Introduction to Regression Analysis

Jerome Dumortier

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Topics Covered

Empirical research approach

- Scientific method

Data analysis and modeling

- Overview of regression models and concepts

Large language models (e.g., ChatGPT)

- Use of artificial intelligence in data analysis

Review of statistical concepts within the context of R/RStudio

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Hypothesis, Model, and Data

The empirical research approach in the social sciences consists of multiple steps:

1 Statement of theory or hypothesis

- Example: *"People increase their consumption when their income increases by less than the income increase."*

2 Specification of the mathematical model

$$\Delta consumption = \beta_0 + \beta_1 \cdot \Delta income$$

3 Obtaining the data

- Real personal consumption expenditures per capita
- Real disposable personal income per capita

4 Estimation of the parameters of the econometric model

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Estimation Results

```
##
## ================================================
##                       Dependent variable:
##                   ----------------------------
##                       diff(consumption)
## ------------------------------------------------
## diff(income)               0.549***
##                            (0.109)
## Constant                  231.954**
##                           (85.745)
## ------------------------------------------------
## Observations                 39
## R2                          0.409
## Adjusted R2                 0.393
## Residual Std. Error   326.315 (df = 37)
## F Statistic           25.567*** (df = 1; 37)
## ================================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Post-Estimation Procedures

6. Hypothesis testing on whether results are aligned with the theory

7. Forecasting or prediction

   - Example: Effect of income increase due to economic growth on consumption

8. Model use for policy purposes

   - Example: Effects of stimulus spending on the economy

Above approach at the core of social science research such as public administration, criminal justice, economics, sociology:

- Influence of trust and attitudes on the organic food purchases
- Relationship between automatic bill payment and electricity consumption

Caveat: Ideological bias in the production of research findings

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Bivariate and Multivariate Regression

Bivariate regression model:

- One dependent (e.g., home value) and one independent variable (e.g., square footage)

$$homevalue_i = \beta_0 + \beta_1 \cdot sqft_i + \epsilon_i$$

- Useful to explain the mechanics of ordinary least square (OLS) models

Multivariate regression model

- One dependent (e.g., home value) and multiple independent variables (e.g., square footage, bedrooms)

$$homevalue_i = \beta_0 + \beta_1 \cdot sqft_i + \beta_2 \cdot bedrooms_i + \epsilon_i$$

Assumptions required for consistent coefficient estimates.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling

Linear Regression

Qualitative Choice
and Limited
Dependent Variables

Panel Data

Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts

Confidence Intervals

Hypothesis Testing

Advanced Multivariate Regression

Concepts associated with the independent variables

- Dummy variables to describe a qualitative characteristic
- Use of natural logarithm to transform the dependent and/or independent variables
- Functional forms including squared terms
- Interaction effects if the marginal effect of one variable dependents on the level of another variable

Model misspecification

- Effects of inclusion of irrelevant or exclusion of relevant variables

Regression diagnostics and tests

Relaxing assumptions

Heteroscedasticity vs. homoscedasticity

- Non-constant variance of the error term
- Tests to detect heteroscedasticity (i.e., Goldfeld-Quandt test, Breusch-Pagan-Godfrey test)

Multicollinearity

- Variance Inflation Factor

Autocorrelation of error terms

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Qualitative Choice Models

Binary choice (i.e., probit and logit) models

- Dependent variable: 1 or 0 ("yes" or "no")
- Example: Recidivism (i.e., committing a crime after release from prison)

Ordered logit

- More than two categories for the dependent variable but ordered
- Example: Level of support for a particular policy (e.g., opposed, neutral, supportive)

Multinomial logit

- More than two categories of the dependent variable but no order
- Example: Commute to campus by bike, bus, car, or on foot

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Limited Dependent Variables

Restrictions placed on the dependent variable

- Censoring (also known as a Tobit model): Observations from a restricted sample of the population
- Truncation: Reporting of the dependent variable above or below a certain value at that value
- Count regression: Positive, integer values (in addition to zero) of the dependent variable

Extension of count regression models

- Hurdle and zero-inflated models given excessive number of zeros in count data

Duration (also known as hazard or survival) models to determine exogenous variables leading to a particular event occurring

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Overview

Panel data regression models

- Observation of the same individual or unit over multiple years
- Fixed effects versus random effects models

Examples

- Household income and consumption patters
- County-level data on crop yields and area allocation (outcome variables) as a function of output and input prices as well as weather data (e.g., growing degree days)
- State-level spending on higher education and education outcomes

Unit of analysis is the household, county, and state in the previous three examples

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Overview

Time series basics

- Trend and seasonality
- Finite-distributed lag models (e.g., adjustment of consumption after an increase income)

Simple forecasting models

- Moving average and exponential smoothing methods

Time series analysis

- Autoregressive and distributed-lag models

Review of Statistical Concepts: Overview

Basic statistical concepts

- Population versus sample, measures of dispersion, sampling variance

Confidence interval (CI)

- CI for a mean

Hypothesis testing

- One-group and two-group samples

Introduction to Regression Analysis

Jerome Dumortier

Empirical Research Approach

Data Analysis and Modeling
Linear Regression
Qualitative Choice and Limited Dependent Variables
Panel Data
Dynamic Models and Time Series

Review of Statistical Concepts

Basic Statistical Concepts
Confidence Intervals
Hypothesis Testing

Population versus Sample

Population versus sample

- The population is characterized by parameters that will always remain unknown.
- Given a sample taken from the population allows us to learn something about the population parameters.
- The sample needs to be drawn at random.

The sample mean is the arithmetic average of the values in a random sample. It is usually denoted

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Measures of Dispersion

Population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample variance

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

Standard deviation is the square root of the variance, i.e., $\sigma = \sqrt{\sigma^2}$ or $s = \sqrt{s^2}$.

Sampling Variance

The sampling variance is expressed as:

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

This is different from the sample variance! The sampling variance represents the variation of a particular statistic, e.g., mean. The larger n, the smaller the variance in the mean of the various drawings.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Overview

Definition

- A 95% confidence interval for a parameter is an interval obtained from a sample that has a 95% probability of producing a interval containing the true value of the parameter.

Computation:

$$\bar{x} \pm t_{df,\alpha} \cdot \frac{s}{\sqrt{n}}$$

Example data of starting salary (in 1,000) after college graduation:

Student	1	2	3	4	5	6	7	8	9	10
Salary	87	43	59	64	59	71	73	49	68	65

Introduction to Regression Analysis

Jerome Dumortier

Empirical Research Approach

Data Analysis and Modeling
Linear Regression
Qualitative Choice and Limited Dependent Variables
Panel Data
Dynamic Models and Time Series

Review of Statistical Concepts

Basic Statistical Concepts
Confidence Intervals
Hypothesis Testing

Calculation of a Confidence Interval

Given the data salary, we have $\bar{x} = 63.8$, $s = 12.44$, and $t_{9,0.025} = 2.262$. Using the equation:

$$\bar{x} \pm t_{df,\alpha} \cdot \frac{s}{\sqrt{n}}$$

and plugging in the data:

$$63.8 \pm 2.262 \cdot \frac{12.44}{\sqrt{10}} = 63.8 \pm 8.90$$

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts

Confidence Intervals

Hypothesis Testing

Confidence Interval with R

```
salary = c(87,43,59,64,59,71,73,49,68,65)
t.test(salary)
```

```
##
##  One Sample t-test
##
## data:  salary
## t = 16.225, df = 9, p-value = 5.695e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  54.90474 72.69526
## sample estimates:
## mean of x
##      63.8
```

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Introduction

A hypothesis is a statement about a parameter taking on a specific value. A hypothesis test is a procedure to verify the statement and the steps are:

1. Formulating the null hypothesis H_0 stating that the parameter takes a specific value:
   - One-sided test: H_0: $\mu \geq \mu_0$ or $\mu \leq \mu_0$
   - Two-sided test: H_0: $\mu = \mu_0$
2. Setting the significance level α, e.g., 1%, 5%, or 10%.
3. Test statistic: Value based on the sample used to **reject** or **fail to reject** the null hypothesis.
4. Critical value and p-value:
   - Critical value represents the threshold between rejecting and failing to reject H_0.
   - p-Value: Probability of observing the parameter given the null hypothesis. Small p-values represent evidence against H_0.

Note that equality is always part of H_0, i.e., $=$, $\leq$, or $\geq$.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Decisions and Errors in Hypothesis Testing

Null Hypothesis	Fail to reject H_0	Reject H_0
H_0 is true	Correct	Type I Error
H_0 is false	Type II Error	Correct

Type I Error:

- Probability of rejecting H_0 when it is true.
- Also known as the significance level of a test denoted with α.

Type II Error:

- Probability of failing to reject H_0 when it is false.

Interpretation of the p-Value

Each statistical software provides a p-value:

- Lowest level of significance at which the null hypothesis can be rejected.
- Represents the probability of observing the sample given that the hypothesis is true. The lower the p-value the more unlikely is the hypothesis.
- The null hypothesis H_0 is rejected if the p-value is smaller than the significance level.

The smaller the p-value, the stronger the evidence against H_0 being true. This is true for any type of hypothesis test.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Two-sided and One-sided Hypothesis Tests

Two-sided test

- H_0: $\mu = \mu_0$ and H_a: $\mu \neq \mu_0$
- Reject H_0 if $|t| > t_{\alpha/2, n-1}$

One-sided test (left-sided)

- H_0: $\mu \leq \mu_0$ and H_a: $\mu > \mu_0$
- Reject H_0 if $|t| > t_{\alpha, n-1}$

One-sided test (right-sided)

- H_0: $\mu \geq \mu_0$ and H_a: $\mu < \mu_0$
- Reject H_0 if $|t| > t_{\alpha, n-1}$

In both cases, $|t|$ refers to the absolute value of the test statistic. Two-sided tests are of importance in regression analysis.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series
Review of
Statistical
Concepts
Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Hypothesis Testing: One-sample vs. Two-sample

One-sample (or one-group) tests

- Population proportion
- Population mean with unknown variance

Two-sample (or two-group) tests

- Population proportions
- Population means (differentiation between equal and unequal variance)
- Paired difference test

Note: Textbooks often include "population mean with *known* variance." This is a highly unlikely case and thus, it is skipped.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

One-Group Proportion

Test statistic:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)/n}}$$

where p_0 is the hypothesized population proportion.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

One-Group Proportion in R

```
##
##  One Sample t-test
##
## data:  gss$owngun
## t = 24.073, df = 45862, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0.3333333
## 95 percent confidence interval:
##  0.3836523 0.3925726
## sample estimates:
## mean of x
## 0.3881124
```

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

One-Group Mean

Unknown variance requires the use of the t-distribution given the following test statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, μ is the hypothesized mean, s is the sample standard deviation, and n is the sample size.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

One-Group Mean in R

```
##
##  One Sample t-test
##
## data:  eggweights$weight
## t = 1.8378, df = 60, p-value = 0.07104
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##  59.90723 62.19113
## sample estimates:
## mean of x
##  61.04918
```

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Two-Group Proportions

Hypothesis test for difference between two population proportions

$$H_0 : p_1 - p_2 = 0$$

```
t.test(gss$owngun~gss$female)
```

```
##
##  Welch Two Sample t-test
##
## data:  gss$owngun by gss$female
## t = 28.58, df = 42297, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is n
## 95 percent confidence interval:
##  0.1217988 0.1397350
## sample estimates:
## mean in group 0 mean in group 1
##       0.4611440       0.3303771
```

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series
Review of
Statistical
Concepts
Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Two-Group Means

Difference between two mean:

$$H_0 : \bar{x}_1 - \bar{x}_2 = 0$$

Means of two dependent populations

- Assumption of equal variance, i.e., $\sigma_1^2 = \sigma_2^2$
- Example: Pre- and post-test
- Pooled-Variance t-test: One estimate of unknown σ^2, i.e., s_p.

Means of two independent populations

- Assumption of unequal variance, i.e., $\sigma_1^2 \neq \sigma_2^2$
- Samples from two different populations
- Separate-Variance t-test: Two estimates for unknown σ_1^2 and σ_2^2.

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Two-Group Means: Equal Variance

```
t.test(indyhomes$price~indyhomes$zip,var.equal=TRUE)
```

```
##
##   Two Sample t-test
##
## data:  indyhomes$price by indyhomes$zip
## t = 2.0005, df = 100, p-value = 0.04816
## alternative hypothesis: true difference in means between group 46228 and group 46
## 95 percent confidence interval:
##    1510.38 363678.01
## sample estimates:
## mean in group 46228 mean in group 46268
##            381600.2            199006.0
```

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Two-Group Means: Unequal Variance

```
t.test(indyhomes$price~indyhomes$zip,var.equal=FALSE)
```

```
##
##   Welch Two Sample t-test
##
## data:  indyhomes$price by indyhomes$zip
## t = 2.0403, df = 51.323, p-value = 0.04648
## alternative hypothesis: true difference in means between group 46228 and group 46
## 95 percent confidence interval:
##    2953.984 362234.402
## sample estimates:
## mean in group 46228 mean in group 46268
##            381600.2            199006.0
```

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Paired Difference Test I

Difference between paired (!) values:

$$D_i = x_{1,i} - x_{2,i}$$

Elimination of variation among subjects. Point estimate for paired difference

$$\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$$

Sample standard deviation

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}}$$

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series
Review of
Statistical
Concepts
Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Paired Difference Test II

Test statistic

$$t_p = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}}$$

Confidence interval

$$\bar{D} \pm t_{\alpha/2}\frac{S_D}{\sqrt{n}}$$

t_p has n-1 degrees of freedom

Introduction
to Regression
Analysis

Jerome
Dumortier

Empirical
Research
Approach

Data Analysis
and Modeling
Linear Regression
Qualitative Choice
and Limited
Dependent Variables
Panel Data
Dynamic Models and
Time Series

Review of
Statistical
Concepts

Basic
Statistical
Concepts
Confidence Intervals
Hypothesis Testing

Textbook Example

Book	Online	Bookstore	Difference
History 1	10.20	11.40	-1.20
History 2	18.95	19.00	-0.05
Economics 1	184.53	200.75	-16.22
Business 1	236.75	247.20	-10.45
Business 2	67.41	71.25	-3.48

Note that $\sum D_i = -31.76$, $\bar{D} = -6.352$, and $s_D = 6.833$.

Introduction to Regression Analysis

Jerome Dumortier

Empirical Research Approach

Data Analysis and Modeling
Linear Regression
Qualitative Choice and Limited Dependent Variables
Panel Data
Dynamic Models and Time Series

Review of Statistical Concepts

Basic Statistical Concepts
Confidence Intervals
Hypothesis Testing

Textbook Example in R

```
online     = c(10.20,18.95,184.53,236.75,67.41)
bookstore  = c(11.40,19.00,200.75,247.20,71.25)
t.test(online,bookstore,paired=TRUE)

##
##   Paired t-test
##
## data:  online and bookstore
## t = -2.0788, df = 4, p-value = 0.1062
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -14.835834    2.131834
## sample estimates:
## mean difference
##          -6.352
```