R Applications

Tan Sein Jone

University of British Columbia

August 7, 2024

Table of contents

Table of Contents

Hypothesis Testing

- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data.
- A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- These two statements are called the null hypothesis and the alternative hypothesis.

Hypothesis Testing

- The null hypothesis is the statement being tested. Usually, the null hypothesis states that there is no effect or no difference.
- The alternative hypothesis is the statement that is accepted if the sample data provide enough evidence that the null hypothesis is false.
- The alternative hypothesis states that there is an effect or a difference.

Hypothesis Testing

- The hypothesis test is conducted by comparing the value of the test statistic to a critical value.
- The critical value is a value that determines whether the null hypothesis can be rejected.
- If the test statistic is more extreme than the critical value, then the null hypothesis is rejected.

Types of Hypothesis Tests

- One-Sample t-Test: Used to test whether the mean of a single sample is significantly different from a specified value.
- Two-Sample t-Test: Used to test whether the means of two independent samples are significantly different from each other.
- Paired t-Test: Used to test whether the means of two paired samples are significantly different from each other.

Types of Hypothesis Tests

- One-Sample z-Test: Used to test whether the mean of a single sample is significantly different from a specified value when the sample size is large.
- Two-Sample z-Test: Used to test whether the means of two independent samples are significantly different from each other when the sample sizes are large.
- Chi-Square Test: Used to test whether the observed frequencies in a contingency table are significantly different from the expected frequencies.

Types of Hypothesis Tests

- One-Way ANOVA: Used to test whether the means of three or more independent samples are significantly different from each other.
- Two-Way ANOVA: Used to test whether the means of two or more independent samples are significantly different from each other, taking into account two independent variables.
- Goodness of Fit Test: Used to test whether the observed frequencies in a sample are consistent with the expected frequencies.

Types of Hypothesis Tests

- Test for Equal Variances: Used to test whether the variances of two or more samples are equal.
- Test for Normality: Used to test whether the data in a sample comes from a normal distribution.
- Test for Independence: Used to test whether the observations in a sample are independent of each other.

Steps in Hypothesis Testing

- Step 1: State the null hypothesis and the alternative hypothesis.
- Step 2: Choose the significance level.
- Step 3: Collect the sample data and calculate the test statistic.
- Step 4: Determine the critical value or the p-value.
- Step 5: Make a decision to reject or fail to reject the null hypothesis.
- Step 6: Interpret the results of the hypothesis test.

Null and Alternative Hypotheses

- The null hypothesis is denoted by H_0 and states that there is no effect or no difference.
- The alternative hypothesis is denoted by H_1 and states that there is an effect or a difference.
- The null hypothesis is the default assumption that is tested against the alternative hypothesis.
- The null hypothesis is rejected if the sample data provide enough evidence that the null hypothesis is false.

Assumptions of Hypothesis Testing

- The sample data are independent and identically distributed.
- The sample data are drawn from a population that is normally distributed.
- The sample data are drawn from a population that has a constant variance.
- The sample data are drawn from a population that is randomly selected.
- The sample data are drawn from a population that is representative of the population of interest.

Significance Level

- The significance level is the probability of rejecting the null hypothesis when it is true.
- The significance level is denoted by α and is usually set to 0.05.
- The significance level is the threshold for determining whether the null hypothesis should be rejected.
- If the p-value is less than the significance level, then the null hypothesis is rejected.

Test Statistic

- The test statistic is a numerical value that is used to determine whether the null hypothesis should be rejected.
- The test statistic is calculated from the sample data and is compared to a critical value or a p-value.
- The test statistic measures the strength of the evidence against the null hypothesis.
- The test statistic is used to make a decision to reject or fail to reject the null hypothesis.

Critical Value

- The critical value is a value that determines whether the null hypothesis can be rejected.
- The critical value is determined by the significance level and the degrees of freedom.
- The critical value is compared to the test statistic to determine whether the null hypothesis should be rejected.
- If the test statistic is more extreme than the critical value, then the null hypothesis is rejected.

P-Value

- The p-value is the probability of observing a test statistic as extreme as the one computed from the sample data, assuming that the null hypothesis is true.
- The p-value is a measure of the strength of the evidence against the null hypothesis.
- The p-value is compared to the significance level to determine whether the null hypothesis should be rejected.
- If the p-value is less than the significance level, then the null hypothesis is rejected.

Type I and Type II Errors

- Type I Error: Occurs when the null hypothesis is rejected when it is true.
- Type II Error: Occurs when the null hypothesis is not rejected when it is false.
- The probability of a Type I Error is denoted by α and is equal to the significance level.
- The probability of a Type II Error is denoted by β and is equal to 1 minus the power of the test.

Power of the Test

- The power of the test is the probability of rejecting the null hypothesis when it is false.
- The power of the test is equal to 1 minus the probability of a Type II Error.
- The power of the test is a measure of the ability of the test to detect an effect or a difference when it exists.
- The power of the test is affected by the sample size, the effect size, and the significance level.

Confidence Interval

- A confidence interval is a range of values that is likely to contain the true value of the population parameter.
- A confidence interval is calculated from the sample data and is used to estimate the population parameter.
- The confidence interval is calculated from the sample mean and the standard error of the mean.
- The confidence interval is used to make inferences about the population parameter with a specified level of confidence.

Confidence Level

- The confidence level is the probability that the confidence interval contains the true value of the population parameter.
- The confidence level is denoted by $(1 - \alpha) \times 100\%$ and is usually set to 95%.
- The confidence level is the proportion of confidence intervals that contain the true value of the population parameter.
- The confidence level is used to make inferences about the population parameter with a specified level of confidence.

Hypothesis Testing in R

- In R, the `t.test()` function is used to perform hypothesis tests.
- The `t.test()` function takes in the sample data and the null hypothesis as arguments.
- The function returns the test statistic, the p-value, and the confidence interval.

Hypothesis Testing in R

- The `t.test()` function can be used to perform one-sample t-tests, two-sample t-tests, and paired t-tests.
- The `t.test()` function can also be used to perform one-sample z-tests and two-sample z-tests.
- The `t.test()` function can be used to perform hypothesis tests for means, proportions, and variances.

Hypothesis Testing in R

```r
# Generate some data
set.seed(123)
x <- rnorm(100, mean = 5, sd = 2)
y <- rnorm(100, mean = 6, sd = 2)

# One-sample t-test
t.test(x, mu = 0)

# Two-sample t-test
t.test(x, y)

# Paired t-test
t.test(x, y, paired = TRUE)
```

Breakdown of Hypothesis Testing in R

- **Set Seed:**
  - `set.seed(123)`: Sets the seed for random number generation to ensure reproducibility of results.
- **Generate Data:**
  - `x <- rnorm(100, mean = 5, sd = 2)`: Generates 100 random numbers from a normal distribution with a mean of 5 and a standard deviation of 2, and assigns them to vector `x`.
  - `y <- rnorm(100, mean = 6, sd = 2)`: Generates 100 random numbers from a normal distribution with a mean of 6 and a standard deviation of 2, and assigns them to vector `y`.
- **One-Sample t-Test:**
  - `t.test(x, mu = 0)`: Performs a one-sample t-test to check if the mean of vector `x` is significantly different from 0.
- **Two-Sample t-Test:**
  - `t.test(x, y)`: Performs a two-sample t-test to check if the means of vectors `x` and `y` are significantly different from each other.

Table of Contents

Regression Analysis

- Regression analysis is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables.
- The goal of regression analysis is to estimate the parameters of the regression model that best fit the data.
- The regression model is a mathematical equation that describes the relationship between the dependent variable and the independent variables.

Types of Regression Models

- Linear Regression: Used to model the relationship between a continuous dependent variable and one or more continuous independent variables.
- Logistic Regression: Used to model the relationship between a binary dependent variable and one or more continuous independent variables.
- Polynomial Regression: Used to model the relationship between a continuous dependent variable and one or more continuous independent variables using polynomial functions.

Types of Regression Models

- Ridge Regression: Used to model the relationship between a continuous dependent variable and one or more continuous independent variables with multicollinearity.
- Lasso Regression: Used to model the relationship between a continuous dependent variable and one or more continuous independent variables with feature selection.
- Elastic Net Regression: Used to model the relationship between a continuous dependent variable and one or more continuous independent variables with both ridge and lasso regularization.

Linear Regression

- Linear regression is a regression model that assumes a linear relationship between the dependent variable and the independent variables.
- The linear regression model is a mathematical equation that describes the relationship between the dependent variable and the independent variables.
- The linear regression model is estimated using the method of least squares, which minimizes the sum of the squared differences between the observed values and the predicted values.

Linear Regression

- The linear regression model is specified by the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$, where y is the dependent variable, $x_1, x_2, \ldots, x_n$ are the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and ϵ is the error term.
- The coefficients of the independent variables are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed values and the predicted values.

Logistic Regression

- Logistic regression is a regression model that is used when the dependent variable is binary.
- The logistic regression model is a mathematical equation that describes the relationship between the dependent variable and the independent variables.
- The logistic regression model is estimated using the method of maximum likelihood, which maximizes the likelihood of the observed data given the model.

Ordinary Least Squares

- Ordinary least squares (OLS) is a method that is used to estimate the parameters of a linear regression model.
- OLS minimizes the sum of the squared differences between the observed values and the predicted values.
- OLS is used to estimate the coefficients of the independent variables that best fit the data.
- OLS is used to estimate the coefficients of the independent variables that minimize the sum of the squared differences between the observed values and the predicted values.

Formula for OLS

- The formula for OLS is $\hat{\beta} = (X^T X)^{-1} X^T y$, where $\hat{\beta}$ is the estimated coefficients of the independent variables, X is the matrix of the independent variables, and y is the vector of the dependent variable.
- The formula for OLS is derived by minimizing the sum of the squared differences between the observed values and the predicted values.
- The formula for OLS is used to estimate the coefficients of the independent variables that best fit the data.

Logistic Regression

- The logistic regression model is specified by the equation $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$, where p is the probability of the dependent variable being 1, $x_1, x_2, \ldots, x_n$ are the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and e is the base of the natural logarithm.
- The coefficients of the independent variables are estimated using the method of maximum likelihood, which maximizes the likelihood of the observed data given the model.

Polynomial Regression

- Polynomial regression is a regression model that is used when the relationship between the dependent variable and the independent variables is not linear.
- The polynomial regression model is a mathematical equation that describes the relationship between the dependent variable and the independent variables using polynomial functions.
- The polynomial regression model is estimated using the method of least squares, which minimizes the sum of the squared differences between the observed values and the predicted values.

Polynomial Regression

- The polynomial regression model is specified by the equation $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon$, where y is the dependent variable, x is the independent variable, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and ϵ is the error term.
- The coefficients of the independent variables are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed values and the predicted values.

Regression Analysis

- The goodness of fit of the regression model is measured using the coefficient of determination, which is the proportion of the variance in the dependent variable that is explained by the independent variables.

- The coefficient of determination ranges from 0 to 1, with higher values indicating a better fit.

- The significance of the regression model is tested using the F-test, which tests whether the regression model is a better fit than a model with no independent variables.

Regression Analysis in R

- In R, the `lm()` function is used to fit linear regression models.
- The `lm()` function takes in the formula for the regression model and the data as arguments.
- The formula specifies the dependent variable and the independent variables in the regression model.

Regression Analysis in R

- The `summary()` function is used to display the results of the regression analysis.
- The `summary()` function displays the estimated coefficients, the standard errors, the t-values, and the p-values of the regression model.
- The `summary()` function also displays the coefficient of determination and the results of the F-test.

Regression Analysis in R

```r
# Generate some data
set.seed(123)
data <- data.frame(
    y = rnorm(100, mean = 5, sd = 2),
    x1 = rnorm(100, mean = 6, sd = 2),
    x2 = rnorm(100, mean = 7, sd = 2)
)

# Fit linear regression model
model <- lm(y ~ x1 + x2, data = data)

# Display results
summary(model)
```

Breakdown of Regression Analysis in R

- **Set Seed:**
  - `set.seed(123)`: Sets the seed for random number generation to ensure reproducibility of results.

- **Generate Data:**
  - `data <- data.frame(y = rnorm(100, mean = 5, sd = 2), x1 = rnorm(100, mean = 6, sd = 2), x2 = rnorm(100, mean = 7, sd = 2)`: Generates 100 random numbers from a normal distribution with specified means and standard deviations, and assigns them to variables `y`, `x1`, and `x2`.

- **Fit Linear Regression Model:**
  - `model <- lm(y   x1 + x2, data = data)`: Fits a linear regression model to predict variable `y` using variables `x1` and `x2` as predictors.

- **Display Results:**
  - `summary(model)`: Displays the results of the linear regression analysis, including the estimated coefficients, standard errors, t-values, p-values, coefficient of determination, and F-test results.

Best Practices for Displaying Results

- Use tables to display the results of the regression analysis.
- Include the estimated coefficients, standard errors, t-values, and p-values in the table.
- Highlight the statistically significant coefficients in the table.
- Include the coefficient of determination and the results of the F-test in the table.
- Use visualizations, such as scatter plots and residual plots, to display the relationship between the dependent variable and the independent variables.

Best Practices for Interpreting Results

- Interpret the estimated coefficients of the independent variables in the regression model.
- Interpret the coefficient of determination, which is the proportion of the variance in the dependent variable that is explained by the independent variables.
- Interpret the results of the F-test, which tests whether the regression model is a better fit than a model with no independent variables.
- Interpret the p-values of the estimated coefficients, which indicate the statistical significance of the coefficients.

Table of Contents

Instrumental Variables

- Instrumental variables are used in econometrics to estimate the causal effect of an independent variable on a dependent variable.
- Instrumental variables are used when the independent variable is correlated with the error term in the regression model.
- Instrumental variables are used to identify the causal effect of the independent variable by removing the correlation between the independent variable and the error term.

Instrumental Variables

- Instrumental variables are variables that are correlated with the independent variable but are uncorrelated with the error term.
- Instrumental variables are used to estimate the causal effect of the independent variable by using the variation in the instrumental variables to identify the causal effect.
- Instrumental variables are used in regression analysis to estimate the parameters of the regression model that best fit the data.

Assumptions of Instrumental Variables

- Relevance: The instrumental variables are correlated with the independent variable.
- Exogeneity: The instrumental variables are uncorrelated with the error term.
- Exclusion: The instrumental variables are not correlated with the dependent variable.

Checking Assumptions

- The relevance assumption can be checked by testing the correlation between the instrumental variables and the independent variable.
- Exogeneity however has to be argued, since it is not testable.
- Exclusion can be checked by testing the correlation between the instrumental variables and the dependent variable.

Local Average Treatment Effect

- The local average treatment effect (LATE) is the causal effect of the independent variable on the dependent variable for the subpopulation of individuals who are affected by the instrumental variables.

Wald Estimator

$$\hat{\beta}_{\mathrm{IV}} = \frac{\mathrm{Cov}(z, y)}{\mathrm{Cov}(z, x)} \tag{1}$$

- The Wald estimator is a method that is used to estimate the parameters of the instrumental variables regression model.
- The Wald estimator is calculated as the ratio of the covariance between the instrumental variables and the dependent variable to the covariance between the instrumental variables and the independent variable.
- The Wald estimator is used to estimate the causal effect of the independent variable on the dependent variable by removing the correlation between the independent variable and the error term.

Instrumental Variables

- The instrumental variables regression model is specified by the equation $y = \beta_0 + \beta_1 x + \epsilon$, where y is the dependent variable, x is the independent variable, β_0 and β_1 are the coefficients of the independent variable, and ϵ is the error term.
- The instrumental variables regression model is estimated using the method of instrumental variables, which removes the correlation between the independent variable and the error term.
- The instrumental variables regression model is used to estimate the causal effect of the independent variable on the dependent variable.

2 Stage Least Squares

- The two-stage least squares (2SLS) method is a method that is used to estimate the parameters of the instrumental variables regression model.
- The 2SLS method is a two-stage procedure that first estimates the parameters of the instrumental variables regression model and then uses the estimated parameters to estimate the causal effect of the independent variable on the dependent variable.
- The 2SLS method is used to estimate the causal effect of the independent variable by removing the correlation between the independent variable and the error term.

2 Stage Least Squares

- The 2SLS method is specified by the following steps:
  - Step 1: Estimate the parameters of the instrumental variables regression model using the instrumental variables.
  - Step 2: Use the estimated parameters to estimate the causal effect of the independent variable on the dependent variable.
- The 2SLS method is used to estimate the causal effect of the independent variable by removing the correlation between the independent variable and the error term.

Examples of IV

- The effect of education on income can be estimated using the number of years of schooling as an instrumental variable.
- Education is correlated with income, but it is also correlated with other factors that affect income, such as ability and motivation.
- The number of years of schooling is correlated with education, but it is not correlated with ability and motivation, making it a good instrumental variable.
- The effect of education on income can be estimated using the number of years of schooling as an instrumental variable to remove the correlation between education and ability and motivation.

Instrumental Variables in R

- The `ivreg()` function estimates the parameters of the regression model using the method of instrumental variables.
- The `ivreg()` function returns the estimated coefficients, the standard errors, the t-values, and the p-values of the regression model.
- The `ivreg()` function is used to estimate the causal effect of the independent variable on the dependent variable by removing the correlation between the independent variable and the error term.

Instrumental Variables in R

```
library(AER)
# Generate some data
set.seed(123)
n <- 100
x <- rnorm(n)
z <- rnorm(n)
y <- 1 + 2 * x + 3 * z + rnorm(n)

# Fit instrumental variables regression model
model <- ivreg(y ~ x | z)
summary(model)
```

Breakdown of Instrumental Variables in R

- **Load Package:**
  - `library(AER)`: Loads the AER package, which contains the `ivreg()` function for estimating instrumental variables regression models.

- **Generate Data:**
  - `x <- rnorm(n)`: Generates `n` random numbers from a normal distribution and assigns them to vector `x`.
  - `z <- rnorm(n)`: Generates `n` random numbers from a normal distribution and assigns them to vector `z`.
  - `y <- 1 + 2 * x + 3 * z + rnorm(n)`: Generates `n` random numbers from a normal distribution and assigns them to vector `y`.

- **Fit Instrumental Variables Regression Model:**
  - `model <- ivreg(y   x | z)`: Fits an instrumental variables regression model to predict variable `y` using variable `x` as a predictor and variable `z` as an instrumental variable.

- **Display Results:**
  - `summary(model)`: Displays the results of the instrumental variables regression analysis, including the estimated coefficients, standard errors, t-values, and p-values