Regression Analysis

Inchan Hwang, inchan@yonsei.ac.kr

2 Jan 2025

1. Linear Regression

What is Regression?

Regression is a statistical method used to model and analyze the relationships between a dependent (outcome) variable and one or more independent (explanatory) variables. Let's start with a simple explanation and delve into the details later. We will create a dataset for 10 students that includes their High School GPA and First-Year GPA. Using this data, we aim to explore the relationship between the two variables.

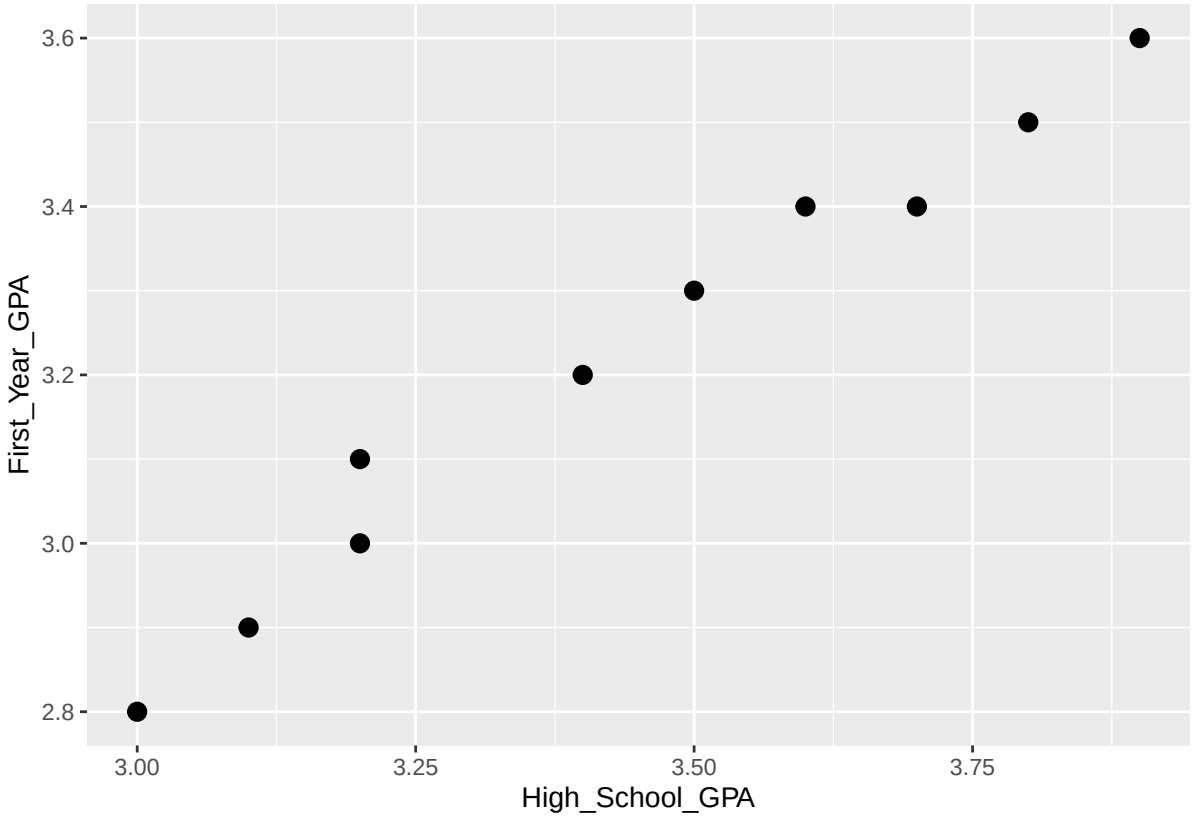
```
regression_data <- data.frame(
  Student = paste("Student", 1:10),
  High_School_GPA = c(3.0, 3.2, 3.5, 3.8, 3.6, 3.1, 3.9, 3.4, 3.7, 3.2), # High school GPA
  First_Year_GPA = c(2.8, 3.0, 3.3, 3.5, 3.4, 2.9, 3.6, 3.2, 3.4, 3.1)  # First-year college
)

regression_data
```

```
##       Student High_School_GPA First_Year_GPA
## 1   Student 1             3.0            2.8
## 2   Student 2             3.2            3.0
## 3   Student 3             3.5            3.3
## 4   Student 4             3.8            3.5
## 5   Student 5             3.6            3.4
## 6   Student 6             3.1            2.9
## 7   Student 7             3.9            3.6
## 8   Student 8             3.4            3.2
## 9   Student 9             3.7            3.4
## 10 Student 10             3.2            3.1
```

Now, let's visually represent the relationship between High School GPA and First-Year GPA using a scatterplot.

```
library(ggplot2)
ggplot(regression_data, aes(x = High_School_GPA, y = First_Year_GPA)) +
geom_point(size=3)
```



The purpose of regression analysis is to represent the relationship between variables as a single line.

Detail Mathematical Explanation

The most common form of regression is linear regression, where the relationship is modeled as a straight line:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where: - Y: Dependent variable (outcome variable)

- X: Independent variable (explanatory variable)

- β_0: Intercept (value of Y when $X = 0$)

- β_1: Slope (change in Y for a one-unit change in X)

- ϵ: Error term (captures variability not explained by X)

This line is referred to as the estimated regression line. The estimated regression line's slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) are calculated using the following formulas:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{V(X)} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where: - Regression slope (β_1): The rate at which Y changes with X.

- Variance (Var(X)): Measures how much X values deviate from their mean.

- Covariance (Cov(X, Y)): Measures how X and Y vary together.

- x_i: Independent variable value for the i-th observation

- y_i: Dependent variable value for the i-th observation

- $\bar{x}$: Mean of the independent variable

- $\bar{y}$: Mean of the dependent variable

- m: Total number of observations

For the regression line to effectively capture the relationship, it must minimize the distance between the data points and the line. This distance is called the residual. The estimated regression line's intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ are calculated such that the sum of squared residuals is minimized. The equation for this calculation is as follows:

$$\min \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2$$

Where: - Y_i: Actual value of the dependent variable

- X_i: Actual value of the independent variable

- $\hat{\beta}_0$: Estimated intercept of the regression line

- $\hat{\beta}_1$: Estimated slope of the regression line

- n: Number of observations

Simple Regression Analysis

We will now apply linear regression to explore the relationship between high school GPA and first-year college GPA:

```
# Performing simple linear regression
simple_regression <- lm(First_Year_GPA ~ High_School_GPA, data = regression_data)
summary(simple_regression) # Display regression results
```

```
##
## Call:
## lm(formula = First_Year_GPA ~ High_School_GPA, data = regression_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.04722 -0.03021 -0.01319  0.02535  0.08333
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.30556    0.16524   1.849    0.102
## High_School_GPA  0.84722    0.04786  17.702 1.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04449 on 8 degrees of freedom
```

```
## Multiple R-squared:  0.9751, Adjusted R-squared:  0.972
## F-statistic: 313.3 on 1 and 8 DF,  p-value: 1.061e-07
```

The regression output provides estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$, allowing us to interpret the relationship between high school GPA and first-year college GPA.

Multiple Regression Analysis

So far, we have analyzed a simple regression where only one variable, high school GPA, was used to predict first-year college GPA.

However, we also need to consider multiple regression, which takes into account control variables that could influence the outcome variable, Y.

In our earlier regression, we concluded that high school GPA affects first-year college GPA. But let's consider a scenario where the dataset includes a variable for Study time. Let's hypothesize about the relationship between high school GPA, Study time, and college GPA:

Hypothesis: Students with higher high school GPAs are generally assumed to have better study habits, which leads to longer and more focused Study time. Therefore, the strong positive relationship between high school GPA and first-year college GPA might not directly reflect the effect of high school GPA itself. Instead, it could result from these better study habits, represented by Study time.

In this case, if we control for Study time, we effectively remove the indirect influence of high school GPA on college GPA that operates through good study habits. To better understand the relationship between high school GPA and first-year college GPA, we will now control for Study time. Controlling for this variable allows us to see whether the relationship between high school GPA and first-year GPA changes after accounting for its effects.

The regression model becomes:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

Where: - Y: First-year college GPA (dependent variable)

- X: High school GPA (main explanatory variable)

- Z: SAT scores (control variable)

- β_2: Coefficient for the control variable

```r
# Adjusting data so Study_time influences results significantly
regression_data$Study_time <- c(1000, 1100, 1200, 1500, 1400, 1350, 1550, 1450, 1600, 900)

# Performing multiple linear regression
multiple_regression <- lm(First_Year_GPA ~ High_School_GPA + Study_time, data = regression_d
summary(multiple_regression) # Display regression results
```

```
##
## Call:
## lm(formula = First_Year_GPA ~ High_School_GPA + Study_time, data = regression_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.053106 -0.024154 -0.004761  0.033242  0.043992
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.757e-01  1.585e-01   1.108   0.3043
## High_School_GPA  9.468e-01  6.658e-02  14.221 2.02e-06 ***
## Study_time      -1.629e-04  8.523e-05  -1.912   0.0975 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03855 on 7 degrees of freedom
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.979
## F-statistic: 210.5 on 2 and 7 DF,  p-value: 5.596e-07
```

Explanation about Results

Initially, the relationship between high school GPA and first-year GPA was observed without controlling for Study time. The results showed a strong positive relationship. After controlling Study time as a control variable, the coefficient for high school GPA decreased substantially, with its p-value increasing, indicating reduced significance. This suggests that Study time explain much of the variation in first-year GPA initially attributed to high school GPA.

2. Logistic Regression

What is Logistic Regression? and Why?

While linear regression is used to predict continuous outcomes, logistic regression is designed for binary dependent variables. For instance, logistic regression can be used for situations where the outcome is either yes or no, pass or fail, or success or failure. The key difference lies in the nature of the dependent variable: Logistic regression is essential when the dependent variable is binary, meaning it has only two possible outcomes. But why must we use logistic regression? Can't we just use the regression we are familiar with?

```r
# Create the dataset
data_select <- data.frame(
  X2urvived = c(1, 0, 1, 0, 1, 0, 0, 1, 1, 0),
  Pclass = c(3, 1, 3, 1, 3, 3, 2, 1, 2, 3)
)

# Check the first few rows
head(data_select)
```

```
##   X2urvived Pclass
## 1         1      3
## 2         0      1
## 3         1      3
## 4         0      1
## 5         1      3
## 6         0      3
```

Here, the outcome variable, Survived, indicates survival (1) or non-survival (0). The independent variable is Pclass, which represents the passenger's ticket class:

1 for first class, 2 for second class, 3 for third class. We will analyze how ticket class influenced survival using regression analysis.

```
data_select$Pclass <- as.factor(data_select$Pclass)

regression <- lm(X2urvived ~ Pclass, data_select)
regression
```

```
##
## Call:
## lm(formula = X2urvived ~ Pclass, data = data_select)
##
## Coefficients:
## (Intercept)      Pclass2      Pclass3
##      0.3333       0.1667       0.2667
```

Why Can't We Use OLS for Binary Outcomes? With just regression, these coefficients are not intuitive for binary outcomes. Our dependent variable, Survived, is binary (1 or 0), but the predicted results are neither 0 nor 1. In fact, OLS may produce predictions below 0 or above 1, which are nonsensical for probabilities.

What we want to understand is the probability that the dependent variable equals 1 (e.g., the probability of survival), given the independent variable. For example:

How much does the survival probability decrease for second-class passengers compared to first-class passengers?

OLS cannot provide a valid interpretation for this probability because:

Regression coefficients are not interpretable as changes in probabilities. Predicted values can exceed the range of [0, 1]. To properly model the relationship between binary outcomes and independent variables, logistic regression is required. This method ensures that the predicted probabilities fall between 0 and 1, and the coefficients can be interpreted as the effect of an independent variable on the log-odds of the outcome.

Odds ratio

In such cases, the dependent variable Y must be transformed to model probabilities effectively. We will use a special form ofY known as the odds ratio. Odds represent the ratio of the probability of an event occurring to the probability of it not occurring:

$$\text{Odds} = \frac{p}{1-p}$$

Where: - p: Probability of the event occurring.

- $1-p$: Probability of the event not occurring.

Odds provide a different way of representing probabilities by comparing the likelihood of an event occurring to the likelihood of it not occurring:

Using odds is advantageous because they are unbounded on the upper end ($0 < \text{Odds} < \infty$), allowing for easier interpretation of multiplicative changes. For example: - If the odds of survival are 2, this means the event is twice as likely to occur than not occur.

- If the odds are 0.5, the event is half as likely to occur compared to not occurring.

Then, why use odds ratio?

Odds ratios (ORs) are used to compare the odds of an event occurring between two groups. Instead of comparing raw probabilities, the odds ratio quantifies the relative likelihood of an event:

$$\text{Odds Ratio} = \frac{\text{Odds in Group 1}}{\text{Odds in Group 2}}$$

This is especially useful in logistic regression because: 1. **Interpretability:** The OR indicates how much more (or less) likely an event is to occur in one group compared to another. - OR > 1: Event is more likely in Group 1. - OR < 1: Event is less likely in Group 1. - OR $= 1$: Event is equally likely in both groups. 2. **Logistic Regression Coefficients:** Logistic regression models the log of the odds, making the OR a natural and interpretable measure of effect size.

Odds, Odds Ratio, and Log-Odds in Logistic Regression

Logistic regression models the **log-odds** (logarithm of the odds) instead of probabilities directly:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Why log-odds? - The log function maps odds ($0 < \text{Odds} < \infty$) to a continuous range ($-\infty < \text{Log-Odds} < \infty$).

- This ensures the logistic regression model can handle the full range of probabilities while maintaining mathematical consistency.

From the log-odds, we can calculate: - **Odds:** By exponentiating the log-odds.

- **Odds Ratio:** As the exponentiated logistic regression coefficient (e^{β_1}), representing the multiplicative change in odds for a one-unit increase in X.

Mathematical Explanation for Logistic Regression

The logistic regression equation is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Where: - p: Probability of the outcome (e.g., probability of passing) - $1-p$: Probability of the opposite outcome (e.g., probability of failing) - β_0, β_1: Regression coefficients - X: Independent variable

Applying Logistic Regression to Example Data

Let us apply logistic regression to a more realistic educational dataset. We will model whether a student passed a standardized math test based on the hours they studied and their attendance rate.

```r
# Generate synthetic educational data
set.seed(123)
n <- 200 # Number of students
data <- data.frame(
  Hours_Studied = runif(n, 0, 10), # Hours of study ranging from 0 to 10
  Attendance_Rate = runif(n, 50, 100) # Attendance rate ranging from 50% to 100%
)

# Outcome variable: Pass or Fail (binary)
data$Passed <- ifelse(
  0.4 * data$Hours_Studied + 0.03 * data$Attendance_Rate + rnorm(n, 0, 0.5) > 5, 1, 0
)

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

Creating a Synthetic Dataset

```
##   Hours_Studied Attendance_Rate Passed
## 1      2.875775        61.93630      0
## 2      7.883051        98.11795      1
## 3      4.089769        80.06829      0
## 4      8.830174        75.75149      1
## 5      9.404673        70.12867      1
## 6      0.455565        94.01233      0
```

The `Passed` variable is binary, indicating whether a student passed (1) or failed (0) the test. The independent variables are: - `Hours_Studied`: Hours spent studying. - `Attendance_Rate`: Attendance rate as a percentage.

Fitting the Logistic Regression Model We will use logistic regression to understand how hours of study and attendance rate influence the likelihood of passing the test.

```r
# Fit logistic regression model
logistic_model <- glm(Passed ~ Hours_Studied + Attendance_Rate, data = data, family = binomi

# Display logistic regression results
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Passed ~ Hours_Studied + Attendance_Rate, family = binomial,
##     data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -21.2934     3.9133  -5.441 5.29e-08 ***
## Hours_Studied     1.7708     0.2958   5.986 2.15e-09 ***
## Attendance_Rate   0.1207     0.0285   4.233 2.30e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 246.017  on 199  degrees of freedom
## Residual deviance:  77.752  on 197  degrees of freedom
## AIC: 83.752
##
## Number of Fisher Scoring iterations: 7
```

Interpreting the Results The logistic regression coefficients can be interpreted in terms of odds:

1. **Intercept (\(\beta_0\)):** The log-odds of passing when both hours studied and attendance rate are zero.
2. **Hours_Studied (\(\beta_1\)):** The change in log-odds of passing for a one-hour increase in study time.
3. **Attendance_Rate (\(\beta_2\)):** The change in log-odds of passing for a 1% increase in attendance rate.

Transforming coefficients into odds ratios provides a more intuitive understanding: - Each additional hour of study increases the likelihood of passing by approximately 5.87 times. - Each 1% increase in attendance rate increases the likelihood of passing by approximately 1.13 times.