SPEC REU R Resources: Applied Introduction to T-Tests, Correlation, and OLS regression

Alix Ziff, Gaea Morales, Zachary Johnson, Claudia Salas Gimenez and Ben Graham

January 2025

Welcome to the first module on regression analysis! This walkthrough provides a very brief introduction to fundamental statistical techniques commonly used in data analysis and research, focusing on correlation, Ordinary Least Squares (OLS) regression, and summary statistics. By the end of this walkthrough, you will understand how to calculate and interpret correlations, run preliminary OLS regressions, and use statistical tools to uncover relationships between variables, assess their strength, and draw data-driven conclusions.

This walkthrough serves as your starting guide to regression analysis in R, building a foundation for more advanced techniques in later modules. If you're new to statistics, this is a math-free introduction that complements more formal econometrics courses. This course is NOT a replacement for a thorough econometrics course.

Initial Setup

Begin by setting up your working directory and loading the necessary libraries.

```
# Set working directory
#setwd("YourFolderPath")

# Load required libraries
library(tidyverse)
library(ggplot2)
```

Data Preparation

For simplicity, for this exercise we'll create practice data with three numeric variables `X1`, `X2` and `X3`. These variables are generated randomly and don't represent real-world values; they're simply generated from normal distributions for illustration.

For ease of understanding, we are going to create some practice data that has no root in real-life observation. We will use the `rnorm()` function to generate normally distributed data. If you're curious, watch this video for details on what a normal distribution is.

The following code generates three normally distributed numeric variables `X1`, `X2` and `X3`. Don't focus too much on the details of the code – it generates three columns of randomly generated numeric values.

```
# Create dummy data
data <- data.frame(
  x1 = 1:100 + rnorm(100, sd=9),
  x2 = 1:100 + rnorm(100, sd=16),
  x3 = 1:100 + rnorm(100, sd=3))
```

```r
# Preview the first few rows of the dataset
head(data)
```

```
##           x1         x2         x3
## 1 -9.1464700  -2.486780 -0.8753317
## 2 -3.0552639  -3.931987  1.2254946
## 3 -5.6379438  23.250800 -0.9100024
## 4  6.6069695 -11.370169 -0.4933276
## 5  6.6462032  -7.730172  4.4460469
## 6 -0.1913463  12.206280  0.5396395
```

Correlation Analysis

In this section, we will calculate a correlation matrix for our three numeric variables. The correlation coefficients in this matrix tells us the strength and direction of the relationship between each pair of variables. The value of a correlation coefficient ranges from -1 to 1, where:

- +1 indicates a perfect positive correlation (as one variable increases, the other variable increases proportionally).
- -1 indicates a perfect negative correlation (as one variable increases, the other variable decreases proportionally).
- 0 indicates no correlation (no linear relationship between the variables).

We will use the `cor()` function to compute the correlation coefficient for `X1`, `X2` and `X3`.

```r
# Calculate correlation matrix
cor_matrix <- cor(data, use = "complete.obs")
## 'use = "complete.obs"' argument specifies to only use rows that have no missing values

print(cor_matrix)
```

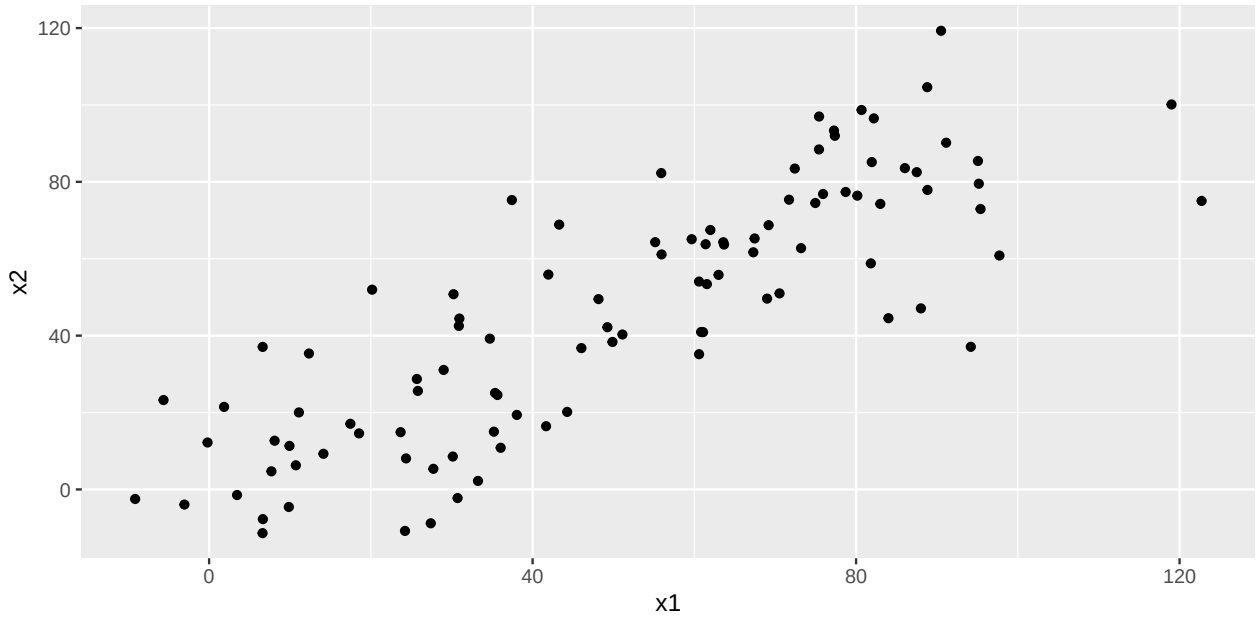
```
##           x1        x2        x3
## x1 1.0000000 0.8289401 0.9477477
## x2 0.8289401 1.0000000 0.8705728
## x3 0.9477477 0.8705728 1.0000000
```

Visualizing Correlations

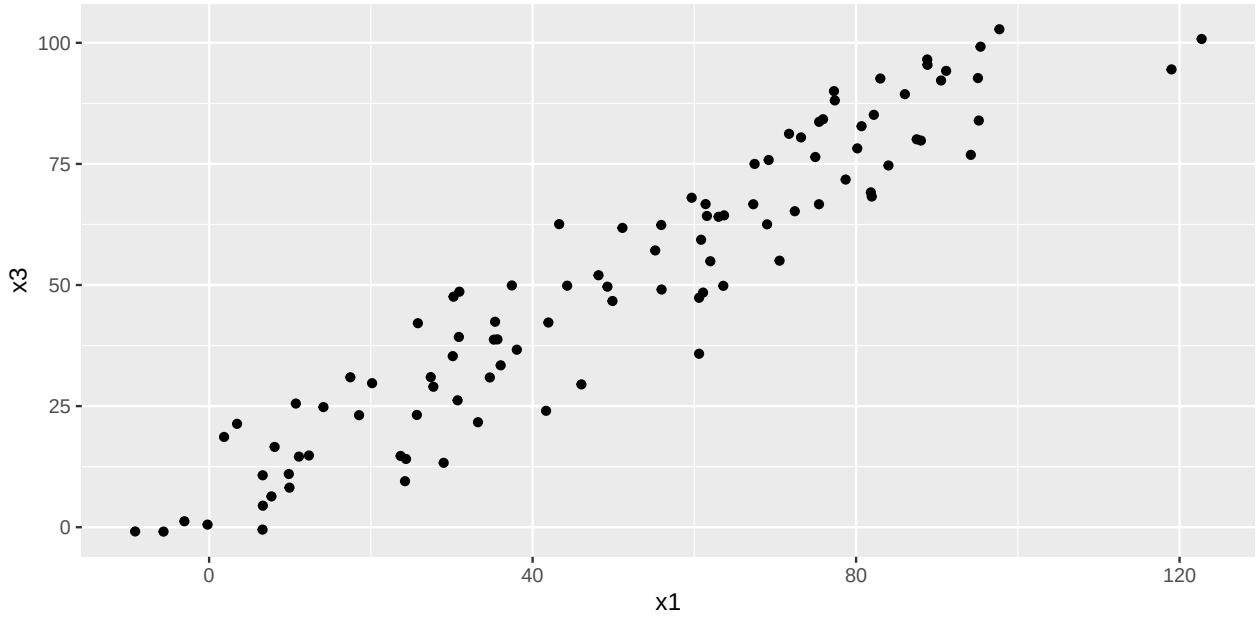
Now, let's say that we want to understand the strength and direction of our correlation. We can visualize the correlation by creating a scatterplot. As seen in M6: Data Visualization II Walkthrough pt.1, scatterplots allow us to analyze the relationship between the two variables, observing any patterns or trends.

- If the correlation coefficient is close to 1, the scatterplot will show points that are tightly clustered around a line with a positive slope. This indicates that as one variable increases, the other one also increases proportionally.
- If the correlation coefficient is close to -1, the scatterplot will show points that are tightly clustered around a line with a negative slope. This indicates that as one variable increases, the other one decreases proportionally.
- If the correlation coefficient is close to 0, the scatterplot will show points that are more randomly dispersed without any clear pattern. This indicates that there is no linear relationship between the variables.

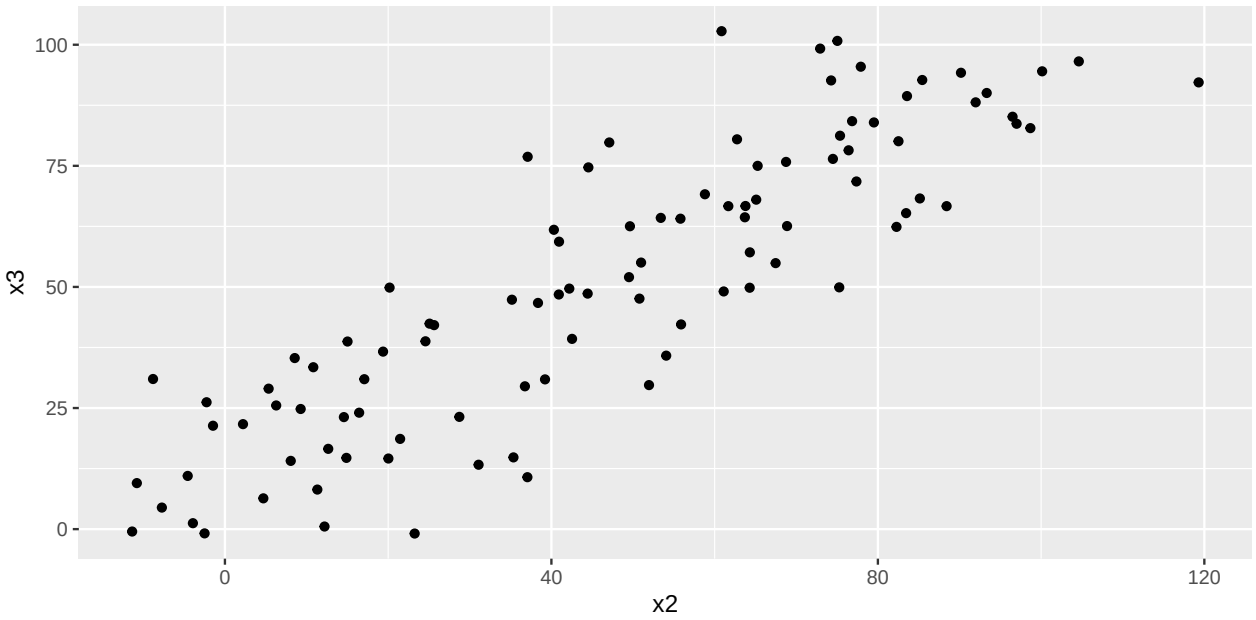
```r
# Create scatterplot to visualize correlation between X1 and X2
ggplot(data = data, aes(x = x1, y = x2)) +
  geom_jitter()
```




```r
# Create scatterplot to visualize correlation between X1 and X3
ggplot(data = data, aes(x = x1, y = x3)) +
  geom_jitter()
```




```r
# Create scatterplot to visualize correlation between X3 and X2
ggplot(data = data, aes(x = x2, y = x3)) +
  geom_jitter()
```



Notice that, like in our correlation matrix, the three scatterplots show clear upward slopes, suggesting a positive correlation between the variables. `X1` and `X3` show a tighter grouping of observations with a clear upward slope, which indicates a high positive correlation between them, as they have the highest correlation (0.9419428).

Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. In simple terms, OLS regression helps us understand how changes in the independent variable(s) are associated with changes in the dependent variable.

The general formula for an Ordinary Least Squares (OLS) regression model is:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i$$

Where:

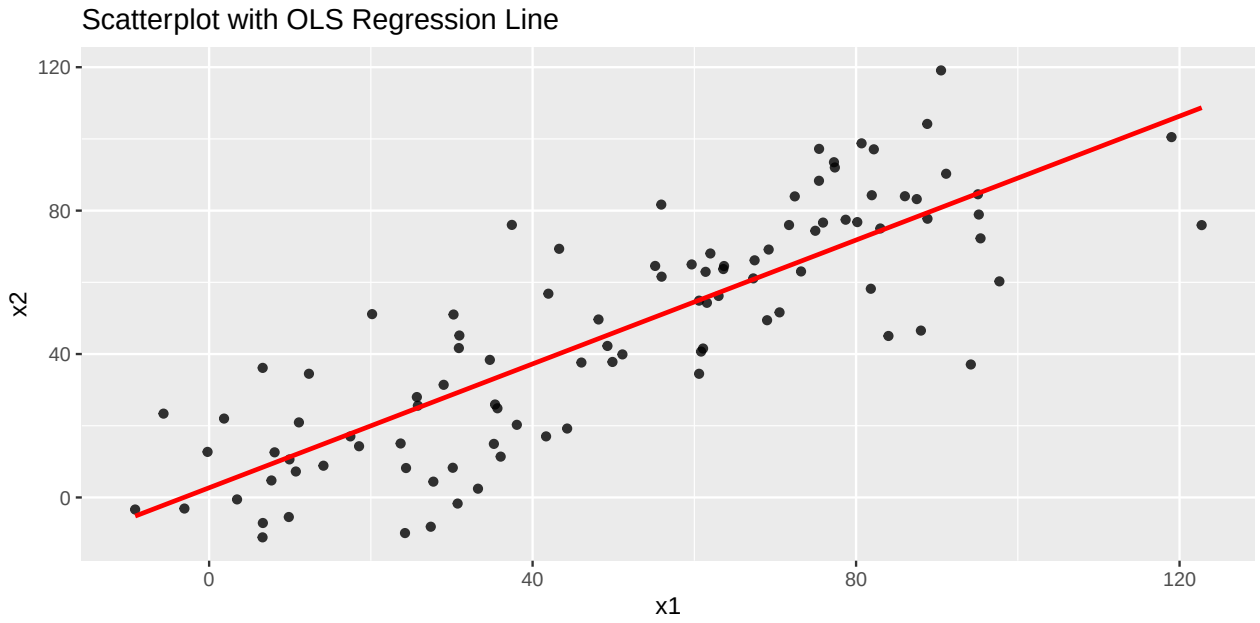
- y_i is the dependent variable for observation i.
- β 0 is the intercept term.
- β 1, β 2, β k are the coefficients for each independent variable `Xi1`; `Xi2`; `Xik`.
- ϵ_i is the error term (the difference between the observed and predicted values) for observation i.

Plotting a Regression

While visualizing the data with a scatterplot allows us to visually inspect the relationship between the variables, adding a regression line provides a visual representation of the linear relationship between the two variables in the plot. In general, as one variable increases in value, does the other increase or decrease? By how much?

```
# Scatterplot with regression line
ggplot(data, aes(x = x1, y = x2)) +
  geom_point(position = position_jitter(height = 1), alpha = 0.8) +
```

```
  geom_smooth(method = "lm", colour = "red", se = FALSE) +
  # Draw line of best-fit with "lm"
  labs(title = "Scatterplot with OLS Regression Line",
       x = "x1",
       y = "x2")
```



Running a Regression

Conducting an OLS regression analysis provides us with quantitative measures to assess the relationship between the variables. The regression coefficients quantify the expected change in the dependent variable for a one-unit change in the independent variable(s). This helps us understand the magnitude and direction of the effect.

Let's proceed with performing an OLS regression on our practice data.

```
# Fit OLS regression model
fit <- lm(x2 ~ x1, data = data, na.action = na.exclude)

# Summarize the regression results
summary(fit)
```

```
##
## Call:
## lm(formula = x2 ~ x1, data = data, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.967 -12.642   0.326  10.755  40.239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.70306    3.48373   0.776     0.44
## x1           0.86369    0.05887  14.671   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 18.02 on 98 degrees of freedom
## Multiple R-squared:  0.6871, Adjusted R-squared:  0.6839
## F-statistic: 215.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

Interpreting Regression Results

With the regression, we want to understand if there is a relationship between `X1` and `X2`. The summary of the OLS regression model provides several important pieces of information. Note that here we are only highlighting a couple of key elements of our regression results to determine if there is relationship between our variables.

- **Coefficients OR Estimates**: The regression coefficients tell us the expected change in the dependent variable (`X2`) for a one-unit change in the independent variable (`X1`).
- **Intercept**: The intercept represents the expected value of `X2` when `X1` is zero.
- **Residuals**: These represent the differences between the observed values and the values predicted by the model. Smaller residuals indicate a better fit. In the scatterplot, these are the distances between each observed value and the predicted value from the line of best fit.
- **Standard Error**: This measures the precision of the coefficient estimates. Smaller standard errors indicate more precise estimates.
- **T-Value**: This statistic helps determine whether the coefficient is significantly different from zero. Higher absolute t-values indicate more significant coefficients.
- **R-squared (R^2)**: It represents the proportion of the variance in the dependent variable that can be explained by the independent variable(s). It ranges between 0 and 1, and an R^2 closer to 1 indicates a better fit.
- **P-Value**: This indicates the strength of the evidence against the null hypothesis. In simpler terms, it is essentially the likelihood that our results occurred by random chance. A smaller p-value (< 0.05) suggests that the relationship between the variables is statistically significant.

Regression Analysis Context

Why do we expect certain variables to correlate with each other? Before running any analysis, we typically look at real-life variables and hypothesize a causal relationship based on our knowledge of how the world works. An example of this would be our intuition that as populations experience economic development (accrue wealth), women may have fewer children (the fertility rate decreases) due to more access to education and employment opportunities, as well as a shift in cultural expectations. This hypothesis would be the basis for a regression analysis.

Hypotheses

Thus, a hypothesis can be defined as an informed guess about how certain variables relate to each other based on existing knowledge or assumptions about how the world works. Often, we begin by stating a null hypothesis. The **null hypothesis (H0)** proposes there is no relationship between the variables (correlation is zero).

H0: $\beta = 0$

The **alternate hypothesis (H1)** proposes there is a significant relationship between the variables (correlation is different than zero).

H1: $\beta \neq 0$

Using the example above, under the null hypothesis, wealth and fertility are unrelated (the slope in a regression would be zero). Under the alternative hypothesis, we predict a negative slope, meaning fertility rates decrease as wealth increases. By fitting a linear model and testing whether the slope is significantly different from zero, we can evaluate which hypothesis is more likely supported by the data.

However, in many cases hypotheses are directional—we don't just predict that two variables are related; we also specify whether the relationship is positive or negative. For example, if we consider economic development and fertility rates, we might hypothesize that as populations become wealthier, women tend to have fewer children over time. This prediction is based on existing knowledge, such as expanded access to education and employment opportunities for women and shifts in cultural expectations associated with economic development.

When formulating such directional hypotheses, the null and alternative hypotheses are framed accordingly:

H0: $\beta \leq$ or ≥ 0

H1: $\beta >$ or < 0

For our example, the null hypothesis suggests that economic development increases fertility rates ($\beta \geq 0$), while the alternative hypothesis posits that fertility rates decrease as economic development increases ($\beta < 0$). By testing these hypotheses, we can assess whether the data supports the predicted direction—a negative relationship between economic development and fertility rates.

Conclusion

In this walkthrough, we explored correlations, fitted OLS regressions, and interpreted regression results while practicing data preparation and visualization. These skills are foundational for analyzing relationships between variables and drawing meaningful conclusions.

Next, you will apply these techniques to groupwork and homework assignments, working with real-world data to solidify your understanding and prepare for more advanced regression methods.