Day 13 - Intro to Regression

Introduction

Trends in time series can be classified as either *stochastic*, *deterministic*, or both. Recently, we have considered *stochastic* models for time series, culminating in the AR(p) process. Stochastic time series models are adept at explaining serial autocorrelation, but can struggle to explain large structural effects, such as trends or seasonality.

To remedy this problem, we often pair stochastic models with models capable of accounting for deterministic trends, such as regression. In today's lab, we will begin to explore time series as a tool for modeling

---

Regression overview

Linear model theory

A ______________________________ is a model of the form

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

where ϵ_i is ____ distributed $N(0, \sigma^2)$. In matrix notation, the above model is equivalent to

$$y = X\beta + \epsilon$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

By properties of normal distributions, the above model is equivalent to

$$y \sim \mathcal{N}(X\beta, \Sigma)$$

where $\Sigma = \sigma^2 \mathcal{I}_n$ and $\mathcal{I}_n$ is an $n \times n$ identity matrix. To fit the model, we must estimate _______________ and __________________. The estimated regression equation is written as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p}$$

Parameter estimates are obtained by minimizing the _________________________________. The error in regression is called a _____________________,

$$e_i = y_i - \hat{y}_i$$

The ____, which minimizes the ______________, is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

Assumptions

1.

2.

3.

4.

Linear models in R

The following functions are useful when working with regression models in R:

- `lm`
- `plot`
- `fitted`
- `resid`
- `predict`
- `model.matrix`

Air Passengers activity

1. Load the `AirPassengers` data set and create a data frame that includes the count of air passengers, the year, the month, and a time index ($t = 1, 2, \dots, n$). Plot the count of air passengers by the time index. Then create a second data set, called `ap_sub`, that contains the air passenger measurements between 1949 and 1959.

> **! Important**
>
> Be sure to treat the month column as a `factor`!

2. Fit a regression model, called `fit`, that models the passenger count by the time index plus the month (be sure month is treated as a `factor`) for the `ap_sub` data frame. Print out a summary of the model, and use the summary to write out the estimated regression model.

3. Recreate the slope coefficient estimates by calculating the OLS estimator by hand using `model.matrix`. The following R functions may be useful:
   - `solve` computes a matrix inverse
   - `t` computes the transpose of a matrix
   - `%*%` computes matrix multiplication

4. Interpret the slope coefficient associated with the time index.

5. Interpret the slope coefficient associated with the adjustment to the intercept for August.

6. Assess the linearity assumption for the regression model by running `plot(fit, which = 1)`

7. Assess the constant variance assumption for the regression model by using the previous plot and running `plot(fit, which = 3)`

8. Assess the normality assumption for the regression model by running `plot(fit, which = 2)`

9. Assess the independence assumption by thinking about the problem and running `acf(resid(fit))` and `pacf(resid(fit))`

10. One way to address the violations of the normality and constant variance assumptions is to log transform the response. To address the violation of the linearity assumption, it may be helpful to create a squared time index variable. Create new columns in the `ap_sub` data frame that log the number of passengers and computes the square of the time index. Then fit a new model, called `fit_log`, that models the logged passenger count by the time index, time index squared, and month.

11. Reassess the assumptions using the new model.

12. If the residuals are positively serially correlated, we will tend to underestimate the standard errors of the regression coefficients. What impact does this have when determining statistical significance of regression coefficients? It may be helpful to know that the t-test provided in the R output is calculated as

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

13. With the exception of the independence assumption, hopefully you are convinced that the model created in 10 is a reasonable model for the `AirPassengers` data set. Next, calculate the fitted values by hand and compare them to the values from `fitted`.

14. Plot the observed and fitted time series on a single plot and comment on how well the model estimates the observed relationship.

15. Calculate the residuals by hand and compare them to the values obtained from `resid`.

16. Use the `fit_log` model to forecast the time series for the year of 1960 (which we had previously excluded) using the `predict` function. The predict function requires a `newdata` argument to obtain forecasts - this data frame must include all the predictors used in the model. Plot the observed, fitted, and forecasted series on a single plot and include a 95% prediction interval for the forecasted series. For more about predicting from `lm` models, run `?predict.lm`.

17. You should notice that the model tends to consistently overestimate the peaks early on, then consistently overestimate the peaks later on. Why do you think that is?

> The model we have fit assumes that the effect of month is constant over time, which does not appear to be the case.

18. One way to remedy the problem described in the previous question is to create an *interaction* between t and the month, which allows the estimated effect of month to depend on t. The code below fits such an interaction model. Recreate the figure from question 16 and comment on the differences.

```r
# create a new fit_log that includes the interaction
fit_log <- lm(log_ap ~ t*month + t2, ap_sub)
```