1 Bootstrapped Bias and CIs

Given a multiple regression model with mean and variance functions given by $E[y|X] = X\beta$ and $\text{Var}[y|X] = \sigma^2 I$, respectively; assume that we are re-sampling cases, such that we have $z^*_b := \{(y^*_{1b}, x^*_{1b}), \ldots, (y^*_{nb}, x^*_{nb})\}$, for $b = 1, \ldots, B$. Alternatively, we may represent each of these bootstrap samples using matrix notation as $z^*_b = (y^*_b, X^*_b)$.

These bootstrap samples will be contrasted with our original data set, which we denote by $z = (y, X)$.

We will now demonstrate how to conduct statistical inference using these bootstrap quantities, and how to obtain a bootstrap estimate of the bias for our OLS estimator.

1.1 Bootstrapped Mean Squared Error

Recall that the MSE is a measure of the expected distance between a given (fixed) parameter $\theta := g(F)$ and an estimator $\hat{\theta}$, which is a function of the data, such that $\hat{\theta} := s(y)$, for some random vector $y := [y_1, \ldots, y_n]^T$. Here, the functions $g(\cdot)$ and $s(\cdot)$ may be different. However, as long as $F$ is fixed, it follows that $\theta = g(F)$ is also a fixed quantity. Using these definitions, the MSE of $\hat{\theta}$ with respect to $\theta$ is defined as follows,

$$\text{MSE}_F[\hat{\theta}, \theta] := E_F[(\hat{\theta} - \theta)^2] = \text{Var}_F[\hat{\theta}] + (E_F[\hat{\theta}] - \theta)^2,$$

where the second term is referred to as the bias of $\hat{\theta}$ as an estimator of $\theta$, and denoted by $b^2_F(\hat{\theta})$.

We can here use the plug-in principle in order to replace the true (but unknown) distribution $F$ by an empirical estimate such as the EDF based on a sample from $F$. This gives

$$\text{MSE}_F[\hat{\theta}, \theta] = E_F[(\hat{\theta} - \theta)^2] = \frac{1}{B} \sum_{b=1}^B \left(\hat{\theta}(z^*_b) - \theta(z)\right)^2,$$

where $z := (y, X)$ is our original data set.

Using the bootstrap estimates described earlier, $z_1, \ldots, z_B$, we can readily estimate the EDF $\hat{F}$ using the $n$ data points available. The first term in the decomposition of the MSE is simply the squared bootstrapped
standard deviation that we have already encountered,
\[ \text{Var}_{\hat{F}}[\hat{\theta}] = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}(z^*_b) - \bar{\theta}^*)^2, \quad \text{and} \quad b^2_{\hat{F}}(\hat{\theta}) = (\bar{\theta}^* - \theta(z))^2; \]

where the bootstrapped true parameter based on the EDF, \( \theta(z) = g(\hat{F}) \), is solely based on the original data set, \( z = (y, X) \). The rationale for the identity \( \theta(z) = g(\hat{F}) \) involves some basic notions of measure theory can be found in standard textbooks on the bootstrap (see Davison and Hinkley, 1997, p.13). Moreover, the bootstrap mean is given by
\[ \bar{\theta}^* := \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}(z^*_b). \]

Thus, we have obtained the following decomposition of the bootstrapped MSE, which is analogous to the one of the theoretical MSE,
\[ \text{MSE}_{\hat{F}}[\hat{\theta}, \theta] = \text{Var}_{\hat{F}}[\hat{\theta}] + b^2_{\hat{F}}(\hat{\theta}). \]

We have seen that the OLS estimators for simple and multiple regressions are unbiased. Nonetheless, one can use the above quantity as a bootstrap estimate for that bias.

**1.2 Bootstrapped Confidence Intervals**

The procedure is here slightly different. We now wish to compute an estimate of the population distribution of a given normalized statistics, such as
\[ t_1 := \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)}, \]

for a given slope parameter in a multiple regression model, where \( \beta_1 \) is the unknown true value of that parameter. Again, the confidence interval for \( \beta_1 \) can also be expressed using the unknown distribution \( F \),
\[ \mathbb{P}_{\hat{F}} \left[ -t(\alpha/2; n - p^*) \leq \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\beta_1)} \leq t(\alpha/2; n - p^*) \right] = 1 - \alpha. \]

which indicates that the value of the true parameter \( \beta_1 \) has a \((1 - \alpha)\) probability of being comprised in that interval. Here, \( t(\alpha/2, n - p^*) \) denotes the \((1 - \alpha/2)\)th-percentile of a \( t \)-distributed random variable with \( n - p^* \) degrees of freedom.

The bootstrap approach for this problem is to use the bootstrap samples, \( z^*_b \) for computing a large number of normalized quantities, which will represent the estimate of the CIs for \( \beta_1 \) under the EDF. That is, we compute
\[ t_1^*(b) := \frac{\hat{\beta}_1(z^*_b) - \beta_1(z)}{\text{se}^*_b(\beta_1)}. \]

Once we have constructed this distribution of bootstrap replicates of the \( t \)-statistic of \( \beta_1 \), it suffices to take the 2.5\textsuperscript{th} and 97.5\textsuperscript{th} percentile of that distribution to obtain a bootstrapped CI of \( \beta_1 \). Note, however, that this distribution needs not be symmetric and therefore we are forced to explicitly compute these two different percentiles.
### Table 1. Classical “confusion” matrix.

<table>
<thead>
<tr>
<th></th>
<th>Accept $H_0$</th>
<th>Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>$\mathbb{P}[X &lt; \theta^*</td>
<td>H_0]$</td>
</tr>
<tr>
<td></td>
<td>True Negative Rate</td>
<td>False Positive Rate</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>Type I Error</td>
</tr>
<tr>
<td></td>
<td>$1 - \alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Population</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_1$</td>
<td>$\mathbb{P}[X &lt; \theta^*</td>
<td>H_1]$</td>
</tr>
<tr>
<td></td>
<td>False Negative Rate</td>
<td>True Positive Rate</td>
</tr>
<tr>
<td></td>
<td>Type II Error</td>
<td>Sensitivity (Power)</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>$1 - \beta$</td>
</tr>
</tbody>
</table>

---

2 Permutation Tests

The classical example of a permutation test is the evaluation of the statistical significance of a correlation coefficient. Let a set of pairs of $n$ realizations from two distinct random variables, denoted $(y_i, x_i)$ with $i = 1, \ldots, n$. The entire data set may be denoted as follows,

$$z = (y, x).$$

The two random variables of interest need not be normally distributed. We will draw permutation samples from this data set by breaking the pairwise associations between the $y_i$’s and the $x_i$’s, such that we uniformly select a permutation of the indexes denoted $\pi$ from the distribution of all permutations on an alphabet of $n$ characters. Such a permutation is a function of the form,

$$\pi : \{1, \ldots, n\} \rightarrow \{y_1, \ldots, y_n\},$$

with for instance, $\pi(2) = y_n$ and $\pi(5) = y_5$. There are $n!$ such permutations of the indexes of the $y_i$’s.

The permutation samples can then be obtained by applying one of these permutations to the original data set. For every $b = 1, \ldots, B$, we have

$$z_b^* := \left\{(y_{\pi(1)}^*, x_1^*), \ldots, (y_{\pi(n)}^*, x_n^*)\right\} =: (y_b^*, x_b^*).$$

Observe that although we are only permuting the indexes of the $y_i$’s, we nonetheless succeed to break the “symmetry” in the pairwise associations between these two sets of realizations. This lack of association is exactly what would be expected under the null hypothesis, since this would yield a lack of correlation.

As for the bootstrap, it is now possible to compute a distribution of permutation test statistics, under the null hypothesis of no association, such that we can derive the correlation coefficients,

$$\hat{r}_b := \text{Cor}(y_b^*, x_b^*), \quad b = 1, \ldots, B.$$  

This can be contrasted with the observed value for that statistic,

$$\hat{r} := \text{Cor}(y, x).$$
Finally, we can compute the permutation p-value, assuming that we are considering single-tailed test,

$$p := \mathbb{P}_{\text{perm}} \left[ \hat{r}_b > |\hat{r}| \right] = \frac{1}{B} \sum_{b=1}^{B} I \{ \hat{r}_b > |\hat{r}| \},$$

We will reject the null hypothesis of no association between $x$ and $y$ if we find $p \leq \alpha$.

The main limitation of this non-parametric inferential method is its reliance on a very large number of permutations. In the above example, if $B = n!$, then we obtain a so-called exact test. Since this number grows very (very) rapidly, practitioners are generally using a random sample from these permutations, and conduct a Monte Carlo approximation of this exact test. However, not using the full set of permutations can lead to a substantial increase in the type I error rate. This inflation of the number of false positives follows from the fact that the values in the tails of the null distribution are, by definition, less likely to be sampled than the ones in the main body of the distribution.

## 3 Polynomial Regression

### 3.1 Quadratic Regression

The linear mean function that we have considered so far can be extended through the inclusion of a quadratic term, such that

$$\mathbb{E}[Y_i | x_i] = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2, \quad i = 1, \ldots, n.$$  

Naturally, the use of a quadratic term does not imply that the observed values will reach a maximum or a minimum in the range of the data available. However, such terms may capture the non-linear nature of the relationship between the $y_i$’s and the $x_i$’s.

For bivariate linear regression, this could be extended to include all the quadratic terms alongside the interaction terms between the two covariates included in the model,

$$\mathbb{E}[Y_i | x_{i1}, x_{i2}] = \beta_0 + \beta_{11} x_{i1} + \beta_{22} x_{i2} + \beta_{12} x_{i1} x_{i2}. \quad (1)$$

Observe that the number of terms in a polynomial regression increases rapidly with the number of covariates in the model. For $p$ covariates, we require $p$ linear terms, $p$ quadratic terms, and $p(p-1)/2$ first-order interaction terms.

Even for two covariates, however, the interpretation of the meaning of the different terms in a quadratic model is extremely difficult. Consider a small perturbation of $X_1$ in the model described in equation (1),

$$\mathbb{E}[Y_i | x_{i1} + \delta, x_{i2}] = \beta_0 + \beta_1 (x_{i1} + \delta) + \beta_{22} x_{i2} + \beta_{12} (x_{i1} + \delta) x_{i2}.$$ 

Subtracting these two mean functions, we obtain three different terms,

$$\mathbb{E}[Y_i | x_{i1} + \delta, x_{i2}] - \mathbb{E}[Y_i | x_{i1}, x_{i2}] = (\beta_1 \delta + \beta_{12} \delta^2) + 2 \beta_{12} x_{i1} \delta.$$

Thus, even when considering a perturbation for a single covariate, the interpretation of the effect is somewhat arduous.

### 3.2 Higher-order Polynomials

Naturally, one can extend this framework by including higher-order terms in the mean function of a polynomial regression. Ignoring interaction terms, we may have up to $K$-th order transformations of $p$ different covariates,

$$\mathbb{E}[Y_i | x_{i1}, \ldots, x_{ip}] = \beta_0 + \sum_{j=1}^{p} \sum_{k=1}^{K} \beta_{jk} x_{ij}^k, \quad i = 1, \ldots, n.$$
Statisticians, however, face a law of diminishing returns. As they include increasingly higher terms in a regression model, the contributions of every single term is likely to decrease, and this is more likely to result in modeling the random variation in the data, which is unlikely to produce accurate out-of-sample predictions.

### 3.3 Criticisms of Polynomial Regression

1. The number of parameters increase rapidly with the number of covariates.

2. There is a greater risk of over-fitting the data, as we add more non-linear transformations of the covariates. As a rule of thumb, it is often said that 80% of the relationship between two variables is likely to be linear, and that only the remaining 20% may be non-linear.

3. The localization of the non-linear effect may be difficult to identify exactly. That is, while the observed variable may exhibit a non-linear effect for some subsets of the values of the predictor, it could be linear for the rest of the values of that predictor.

4. Finally and most importantly, the interpretation of the high-order terms and interactions may be difficult. This is a general criticism that can be leveled to any non-linear model with several variables. This difficulty of interpretation partly accounts for the enduring popularity of standard linear models.

### References