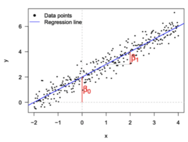
Lecture 1

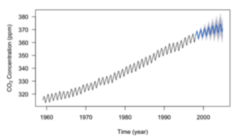
Course Information and Review

Reading: Forecasting, Time Series, and Regression (4th edition) by Bowerman, O'Connell, and Koehler [Link]: Chapters 1 and 2

MATH 4070: Regression and Time-Series Analysis

Whitney Huang
Clemson University





Notes

Agenda

1. About the Instructor
2. Class Policies
3. Review

Notes

About the Instructor

Notes

Instructor Background

- Assistant Professor of Applied Statistics and Data Science
- Born in Laramie, WY, and raised in Taiwan


- Obtained a B.S. in Mechanical Engineering; transitioned to Statistics in graduate school


- Earned a Ph.D. in Statistics from Purdue University in 2017

Notes

How to Reach Me?

- **Email** ✉: wkhuang@clemson.edu
  Please include [MATH 4070] in your email subject line
- **Office:** O-221 Martin Hall
- **Office Hours:** Tue., Wed., and Thurs., 1:45 pm - 2:30 pm, and by appointment

Notes

Class Policies

Notes

Logistics

- There will be some (4-6) homework assignments:
  - To be uploaded to Canvas by 11:59 pm ET on the due dates
  - Worst grade will be dropped
- There will be three 60-minute exam. The (tentative) dates are: Sep. 24, Tuesday; Oct. 22, Tuesday; Nov. 21, Thursday
- There will be a final project. It could be a **data analysis**, a **simulation study**, **methodological or theoretical research**, or a **report on a research article** of interest to you

Notes

Evaluation

Grades will be weighted as follows:

Homework	30%
Exam I	15%
Exam II	15%
Exam III	20%
Final Project	20%

Final course grades will be assigned using the following grading scheme:

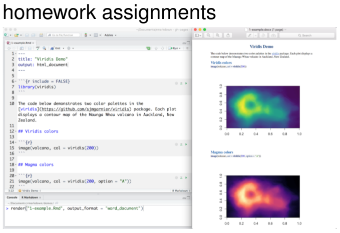
>= 90.00	A
80.00 ~ 89.99	B
70.00 ~ 79.99	C
60.00 ~ 69.99	D
<= 59.99	F

Notes

Computing

We will use software to perform statistical analyses. Specifically, we will be using R/Rstudio

- a **free/open-source** programming language for statistical analysis
- available at https://www.r-project.org/ (R); https://rstudio.com/ (Rstudio)
- I strongly encourage you to use **R Markdown** for homework assignments



Notes

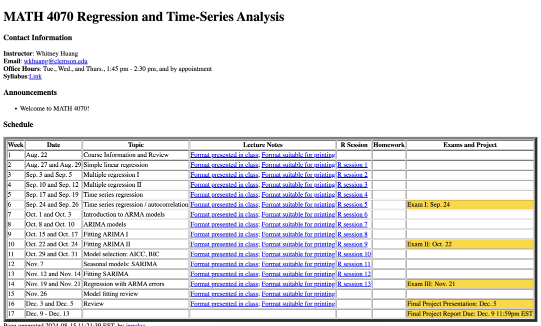
Course Materials at CANVAS

- Course syllabus / announcements
- Lecture slides/notes/videos
- R Codes
- Data sets

Notes

Course Website

Link: https://whitneyhuang83.github.io/MATH4070/Schedule.html



MATH 4070 Regression and Time-Series Analysis

Contact Information

Instructor: Whitney Huang
Email: wkhuang@clemson.edu
Office Hours: Tue., Wed., and Thurs., 1:45 pm - 2:30 pm, and by appointment
Syllabus:Link

Announcements

- Welcome to MATH 4070!

Schedule

Week	Date	Topic	Lecture Notes	R Session	Homework	Exams and Project
1	Aug. 22	Course Information and Review	Format presented in class; Format suitable for printing			
2	Aug. 27 and Aug. 29	Simple linear regression	Format presented in class; Format suitable for printing	R session 1		
3	Sep. 3 and Sep. 5	Multiple regression I	Format presented in class; Format suitable for printing	R session 2		
4	Sep. 10 and Sep. 12	Multiple regression II	Format presented in class; Format suitable for printing	R session 3		
5	Sep. 17 and Sep. 19	Time series regression	Format presented in class; Format suitable for printing	R session 4		
6	Sep. 24 and Sep. 26	Time series regression / autocorrelation	Format presented in class; Format suitable for printing	R session 5		Exam I: Sep. 24
7	Oct. 1 and Oct. 3	Introduction to ARMA models	Format presented in class; Format suitable for printing	R session 6		
8	Oct. 8 and Oct. 10	ARIMA models	Format presented in class; Format suitable for printing	R session 7		
9	Oct. 15 and Oct. 17	Fitting ARIMA I	Format presented in class; Format suitable for printing	R session 8		
10	Oct. 22 and Oct. 24	Fitting ARIMA II	Format presented in class; Format suitable for printing	R session 9		Exam II: Oct. 22
11	Oct. 29 and Oct. 31	Model selection: AICC, BIC	Format presented in class; Format suitable for printing	R session 10		
12	Nov. 7	Seasonal models: SARIMA	Format presented in class; Format suitable for printing	R session 11		
13	Nov. 12 and Nov. 14	Fitting SARIMA	Format presented in class; Format suitable for printing	R session 12		
14	Nov. 19 and Nov. 21	Regression with ARMA errors	Format presented in class; Format suitable for printing	R session 13		Exam III: Nov. 21
15	Nov. 26	Model fitting review	Format presented in class; Format suitable for printing			
16	Dec. 3 and Dec. 5	Review	Format presented in class; Format suitable for printing			Final Project Presentation: Dec. 5
17	Dec. 9 - Dec. 13					Final Project Report Due: Dec. 9 11:59pm EST

Page generated 2024-08-15 11:21:39 EST, by jemdoc.

Course Information and Review
School of Mathematical and Statistical Sciences, Clemson University
About the Instructor
Class Policies
Review

Notes

Tentative Schedule

Week	Dates	Topic
1	8/22	Overview of the course
2	8/27-29	Simple linear regression
3	9/3-5	Multiple regression I
4	9/10-12	Multiple regression II
5	9/17-19	Time series regression
6	9/24-26	TS regression/ autocorrelation
7	10/1-2	Intro to ARMA models
8	10/8-10	ARIMA models
9	10/17	Fitting ARIMA I
10	10/22-10/24	Fitting ARIMA II
11	10/29-10/31	Model selection: AICC, BIC
12	11/7	Seasonal models: SARIMA
13	11/12-14	Fitting SARIMA
14	11/19-21	Regression with ARMA errors
15	11/26	Model fitting review
16	12/3-5	Review and Project Presentation

Course Information and Review
School of Mathematical and Statistical Sciences, Clemson University
About the Instructor
Class Policies
Review

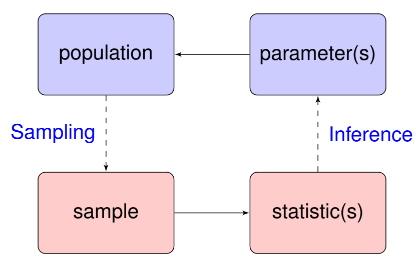
Notes

Review

Notes

Population (parameters) vs. Sample (statistics)

- We use parameters to describe the population and statistics to describe the sample
- **Statistical Science** involves using sample information to infer about populations



Notes

Example

Population is Clemson students and variable Y is IQ

- μ is the average IQ of all Clemson students (we don't know this)
- σ^2 is the variance of IQ in the whole student body (don't know this either)
- Randomly select $n = 36$ students and administer an IQ test to them. Suppose the average IQ score in the sample is 116, with a sample variance of 256
- Note that different samples yield different sample means and variances, but the population mean and variance remain constant. This variation in sample means reflects the sampling properties of the sample mean

Notes

Some Properties of the Sample Mean

Consider a random sample: $Y_1, Y_2, \cdots, Y_n$

- For any outcome of the sample, $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$, where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
- The theoretical average of the sample mean is the population mean:
$$\mathbb{E}[\bar{Y}] = \mu$$
$\Rightarrow$ average over all possible sample means we get the population mean
- The variance of the sample mean is
$$\text{Var}(\bar{Y}) = \text{E}\left[(\bar{Y} - \mu)^2\right] = \frac{\sigma^2}{n}$$
$\Rightarrow$ the average "distance" between $\bar{Y}$ and μ is $\frac{\sigma}{\sqrt{n}}$

Notes

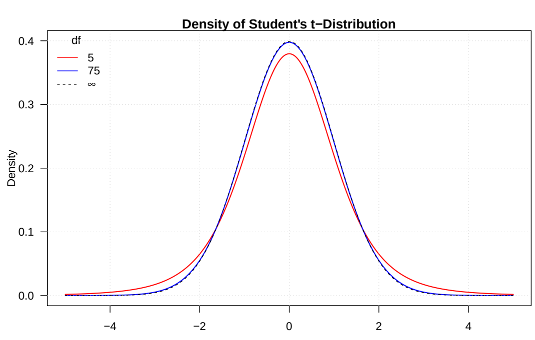
Statistical Inference

Statistical inference is the process of using sample data to draw conclusions about a population

- Tools
  - Confidence intervals
  - Hypothesis tests
- These require distributional assumptions
- If our population variable has a normal distribution, for each sample
$$t = \frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}}$$
is a draw from a t-distribution with degrees of freedom $(\text{df}) = n - 1$

Notes

Stundet-t Distribution



Notes

Inference on μ for Normal Samples: Confidence Interval

95% confidence interval:

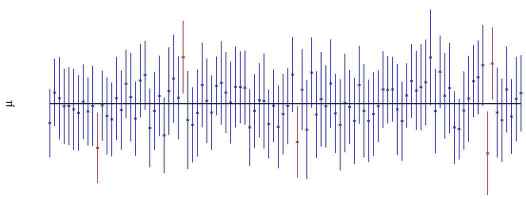
$$\left(\bar{Y} - t_{0.975,df=n-1}\frac{s}{\sqrt{n}}, \bar{Y} + t_{0.975,df=n-1}\frac{s}{\sqrt{n}}\right),$$

where $t_{0.975,\mathrm{df}=n-1}$ denotes the 0.975 quantile of the t distribution with $\mathrm{df} = n - 1$.

- This interval contains μ in 95% of samples, meaning each (random) sample has a 95% chance that its CI includes μ (see next slide for a demonstration)
- The interval gives a likely range for μ. For example, if the interval is $(3.4, 8.6)$, it is unlikely that $\mu < 3$ or $\mu > 10$

Notes

A Demonstration of Confidence Intervals



- The black horizontal line represents the true population mean μ, which is unknown but fixed
- Each vertical line represents a confidence interval around a sample mean, constructed from different samples drawn from the same population

Notes

Inference on μ for Normal Samples: Hypothesis Test

Say you want to conclude that the average IQ of Clemson students is greater than 110.

Null hypothesis $H_0 : \mu \le 110;$

Alternative hypothesis $H_1 : \mu \ge 110.$

Note:

- The alternative hypothesis is what we want to show
- The hypotheses do not depend on any sample

Now take a sample of $n = 36$ students: $\bar{y} = 112$ and $s = 16$. If μ were 110 (H_0)

$$t = \frac{\bar{y} - 110}{\frac{16}{\sqrt{36}}} = 0.75, \text{ and } \mathbb{P}(t_{35} > 0.75) = 0.229.$$

$\Rightarrow$ there is up to a 22.9% chance that $\bar{y} \ge 112$ if $\mu \le 110$. Not too convincing. Can't conclude that $\mu \ge 110$ from this sample

Notes

Hypothesis Test Cont'd

$$\text{Null hypothesis } H_0 : \mu \leq 110;$$
$$\text{Alternative hypothesis } H_1 : \mu \geq 110.$$

- If instead $n = 36, \bar{y} = 116$ and $s = 16$. If μ were 110 (H_0)

  $$t = \frac{116 - 110}{\frac{16}{\sqrt{36}}} = 2.25, \text{ and } \mathbb{P}(t_{35} > 2.25) = 0.0154.$$

  $\Rightarrow$ If $\mu \leq 110$, the chance of getting $\bar{y} \geq 116$ is at most 0.0154. Since this is **unlikely**, we reject H_0 and conclude that $\mu \geq 110$. This outcome provides strong evidence that the average population IQ exceeds 110

- Here, the p-value = 0.0154. A small p-value indicates the likelihood of obtaining our result (in the direction of H_1) if H_0 were true, suggesting that H_0 should be rejected in favor of H_1

Notes

A Connection to Calculus: Mean Squared Error

Course Information and Review
About the Instructor
Class Policies
Review

Consider taking a measurement Y (random variable). If we were to approximate Y with a single number, what would be the best choice?

Consider minimizing

$$g(c) = \mathbb{E}\left[(Y - c)^2\right] = \mathbb{E}[Y^2] + c^2 - 2c\mathbb{E}[Y].$$

Take the derivative on the left hand side and solve $g'(c_0) = 0$ to solve for minimum

Solution

$$c_0 = \mathbb{E}[Y] = \mu$$

$\Rightarrow$ we say μ is the best mean squared error (MSE) constant predictor of Y

Notes

A Little Linear Algebra

Course Information and Review
About the Instructor
Class Policies
Review

Recall that for real-valued vectors

$$\mathbf{u} = (u_1, u_2, \cdots, u_n)^T, \quad \mathbf{v} = (v_1, v_2, \cdots, v_n)^T,$$

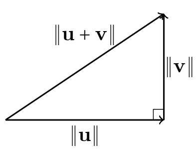
where the superscript T denotes the transpose. The inner product between $\mathbf{u}$ and $\mathbf{v}$ is

$$\mathbf{u}^T\mathbf{v} = \sum_{i=1}^{n} u_i v_i.$$

The vectors are orthogonal if the inner product is 0, and in that case

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2,$$

where $\|\mathbf{u}\|^2 = \mathbf{u}^T\mathbf{u} = \sum_{i=1}^{n} u_i^2$.



Notes

A Connection to Linear Algebra

Consider the sample outcome as a vector:

$$\mathbf{y} = (y_1, y_2, \cdots, y_n)^T.$$

Approximate each component by μ, estimated by $\bar{y}$.

$$\mathbf{y} - \boldsymbol{\mu} = (\hat{\mathbf{y}} - \boldsymbol{\mu}) + (\mathbf{y} - \hat{\mathbf{y}}),$$

where $\hat{\mathbf{y}} = (\bar{y}, \bar{y}, \cdots, \bar{y})^T$ and $\boldsymbol{\mu} = (\mu, \mu, \cdots, \mu)^T$.

Since the first and second vector on the RHS are orthogonal (why?):

$$\|\mathbf{y} - \boldsymbol{\mu}\|^2 = \|\hat{\mathbf{y}} - \boldsymbol{\mu}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

Notes

A Connection to Linear Algebra: Remarks

- $\mathbf{y}$ consists of ordinary n-vectors of real numbers
- The vector $\hat{\mathbf{y}} - \boldsymbol{\mu}$ is a one-dimensional object since all its components have the same value
- The vector $\mathbf{y} - \hat{\mathbf{y}}$ is an $n-1$ dimensional object since its components sum to 0 (one linear restriction)
- The sample variance is related to the squared norm of $\mathbf{y} - \hat{\mathbf{y}}$:

  $$s^2 = \frac{(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})}{n-1}$$

  Notice that the denominator (df) represents the dimension of $\mathbf{y} - \hat{\mathbf{y}}$.

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

Notes

Chi-Square Distribution

Let $Y_1, Y_2, \cdots, Y_n$ be independent with

$$Y_j \sim \mathrm{N}(\mu_j, \sigma^2).$$

Then

$$\chi^2 = \sum_{j=1}^{n} \left(\frac{Y_j - \mu_j}{\sigma}\right)^2$$

has a chi-square distribution with n degrees of freedom.
Note that the df is the dimension of outcomes of the data vector.
Now say $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^T$ takes outcomes in k-dimensions $(k < n)$ with

$$\mathbb{E}(\hat{\mathbf{y}}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T, \qquad (\hat{\mathbf{y}} - \boldsymbol{\mu})^T (\mathbf{y} - \hat{\mathbf{y}}) = 0$$

Then

- $\frac{(n-k)\hat{\sigma}^2}{\sigma^2} = \frac{(\hat{\mathbf{y}}-\boldsymbol{\mu})^T(\hat{\mathbf{y}}-\boldsymbol{\mu})}{\sigma^2} \sim \chi^2_{\text{df} = n-k}$; $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$
- $\hat{\mathbf{y}}$ is independent of $\hat{\sigma}^2$

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

Notes

F- and t- Distributions

Let $Y_1, Y_2, \cdots, Y_n$ be independent with

$$Y_j \sim \mathrm{N}(\mu_j, \sigma^2),$$

$\hat{\mathbf{y}}$ takes outcomes in k-dimensions $(k < n)$ with

$$\mathbb{E}(\hat{\mathbf{y}}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T, \qquad (\hat{\mathbf{y}} - \boldsymbol{\mu})^T (\mathbf{y} - \hat{\mathbf{y}}) = 0$$

Then for any real vector $\mathbf{a} = (a_1, a_2, \cdots, a_n)^T$,

$$T = \frac{\sum_{i=1}^n a_i(\hat{y}_i - \mu_i)}{\hat{\sigma}\sqrt{\sum_{i=1}^n a_i^2}} = \frac{(\hat{\mathbf{y}} - \boldsymbol{\mu})^T \mathbf{a}}{\sqrt{\hat{\sigma}^2 \mathbf{a}^T \mathbf{a}}}$$

is a draw from a t-distribution with $\mathrm{df} = n - k$

Also,

$$F = \frac{(\hat{\mathbf{y}} - \boldsymbol{\mu})^T (\hat{\mathbf{y}} - \boldsymbol{\mu})/k}{\hat{\sigma}^2}$$

is a draw from an F-distribution with $\mathrm{df}_1 = k$ and $\mathrm{df}_2 = n - k$[1]

[1] Note: the textbook uses s^2 to denote the estimated varaince.

Notes

Example: 2 Sample t-Test

Let's assume that we have two independent samples, each with a sample size of $n = 10$, and we want to infer the mean difference $\mu_M - \mu_F$:

- Set $\mathbf{a} = (\frac{1}{10}, \frac{1}{10}, \cdots, \frac{1}{10}, \frac{-1}{10}, \frac{-1}{10}, \cdots, \frac{1}{10})^T$ and let

$$T = \frac{\hat{\mu}_F - \hat{\mu}_M - (\mu_M - \mu_F)}{\hat{\sigma}\sqrt{\frac{2}{10}}}$$

- Reject $\mathrm{H}_0 : (\mu_M - \mu_F) \leq 0$ if the p-value < 0.05, where $T_{\text{obs}} = \frac{\hat{\mu}_m - \hat{\mu}_F}{\hat{\sigma}\sqrt{\frac{2}{10}}}$, and

$$p\text{-value} = \mathbb{P}(t_{n-2} > T_{\text{obs}}).$$

- A 95% confidence interval for $(\mu_M - \mu_F)$ is

$$(\hat{\mu}_M - \hat{\mu}_F) \pm t_{0.975, \mathrm{df}=n-2}\hat{\sigma}\sqrt{\frac{2}{10}}$$

Notes

Review of Main Concepts

- Population parameters are inferred from data using statistics as estimators.
- Statistics are random variables when the data is a random sample.
- The mean is the best MSE predictor. The mean vector $\hat{\mathbf{y}}$ can be estimated from a data vector, with variance estimated by $s^2 = \frac{(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}})}{(n-k)}$.
- The t- and F-distributions arise from independent sampling from normal distributions with equal variance. The df of $\hat{\mathbf{y}}$ is k, and the df of the variance estimate determines the df of the t-distribution $(n - k)$.

Notes

Standard Error for Normal Models

Let $Y_1, Y_2, \cdots, Y_n$ be independent with $Y_j \sim \mathrm{N}(\mu_j, \sigma^2)$

$$\mathbb{E}(\hat{\mathbf{y}}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T;$$
$$(\hat{\mathbf{y}} - \mu)^T(\mathbf{y} - \hat{\mathbf{y}}) = 0;$$
$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})}{n-k}$$

For $\hat{\theta} = \sum_{i=1}^n a_i \hat{y}_i$

$$\sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{\sigma^2 \mathbf{a}^T \mathbf{a}}$$

The standard error of $\hat{\theta}$ is

$$\mathrm{se}(\hat{\theta}) = \sqrt{\hat{\sigma}^2 \mathbf{a}^T \mathbf{a}}$$

Course Information and Review

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

About the Instructor

Class Policies

Review

Notes

t-Distribution Revisited

Under the setup from the previous slide:

$$\mathbb{E}(\hat{\theta}) = \theta = \sum_{i=1}^n a_i \mu_i.$$

Then

$$T = \frac{\hat{\theta} - \theta}{\mathrm{se}(\hat{\theta})}$$

has a t-distribution with $\mathrm{df} = n - k$

Course Information and Review

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

About the Instructor

Class Policies

Review

Notes

Two Sample t-Test Revisited

Take two independent random samples

$$Y_1, Y_2, \cdots, Y_n \sim \mathrm{N}(\mu_1, \sigma^2), \quad X_1, X_2, \cdots, X_m \sim \mathrm{N}(\mu_2, \sigma^2)$$

Estimate the means as

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}; \quad \bar{X} = \sum_{j=1}^m \frac{X_j}{m}$$

Estimate the variance with

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{j=1}^m (X_j - \bar{X})^2}{n+m-2} = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

By independent of the two samples

$$\mathrm{Var}(\bar{Y} - \bar{X}) = \mathrm{Var}(\bar{Y}) + \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$$

$$\mathrm{se}(\bar{Y} - \bar{X}) = s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

Course Information and Review

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

About the Instructor

Class Policies

Review

Notes

Two Sample t-Test

From the previous slide, we have

$$T = \frac{(\bar{Y} - \bar{X}) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

has a t-distribution with $\mathrm{df} = n + m - 2$

Notes

Summary

In this lecture, we reviewed:

- Statistical Inference: Confidence Intervals and Hypothesis Testing
- The t-distribution, F-distribution, χ^2 distribution, and their applications
- Two-sample t-tests

In the next lecture, we will begin exploring Regression Analysis, starting with Simple Linear Regression

Notes

Notes