Time Series Analysis

Basics of Time Series Analysis

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Basic Statistical Modeling Concepts

Georgia Tech

About This Lesson

Georgia Tech

Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \ldots, x_{1,p}), y_1\}, \ldots, \{(x_{n,1}, \ldots, x_{n,p}), y_n\}$
Model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \ i = 1, \ldots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent random variables
- $\varepsilon_i \sim$ Normally distributed *for confidence/prediction intervals, hypothesis testing*

Georgia Tech

Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \ldots, x_{1,p}), y_1\}, \ldots, \{(x_{n,1}, \ldots, x_{n,p}), y_n\}$

Model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \ i = 1, \ldots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent random variables
- $\varepsilon_i \sim$ Normally distributed *for confidence/prediction intervals, hypothesis testing*

The model parameters are: $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$

- **Unknown regardless how much data are observed**
- **Estimated given the model assumptions**
- **Estimated based on data**

Georgia Tech

Parameter Estimation $(\beta_0, \beta_1, \ldots, \beta_p), \sigma^2$

To estimate $(\beta_0, \beta_1, \ldots, \beta_p)$, find values $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ that minimize squared error:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_p x_{i,p})\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

By linear algebra (Orthogonal Decomposition Theorem) or differentiation:

$$\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$$

So

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Parameter Estimation (cont'd)

The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, so

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

where $\mathbf{H} \equiv \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ is called the **hat matrix** because multiplying $\boldsymbol{y}$ by $\mathbf{H}$ gives $\hat{\boldsymbol{y}}$.

The residuals are:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$$

To estimate σ^2,

$$\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}/(n - p - 1)$$

Georgia Tech

Model Interpretation

The Least Squares estimated coefficients have specific interpretations:

$\widehat{\boldsymbol{\beta}}_0$ The estimated expected value of the response variable when all predicting variables equal zero.

$\widehat{\boldsymbol{\beta}}_i$ The estimated expected change in the value of the response variable associated with one unit of change in the value of the i^{th} predicting variable (i.e., associated with a one-unit change in $\boldsymbol{x}_i$, where i is any of $1, \ldots, p$), holding all other predictors in the model fixed (i.e., holding fixed $\boldsymbol{x}_j$ for $j = 1, \ldots, p$ where $j \neq i$).

Sampling Distribution

$$\mathrm{E}(\hat{\beta}) = \beta$$
$$\mathrm{V}(\hat{\beta}) = \sigma^2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} = \Sigma$$

Furthermore, $\widehat{\boldsymbol{\beta}}$ is a linear combination of $\{y_1,...,y_n\}$. If we assume that $\varepsilon_i \sim$ Normal $(0, \sigma^2)$, then $\widehat{\boldsymbol{\beta}}$ is also distributed as $\widehat{\boldsymbol{\beta}} \sim$ Normal $(0, \sigma^2)$.

$$\widehat{\boldsymbol{\beta}} \sim \mathbf{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

Sampling Distribution (cont'd)

$$\hat{\boldsymbol{\beta}} \sim \mathbf{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

σ^2 is unknown!

Replace σ^2 with $\hat{\sigma}^2 = \text{MSE}$

$$\hat{\sigma}^2 = \frac{\sum \hat{\sigma}_i^2}{n-p-1} \sim \chi^2_{n-p-1}$$

(chi-squared distribution with $n-p-1$ degrees of freedom)

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{V}\left(\hat{\beta}_j\right)}} \sim t_{n-p-1}$$

(*t*-distribution with $n-p-1$ degrees of freedom)

Georgia Tech

Confidence Interval Estimation

We can derive confidence intervals for β_j using this t sampling distribution:

$$\hat{\beta}_j \pm \left(t_{\alpha/2,\, n-p-1}\right)\left(\mathrm{SE}(\hat{\beta}_j)\right)$$

Is β_j statistically significant?

- Check whether zero is in the confidence interval

Why is this a *t*-interval?

Testing Statistical Significance

To test for statistical significance of β_j given all other predicting variables in the model, use a t-test for $\mathrm{H_0}$ and $\mathrm{H_a}$:

$$\mathrm{H_0}: \beta_j = 0$$
vs.
$$\mathrm{H_a}: \beta_j \neq 0$$

$$t\text{−value} = \frac{\hat{\beta}_j - 0}{\mathrm{SE}(\hat{\beta}_j)}$$

- Reject $\mathrm{H_0}$ if |t-value| gets too large
- Interpret rejecting the null hypothesis as β_j being statistically significant

Georgia Tech

General Hypothesis Testing

How will the procedure change if

$$H_0: \beta_j = b \quad \text{vs.} \quad H_a: \beta_j \neq b$$

for some known null value b?

we test

$$t\text{−value} = \frac{\hat{\beta}_j - b}{\text{SE}(\hat{\beta}_j)}$$

- Reject H_0 if $|t$-value$|$ is large
  - For significance level α, if $|t\text{−value}| > t_{\alpha/2,\, n-p-1}$ ⟶ reject H_0
- Alternatively, compute a p-value based on the probability that the t distribution is greater than the t-value:

$$\text{p-value} = 2\text{Prob}(T_{n\text{-}p\text{-}1} > |t\text{-value}|)$$

- If p-value is small (e.g., < 0.01) ⟶ reject H_0

Georgia Tech

Testing Subsets of Coefficients

Partial F-test:

- Consider a full model with two sets of predictors, X_1, ..., X_p (perhaps controlling factors) and (Z_1, ..., Z_q) (perhaps additional explanatory factors):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q + \varepsilon$$

- Test whether any of the Z factors add explanatory power to the model:

$$\mathbf{H_0}\text{: } \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0 \quad \textbf{vs.} \quad \mathbf{H_a}\text{: } \alpha_{\mathrm{i}} \neq 0 \text{ for at least one } \alpha_{\mathrm{i}}, \mathrm{i} = 1, \cdots, q$$

$$\mathrm{F-statistic} = \mathrm{F}_{partial} = \frac{\mathrm{SSReg}(Z_1, \ldots, Z_q | X_1, \ldots, X_p)/q}{\mathrm{SSE}(Z_1, \ldots, Z_q, X_1, \ldots, X_p)/(n-p-q-1)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large (F-statistic $> \mathrm{F}_{\alpha,\, q,\, n\text{-}p\text{-}q\text{-}1}$)
  - At least one coefficient is different from zero at the α significance level

Georgia Tech

Normality Transformation

Problem: Constant variance or/and normality assumption

Solution: Transform the response variable from $\boldsymbol{y}$ to $\boldsymbol{y}^*$ via

$$\boldsymbol{y}^* = \boldsymbol{y}^{\lambda}$$

where the value of λ depends on how $\mathrm{Var}(\boldsymbol{y})$ changes as x changes.

$\sigma_y(x) \propto const$	$\lambda = 1$	(don't transform)
$\sigma_y(x) \propto \sqrt{\mu_x}$	$\lambda = 1/2$	$\boldsymbol{y}^* = \sqrt{\boldsymbol{y}}$
$\sigma_y(x) \propto \mu_x$	$\lambda = 0$	$\boldsymbol{y}^* = \ln(\boldsymbol{y})$
$\sigma_y(x) \propto \mu_x^2$	$\lambda = -1$	$\boldsymbol{y}^* = \frac{1}{\boldsymbol{y}}$

Georgia Tech

Outliers in Regression

A data point far from the majority of the data (in *y* and/or any *x*) may be called an *outlier*, especially if it does not follow the general trend of the rest of the data.

- Data points that are far from the means of the *X*s or near the edge of the observation space are called *leverage points*.
- A data point that is far from the means of *y* and/or an *x* is called an *influential point* if it influences the fit of the regression.
- Excluding a leverage point may or may not the regression fit significantly, thus a leverage point may or may not be an influential point.

The upshot: Sometimes there are good reasons to exclude subsets of data (e.g., errors in data entry or experimental errors). Sometimes an outlier belongs in the data. Outliers should always be examined.

R^2: Coefficient of Determination

A measure that efficiently summarizes how well the *X*s can be used to predict *Y* is **R^2** (called *R-squared* or the *coefficient of determination*):

$$R^2 = 1 - SSE/SST$$

where

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R^2 is interpreted as the proportion of total variability in *Y* that can be explained by the linear regression model.

Georgia Tech

Multicollinearity

Recall that finding the ordinary least squares estimator of $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

depends on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ being invertible (nonsingular or nondegenerate). From linear algebra, a square matrix is invertible if and only if its columns are linearly independent (i.e., no column is a linear combination of the others).

If that doesn't hold, the ordinary least squares estimator of $\widehat{\boldsymbol{\beta}}$ doesn't exist. That's probably due to a specification error where one or more predictors should be eliminated as redundant (e.g., if years and number of rings were included in a model for trees).

Even if the columns of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ are linearly independent, some problems might arise if the value of one predictor can be closely estimated from the other predictors. We call this condition *multicollinearity* or *near collinearity*.

Multicollinearity (cont'd)

- Indications that near collinearity is present:
  - The estimated coefficients $\widehat{\boldsymbol{\beta}}$ are unstable: When the value of one predictor changes slightly, the fitted regression coefficients change dramatically
  - The standard error of $\widehat{\boldsymbol{\beta}}$ is artificially large
  - The overall F statistic is significant, but individual *t*-statistics are not
- Prediction may be affected
  - The relationship to the response may change widely
- Some computational algorithms are sensitive to multicollinearity
- <u>But no</u> inflation or deflation in $\mathbf{R^2}$

Summary

Georgia Tech