Introduction and Review

February 7, 2020

The material in this set of notes was initially based on Chapter 2 of Robert Shumway and David Stouffer's Time Series Analysis and Its Applications: With R Examples, but has since grown to include additional material. Notes on the parametric bootstrap are based on Efron and Tibshirani's 1993 textbook An Introduction to the Bootstrap. Chapter 7 of Simon Wood's Core Statistics is also a helpful and relevant reference.

Notation

- $\mathbb{E}[x]$ refers to the expectation of x;
- $\mathbb{V}[x]$ refers to the variance of x;
- $\mathbb{Z}$ refers to the set of all positive and negative integers, $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$;
- Bolded lowercase letters denote column vectors;
- Bolded uppercase letters denote matrices;
- $x \sim \mathcal{N}(\mu, \sigma^2)$ indicates that x is normally distributed with mean μ and variance σ^2;
  - Another way of writing this is $x = \sigma v + \mu$, where $v \sim \mathcal{N}(0, 1)$. I will sometimes use this notation to describe the distribution of x.

- $x \sim \mathcal{T}_k$ indicates that x is central t-distributed with k degrees of freedom;
- $x \sim \chi^2_k$ indicates that x is chi-square distributed with k degrees of freedom;
- $x \sim \mathcal{F}_{k,j}$ indicates that x is central F-distributed with k and j degrees of freedom.
- $||\boldsymbol{x}||_2^2 = \sum_{i=1}^n x_i^2$;
- $x \approx y$ denotes that x is approximately equal to y.
- $\mathbb{1}_{\{x=1\}}$ and $(x = 1)$ are two different ways that I will denote an indicator function, which is equal to 1 if the statement contained in the brackets $\{\cdot\}$ or parentheses $(\cdot)$ is true and equal to 0 otherwise.

Basic Idea!

Most (univariate) time series analysis problems boil down to observing an $n \times 1$ vector $\boldsymbol{y} = (y_1, \ldots, y_n) = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu}$ is a fixed but unknown mean and $\boldsymbol{\epsilon}$ are mean zero random errors, and:

- **Estimating $\boldsymbol{\mu}$**;
- **Predicting** future values $y_{n+1}, \ldots, y_{n+k}$.

Time series analysis problems differ from classical statistical problems because elements of $\boldsymbol{y}$ are ordered in time. Several examples of time series data and problems are given in Chapter 1 of S&S, pages 4-11.

Because elements of $\boldsymbol{y}$ are ordered in time, consecutive elements of $\boldsymbol{y}$ may be correlated and classical statistical methods may not work well. This is easiest to see via example. Suppose we assume $\mu_{x,t} = \mu$ for all $t = 1, \ldots, n$, and we are interested in estimating μ. Ignoring the time series aspect of $\boldsymbol{y}$, we assume ϵ_j are independent and identically distributed with known variance σ^2. The classical approach would be to compute a point estimate of μ, $\hat{\mu} = \sum_{t=1}^n y_t/n$ and corresponding standard error, σ^2/n. Is this accurate?

The classical approach gives a **incorrect** standard error if elements of $\boldsymbol{y}$ are correlated. What would be the correct standard error?

$$\begin{aligned}\mathbb{E}\left[(\hat{\mu}-\mu)^2\right] =& \mathbb{E}\left[\left(\sum_{t=1}^{n} y_t - \mu\right)^2 / n^2\right] \\ =& \sigma^2/n + \sum_{t=1}^{n}\sum_{t'=1,t'\neq t}^{n} \mathbb{E}\left[(y_t-\mu)(y_{t'}-\mu)\right]/n^2.\end{aligned}$$

The correct standard error depends on covariances of elements of $\boldsymbol{y}$,

$$\mathbb{E}\left[(y_t-\mu)(y_{t'}-\mu)\right] = \mathbb{E}\left[\epsilon_t \epsilon_{t'}\right],$$

which may be nonzero if elements of $\boldsymbol{y}$ are ordered in time!

Regression Review (S&S 2.1-2.2)

Many methods for time series analysis build on **linear regression**. We perform linear regression when we are interested in expressing an $n \times 1$ response vector $\boldsymbol{y}$ as a linear function of r $n \times 1$ covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$, i.e. we want to find regression coefficients $\beta_1, \ldots, \beta_r$ such that $\boldsymbol{y} \approx \beta_1 \boldsymbol{x}_1 + \cdots + \beta_r \boldsymbol{x}_r$. If $\boldsymbol{y}$ is a time series, then covariates might include:

- Indicators for distinct time periods different elements of $\boldsymbol{y}$ belong to;
- A vector $\boldsymbol{t}$, where t_i is the time y_i was observed or the order of y_i in the sequence;
- Nonlinear functions of elements of $\boldsymbol{t}$, e.g. $z_{ij} = \sin(t)$ for some $j \in \{1, \ldots, p\}$;
- Lagged values of $\boldsymbol{y}$;
- Lagged values of a different but related time series.

We will very rarely be able to describe $\boldsymbol{y}$ as an exactly linear function of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$. Instead, we try to find the "best" way of writing $\boldsymbol{y}$ as a nearly linear function of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$

by computing the regression coefficients $\boldsymbol{\beta}$ that solve:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \beta_1 \boldsymbol{x}_1 - \cdots - \beta_r \boldsymbol{x}_r||_2^2 . \tag{1}$$

This is easier to express concisely in matrix form. Letting $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r]$ be the $n \times r$ matrix of regression coefficients, $\boldsymbol{\beta}$ equivalently solves:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 . \tag{2}$$

We refer to the quantity $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ as the **residual sum of squares (RSS)**, as it measures how much of the variability of $\boldsymbol{y}$ remains after subtracting off a linear function of the covariates. We can minimize (2) by differentiating; the minimizing value $\hat{\boldsymbol{\beta}}$ will satisfy:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0} \implies \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}.$$

If the matrix $\boldsymbol{X}$ is full rank with rank r, then the minimizing value is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{y}. \tag{3}$$

If we want to say more about $\hat{\boldsymbol{\beta}}$, we need to make some more assumptions. First, note that we can always decompose the observed response $\boldsymbol{y}$ into a linear part $\boldsymbol{X}\boldsymbol{\beta}$ and a remainder $\boldsymbol{\epsilon}$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4}$$

If we assume:

- $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}$, then $\hat{\boldsymbol{\beta}}$ is **unbiased**, i.e. $\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$.
- $\epsilon_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, then:
  - ($\star$) $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$;
  - ($*$) $\hat{\boldsymbol{\beta}} \sim \text{normal}\left(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$;
  - ($\dagger$) $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \sim \text{normal}\left(\boldsymbol{0}, \sigma^2 \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\right)$;
  - ($\circ$) $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are independent.

Time series methods rely extensively on likelihood based inference, so we pause to derive $(\star)$. If $\epsilon_j \overset{i.i.d.}{\sim} \text{normal}\,(0, \sigma^2)$ then rearranging (4) gives $\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta} \overset{i.i.d.}{\sim} \text{normal}\,(0, \sigma^2)$, where $\boldsymbol{x}_i$ is the i-th row of $\boldsymbol{X}$. This yields the likelihood:

$$l\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2\right\}.$$

Finding the value of $\boldsymbol{\beta}$ that maximizes the likelihood is equivalent to finding the value of $\boldsymbol{\beta}$ that minimizes the negative log likelihood, which corresponds to a constant plus the residual sum of squares (2). As a side note - we can actually eliminate σ^2 from the log-likelihood by *profiling* it out, thus making the connection between maximizing the log likelihood and computing the residual sum of squares even clearer. The negative log-likelihood is given by:

$$\frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2.$$

We want to minimize this with respect to $\boldsymbol{\beta}$ and σ^2, and this can be achieved by *first* minimizing over σ^2 for fixed $\boldsymbol{\beta}$, and plugging in the minimizing value of σ^2 to get a minimization problem that depends on $\boldsymbol{\beta}$ alone. This takes advantage of the observation that:

$$\min_{\sigma^2,\boldsymbol{\beta}} \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 = \max_{\boldsymbol{\beta}}\left(\max_{\sigma^2} \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2\right).$$

There's a nice closed form solution to the minimization with respect to σ^2 for fixed $\boldsymbol{\beta}$. Taking derivatives with respect to σ^2 and rearranging, we get:

$$\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 = 0 \implies \sigma^2 = \frac{\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2}{n}.$$

Then plugging this expression for the minimizing value of σ^2 for fixed $\boldsymbol{\beta}$, we get a minimiza-

tion problem over just $\boldsymbol{\beta}$:

$$\begin{aligned}\min_{\sigma^2,\boldsymbol{\beta}} \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 &= \min_{\boldsymbol{\beta}}\left(\min_{\sigma^2}\frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2\right) \\ &= \min_{\boldsymbol{\beta}}\left(\frac{n}{2}\right)\log\left(2\pi\left(\frac{\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2}{n}\right)\right) + \frac{\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2}{2\left(\frac{\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2}{n}\right)} \\ &= \min_{\boldsymbol{\beta}}\left(\frac{n}{2}\right)\log\left(2\pi\left(\frac{\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2}{n}\right)\right) + \frac{n}{2}.\end{aligned}$$

Now only one term depends on the residual sum of squares $\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2$, and since $\log(\cdot)$ is a strictly increasing function, this is equivalent to minimizing the residual sum of squares. This isn't very useful to us now, because the log-likelihood is already easy to maximize, and because the maximum likelihood estimate of $\boldsymbol{\beta}$ does not depend on σ^2. However we will find the idea of profiling very useful when we start considering maximum likelihood estimation for time series models, in which elements of the error vector $\boldsymbol{\epsilon}$ are correlated and we have more parameters to optimize over.

Returning to the claims we can make about the least squares estimate $\hat{\boldsymbol{\beta}}$ when we assume that $\epsilon_j \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, $(*)$, $(\dagger)$, and $(\circ)$ are very useful. They allow us not only to compute standard errors and confidence intervals for $\hat{\boldsymbol{\beta}}$ but also to test the null hypothesis that β_i is exactly equal to a specific value or that a subset of $p-p_1$ elements $\boldsymbol{\beta}_{-1} = \left(\beta_{t_1},\ldots,\beta_{t_{p-p_1}}\right)$ are jointly exactly equal to $\mathbf{0}$.

Standard practice for constructing standard errors and confidence intervals is to use $(*)$, plugging in an unbiased estimator of the variance:

$$s^2 = \frac{\left|\left|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right|\right|_2^2}{n-p}. \tag{5}$$

Note that this is *not* the maximum likelihood estimate of σ^2 - the maximum likelihood estimator $\hat{\sigma}^2 = \left|\left|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right|\right|_2^2/n$ is biased.

It follows from $(*)$, $(\dagger)$, and $(\circ)$ that

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{ii}}} \sim \mathcal{T}_{n-p}. \tag{6}$$

This gives us a way of testing the null hypothesis that β_i is exactly equal to a specific value because it tells us the approximate distribution of $\hat{\beta}_i$ for specific values of β_i. We call such tests ***t*-tests** or ***z*-tests**.

Similarly, letting $\boldsymbol{X}_1 = \left[\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_{p_1}}\right]$ be the design matrix containing the p_1 columns corresponding to elements of $\boldsymbol{\beta}_1$ and letting $\hat{\boldsymbol{\beta}}_1$ be the linear regression estimate of $\boldsymbol{\beta}_1$ from regressing $\boldsymbol{y}$ on just the p_1 columns of $\boldsymbol{X}$ contained in $\boldsymbol{X}_1$, it follows from $(*)$, $(\dagger)$, and $(\circ)$ that

$$F_{p-p_1,n-p} = \left(\frac{\left|\left|\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right|\right|_2^2 - \left|\left|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right|\right|_2^2}{\left|\left|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right|\right|_2^2} \right) \left(\frac{n-p}{p-p_1} \right) \sim \mathcal{F}_{p-p_1,n-p}. \tag{7}$$

This gives us a way of testing the null hypothesis that the $p - p_1$ elements of $\boldsymbol{\beta}_{-1}$ are jointly exactly equal to $\mathbf{0}$ by giving us an approximate distribution of $F_{p-p_1,n-p}$ under the null. We call such tests ***F*-tests**.

Parametric Bootstrap for Standard Errors, Confidence Intervals, and Testing
The derivation of standard errors, confidence intervals, and testing given above are all based on the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{normal}\left(\mathbf{0}, \sigma^2\boldsymbol{I}_n\right). \tag{8}$$

In all cases, we are computing quantities that tell us about the expected variability of a function of the data $\boldsymbol{y}$ and $\boldsymbol{X}$ across repeated realizations of data $\boldsymbol{y}$, assuming a model of the form (8) holds. In general, we'll denote the quantity we are interested in understanding the expected variability of as $c = g\left(\boldsymbol{y}, \boldsymbol{X}\right)$.

For example, when we are interested in computing the standard error of $\hat{\beta}_j$, we would be interested in understanding the variability of $c = g\left(\boldsymbol{y}, \boldsymbol{X}\right) = \hat{\beta}_j$. As another example,

if we were interested in testing the null hypothesis that $\beta_j = 0$, we would be interested in understanding the variability of $c = g(\boldsymbol{y}, \boldsymbol{X}) = \frac{\hat{\beta}_i}{s\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{-1}_{ii}}}$.

A natural approach is to simulate synthetic realizations of the observed data from the approximate model,

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{normal}\left(\boldsymbol{0}, s^2\boldsymbol{I}_n\right), \tag{9}$$

plugging in our estimates $\hat{\boldsymbol{\beta}}$ and s^2 for the unknown parameters $\boldsymbol{\beta}$ and σ^2. This approach is commonly referred to as the **Parametric Bootstrap**, and it is especially useful when we are interested in understanding the expected variability of functions of the unknown parameters for which the distribution under (8) is either only known approximately, e.g. for large n relative to p, or is not known at all. Intuitively, the parametric bootstrap will work well as long as $\hat{\boldsymbol{\beta}}$ and s^2 are "good enough" estimates of $\boldsymbol{\beta}$ and σ^2 and as long as the distribution of the quantity we are interested c in does not vary "too much" as a function of $\boldsymbol{\beta}$ and σ^2.

In practice, the parametric bootstrap proceeds as follows:

- Estimate $\boldsymbol{\beta}$ and σ^2 from the model given by (8), and compute $c = g(\boldsymbol{y}, \boldsymbol{X})$.
- Set a desired number of bootstrap samples, n_{boot}.
- For k in $1, \ldots, n_{boot}$:
  - Draw $\boldsymbol{y}^{(k)} \sim \text{normal}\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}, s^2\boldsymbol{I}_n\right)$;
  - Compute $c^{(k)} = g\left(\boldsymbol{y}^{(k)}, \boldsymbol{X}\right)$.

This procedure produces n_{boot} simulated values of the quantity of interest $c^{(1)}, \ldots, c^{(n_{boot})}$ from the approximate model given by (9), which can be used for many purposes!

- An estimate of the standard error of c can be obtained by computing the sample standard deviation of the simulated values $c^{(1)}, \ldots, c^{(n_{boot})}$.
- An approximate level-α test of the null hypothesis that c is equal to zero under the

model (9) can be obtained by comparing c to the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated values $c^{(1)}, \ldots, c^{(n_{boot})}$.

- This is not the only or best way to construct a $1-\alpha$ interval for a quantity c based on bootstrap samples $c^{(1)}, \ldots, c^{(n_{boot})}$, but it is very intuitive so we will use it here. You can find more information about alternative methods in Bootstrap Methods and their Applications (1997) by Davison and Hinkley, or other more recent textbooks or papers on the construction of bootstrap-based confidence intervals.

Model Selection Tests are very useful for **model selection**, i.e. for choosing the covariates to include in our model. Model selection is especially relevant in linear regression methods for time series analysis, e.g. we may need to decide which lagged values of $\boldsymbol{y}$ to include as covariates. Letting $\boldsymbol{X}_k$ refer to a matrix containing k covariates and $\boldsymbol{\beta}_k$ and $\hat{\boldsymbol{\beta}}_k$ the corresponding regression coefficients and their linear regression estimates, several popular methods for performing model selection when performing linear regression are:

($*$) Perform an t- or z-test comparing *nested* models with k and $k' = k+1$ covariates.

($*$) Perform an F-test comparing *nested* models with k and k' covariates.

($\star$) Compute **Akaike's Information Criterion (AIC)**

$$AIC = \ln\left(\frac{\left|\left|\boldsymbol{y} - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k\right|\right|_2^2}{n}\right) + \frac{n+2k}{n} \tag{10}$$

for models with k and k' covariates, and choose the model with the lower AIC value.

($\star$) Compute **AIC, Bias Corrected (AICc)**

$$AICc = \ln\left(\frac{\left|\left|\boldsymbol{y} - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k\right|\right|_2^2}{n}\right) + \frac{n+k}{n-k-2} \tag{11}$$

for models with k and k' covariates, and choose the model with the lower $AICc$ value.

($\star$) Compute **Schwarz's/Bayesian Information Criterion (SIC/BIC)**

$$SIC = \ln\left(\frac{\left|\left|\boldsymbol{y} - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k\right|\right|_2^2}{n}\right) + \frac{k\log(n)}{n} \tag{12}$$

for models with k and k' covariates, and choose the model with the lower SIC value.

($\star$) Choose any measure of model performance, denoted as $f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})$, define several equally sized, possibly overlapping subsets of the data (training datasets), for each subset fit the models and evaluate the performance of each model on the remaining data (test data), and choose the model that performs best on average across all of the training datasets. This is commonly referred to as **cross-validation**. There are many different ways of performing cross-validation:

- Leave-k-out cross-validation refers to a type of cross-validation that constructs l training data sets by choosing k elements of the full data, and constructing the training data to be the full data with the k chosen elements removed, and the test data to be the chosen k elements. In this case, the average measure of performance is an average across l training data sets.
  - When $k = 1$, we will often set $l = n$ because there are n ways to remove a single observation from the full data. When $k > 1$, we will often set $\binom{l<}{n,k}$ and choose the l subsets of k elements each at random, because it can become computationally prohibitive to average across all $\binom{n}{k}$ possible ways to remove k observations from the full data.
- k-fold cross-validation refers to a type of cross-validation that forms k training datasets by dividing the data into k subsets of size n/k, and then defining each test dataset to be one of the subsets of n/k observations in the full data, and each training dataset as the remaining observations. In this case, the average measure

of performance is an average across k training data sets.

* This corresponds to leave-$\binom{n}{n/k}$-out cross-validation with $l = k$ total subsets/training datasets.

– A common measure of performance is squared error loss, $f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$.

Note that the methods marked with $(*)$ require that the two models be *nested*, i.e. the columns in $\boldsymbol{X}_k$ must be a subset of the columns in $\boldsymbol{X}_{k'}$ or vice versa. The procedures denoted with $(\star)$ are not. Whether AIC, AICc, or BIC is most appropriate for a given problem is problem-specific; AICc can perform better than AIC when n is relatively small, and SIC/BIC can perform better than AIC when the number of covariates k is relatively large. Cross-validation also does not require that the two models be nested, and can be performed in many different ways depending on the choice of k and the choice of a measure of model performance. A very popular choice is 10-fold cross-validation using squared error loss to measure performance on the test data, $f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$.