widehat{\beta}_0), \mathsf{MSE}(\widehat{\beta}_1).$$

## 1 Empirical Bootstrap

We may apply the idea of empirical bootstrap to the regression problem. In this case, the empirical bootstrap is also called *paired bootstrap*. Given the original sample $(X_1, Y_1), \cdots, (X_n, Y_n)$, we generate a new sets of IID observations

$$(X_1^*, Y_1^*), \cdots, (X_n^*, Y_n^*)$$

such that for each $\ell$,

$$P(X_\ell^* = X_i, Y_\ell^* = Y_i) = \frac{1}{n}, \quad \forall\ i = 1, \cdots, n.$$

Namely, we treat $(X_i, Y_i)$ as one object and we sample with replacement $n$ times from these $n$ objects to form a new bootstrap sample. Thus, each time we generate a set of $n$ new observations from the original dataset.

Assume we repeat the entire process $B$ times, we would obtain

$$\begin{gathered}(X_1^{*(1)}, Y_1^{*(1)}), \cdots, (X_n^{*(1)}, Y_n^{*(1)}) \\ (X_1^{*(2)}, Y_1^{*(2)}), \cdots, (X_n^{*(2)}, Y_n^{*(2)}) \\ \vdots \\ (X_1^{*(B)}, Y_1^{*(B)}), \cdots, (X_n^{*(B)}, Y_n^{*(B)}).\end{gathered}$$

For each bootstrap sample, say $(X_1^{*(\ell)}, Y_1^{*(\ell)}), \cdots, (X_n^{*(\ell)}, Y_n^{*(\ell)})$, we fit the linear regression, leading to a bootstrap estimate of the fitted coefficients $\widehat{\beta}_0^{*(\ell)}, \widehat{\beta}_1^{*(\ell)}$. Thus, the $B$ bootstrap samples leads to

$$(\widehat{\beta}_0^{*(1)}, \widehat{\beta}_1^{*(1)}), \cdots, (\widehat{\beta}_0^{*(B)}, \widehat{\beta}_1^{*(B)}),$$

$B$ sets of fitted coefficients. We then estimate the variance and the MSE by

$$\begin{aligned}\widehat{\mathsf{Var}}_B(\widehat{\beta}_0) &= \frac{1}{B}\sum_{\ell=1}^{B}\left(\widehat{\beta}_0^{*(\ell)} - \bar{\beta}_0^*\right)^2, \quad \bar{\beta}_0^* = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{\beta}_0^{*(\ell)}, \\ \widehat{\mathsf{MSE}}_B(\widehat{\beta}_0) &= \frac{1}{B}\sum_{\ell=1}^{B}\left(\widehat{\beta}_0^{*(\ell)} - \widehat{\beta}_0\right)^2, \\ \widehat{\mathsf{Var}}_B(\widehat{\beta}_1) &= \frac{1}{B}\sum_{\ell=1}^{B}\left(\widehat{\beta}_1^{*(\ell)} - \bar{\beta}_1^*\right)^2, \quad \bar{\beta}_1^* = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{\beta}_1^{*(\ell)}, \\ \widehat{\mathsf{MSE}}_B(\widehat{\beta}_1) &= \frac{1}{B}\sum_{\ell=1}^{B}\left(\widehat{\beta}_1^{*(\ell)} - \widehat{\beta}_1\right)^2.\end{aligned}$$

How about the confidence intervals? We can simply construct them using the variance estimate:

$$\begin{aligned}C.I.(\beta_0) &= \widehat{\beta}_0 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathsf{Var}}_B(\widehat{\beta}_0)}, \\ C.I.(\beta_1) &= \widehat{\beta}_1 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathsf{Var}}_B(\widehat{\beta}_1)}.\end{aligned}$$

This follows from the fact that the fitted coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are roughly normally distributed around the true values $\beta_0$ and $\beta_1$, i.e., there exist $\sigma_0^2$ and $\sigma_1^2$ such that

$$\sqrt{n}(\widehat{\beta}_0 - \beta_0) \xrightarrow{D} N(0, \sigma_0^2), \quad \sqrt{n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{D} N(0, \sigma_1^2).$$

# 2 Residual Bootstrap

Although the empirical bootstrap works well in theory, in practice it might lead to a bad result especially in the presence of influential observations (some $X_i$ very far away from the others). When we do an empirical bootstrap, if we do not select those points, the regression coefficients can be very different.

To resolve this problem, we may use the *residual bootstrap*. Recall that the residuals are

$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i.$$

If we compare the residuals to $\epsilon_i$'s in the regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \Longrightarrow \epsilon_i = Y_i - \beta_0 - \beta_1 X_i.$$

Essentially, each $e_i$ mimics the role of $\epsilon_i$ when the fitted coefficients $\widehat{\beta}_0, \widehat{\beta}_1$ are close to $\beta_0, \beta_1$. The residual bootstrap make good use of this property.

The residual bootstrap first generates IID

$$\widehat{\epsilon}_1^*, \cdots, \widehat{\epsilon}_n^*$$

such that for each $\widehat{\epsilon}_\ell^*$,

$$P(\widehat{\epsilon}_\ell^* = e_i) = \frac{1}{n}, \quad \forall n = 1, \cdots, n.$$

And then generates a new bootstrap sample

$$(X_1^*, Y_1^*), \cdots, (X_n^*, Y_n^*)$$

via

$$X_i^* = X_i, \quad Y_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\epsilon}_i^*. \tag{1}$$

Namely, we fixed the covariate $X_i$ but generate a new value of $Y_i$ using the fitted regression function and the 'noise' from sampling the residuals with replacement.

All the estimate of the variance, MSE, and construction of the CI are the same as the empirical bootstrap. But now we are using the bootstrap samples generated by (1).

# 3 Wild Bootstrap

In addition to the above two approaches, there is another bootstrap for regression–the *wild bootstrap*.

The wild bootstrap is to the residual bootstrap in the sense that we fix the covariates $X_i^* = X_i$ for each $i$ and resample the value of $Y_i$ using the residual $e_i$.

The wild bootstrap first generate IID random variables $V_1, \cdots, V_n \sim N(0, 1)$ and then generate the bootstrap sample

$$(X_1^*, Y_1^*), \cdots, (X_n^*, Y_n^*)$$

by

$$Y_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + V_i \cdot e_i, \quad X_i^* = X_i.$$

Note that the distribution of $V_i$ can be non-Gaussian[1].

Why do we want to use the wild bootstrap over the residual bootstrap? The main reason is that when the variance of error $\mathsf{Var}(\epsilon_i | X_i)$ depends on the value of covariates $X_i$ (this is called *heteroskedasticity*), the residual bootstrap will be unstable because the residual bootstrap will swap all the residuals regardless of the value of covariate. On the other hand, for $i$-th observation, the wild bootstrap uses the residual of itself only.

[1] https://en.wikipedia.org/wiki/Bootstrapping_(statistics)#Wild_bootstrap