Introduction and Review

January 22, 2019

The material in this set of notes is based on Chapters 1-2 of S&S.

Notation

- $\mathbb{E}[x]$ refers to the expectation of x;
- $\mathbb{V}[x]$ refers to the variance of x;
- Bolded lowercase letters denote column vectors;
- Bolded uppercase letters denote matrices;
- $x \sim \mathcal{N}\left(\mu, \sigma^2\right)$ indicates that x is normally distributed with mean μ and variance σ^2;
- $x \sim \mathcal{T}_k$ indicates that x is central t-distributed with k degrees of freedom;
- $x \sim \xi_k^2$ 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.

Basic Idea!

Most (univariate) time series analysis problems boil down to observing an $n \times 1$ vector $\boldsymbol{x} = (x_1, \dots, x_n) = \boldsymbol{\mu}_x + \boldsymbol{w}$, where $\boldsymbol{\mu}_x$ is a fixed but unknown mean and $\boldsymbol{w}$ are mean zero random errors, and:

- **Estimating $\boldsymbol{\mu}_x$.**
- **Predicting** future values $x_{n+1}, \dots, x_{n+k}$.

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

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

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

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

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

$$\mathbb{E}\left[\left(x_t - \mu_x\right)\left(x_{t'} - \mu_x\right)\right] = \mathbb{E}\left[w_t w_{t'}\right],$$

which may be nonzero if elements of $\boldsymbol{x}$ 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 $\boldsymbol{x}$ as a linear function of q $n \times 1$ covariate vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_q$, i.e. we want to find regression coefficients $\beta_1, \dots, \beta_q$ such that $\boldsymbol{x} \approx \beta_1 \boldsymbol{z}_1 + \cdots + \beta_q \boldsymbol{z}_q$. If $\boldsymbol{x}$ is a time series, then covariates might include:

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

We will very rarely be able to describe $\boldsymbol{x}$ as an exactly linear function of $\boldsymbol{z}_1, \dots, \boldsymbol{z}_q$. Instead, we try to find the "best" way of writing $\boldsymbol{x}$ as a nearly linear function of $\boldsymbol{z}_1, \dots, \boldsymbol{z}_q$ by computing the regression coefficients $\boldsymbol{\beta}$ that solve:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{x} - \beta_1 \boldsymbol{z}_1 - \cdots - \beta_q \boldsymbol{z}_q||_2^2 . \tag{1}$$

This is easier to express concisely in matrix form. Letting $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_q]$ be the $n \times q$ matrix of regression coefficients, $\boldsymbol{\beta}$ equivalently solves:

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{x} - \boldsymbol{\beta}\boldsymbol{Z}||_2^2 . \tag{2}$$

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

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

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

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

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 $\boldsymbol{x}$ into a linear part $\boldsymbol{Z}\boldsymbol{\beta}$ and a remainder $\boldsymbol{w}$:

$$\boldsymbol{x} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{w}. \tag{4}$$

If we assume:

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

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

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

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).

($*$), ($\dagger$), 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 q_1 elements $\boldsymbol{\beta}_1 = \left(\beta_{t_1}, \dots, \beta_{t_{q_1}}\right)$ are jointly exactly equal to a specific value.

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

$$s_w^2 = \frac{\left|\left|\boldsymbol{x} - \boldsymbol{Z}\boldsymbol{\beta}\right|\right|_2^2}{n-q}. \tag{5}$$

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

$$t_{n-q} = \frac{\hat{\beta}_i - \beta_i}{s_w\sqrt{\left(\boldsymbol{X}'\boldsymbol{X}\right)_{ii}^{-1}}} \sim \mathcal{T}_{n-q}. \tag{6}$$

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**.

Similarly, letting $\boldsymbol{Z}_1 = \left[\boldsymbol{z}_{t_1}, \dots, \boldsymbol{z}_{t_{q_1}}\right]$ be the design matrix containing the q_1 columns corresponding to elements of β_1 and letting $\hat{\boldsymbol{\beta}}_1$ be the linear regression estimate of β_1 from regressing $\boldsymbol{x}$ on just the q_1 columns of $\boldsymbol{Z}$ contained in $\boldsymbol{Z}_1$, it follows from $(*)$, $(\dagger)$, and $(\circ)$ that

$$F_{q-q_1,n-q} = \left(\frac{\left|\left|\boldsymbol{x} - \hat{\boldsymbol{\beta}}_1\boldsymbol{Z}_1\right|\right|_2^2 - \left|\left|\boldsymbol{x} - \hat{\boldsymbol{\beta}}\boldsymbol{Z}\right|\right|_2^2}{\left|\left|\boldsymbol{x} - \hat{\boldsymbol{\beta}}\boldsymbol{Z}\right|\right|_2^2}\right)\left(\frac{n-q}{q-q_1}\right) \sim \mathcal{F}_{q-q_1,n-q}. \tag{7}$$

This gives us a way of testing the null hypothesis that the elements of $\boldsymbol{\beta}_1$ are jointly exactly equal to a specific value because it tells us the approximate distribution of $\hat{\boldsymbol{\beta}}_1$ for specific values of $\boldsymbol{\beta}_1$. We call such tests ***F*-tests**.

F-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 lagged values of $\boldsymbol{x}$ as covariates. Letting $\boldsymbol{Z}_k$ 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 F-test comparing *nested* models with k and k' covariates.

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

$$AIC = \ln\left(\frac{\left|\left|\boldsymbol{x} - \boldsymbol{Z}_k\hat{\boldsymbol{\beta}}_k\right|\right|_2^2}{n}\right) + \frac{n+2k}{n} \tag{8}$$

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

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

$$AICc = \ln\left(\frac{\left|\left|\boldsymbol{x} - \boldsymbol{Z}_k\hat{\boldsymbol{\beta}}_k\right|\right|_2^2}{n}\right) + \frac{n+k}{n-k-2} \tag{9}$$

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

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

$$SIC = \ln\left(\frac{\left|\left|\boldsymbol{x} - \boldsymbol{Z}_k\hat{\boldsymbol{\beta}}_k\right|\right|_2^2}{n}\right) + \frac{k\log(n)}{n} \tag{10}$$

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

Note that ($*$) requires that the two models be *nested*, i.e. the columns in $\boldsymbol{Z}_k$ must be a subset of the columns in $\boldsymbol{Z}_{k'}$ 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 k is relatively large.