

1.10: Linear Regression

Introduction to Regression

In a previous chapter, we discussed correlation analysis, which helps us understand the degree of association between two or more variables. Regression analysis is closely linked to correlation analysis, but it offers a more sophisticated way to explore the relationships among variables. Regression is a broad term encompassing a set of statistical methods used for modeling the relationship between a dependent variable and one or more independent variables, such as simple linear regression, multiple linear regression, polynomial regression, logistic regression, and so on. According to Vogt and Johnson (2011), regression analysis serves three primary purposes:

- Predicting the change in a dependent variable for each one-unit increase in an independent variable.
- Predicting the change in a dependent variable associated with a one-unit change in a specific independent variable, while controlling for other independent variables.
- Assessing how much better we can explain or predict a dependent variable by considering all the independent variables together.

Regression analysis is a powerful tool for understanding and quantifying the relationships between variables and making predictions based on those relationships. In this chapter, we will review two forms of linear regression: simple linear regression and multiple linear regression.

Simple Linear Regression Vs. Multiple Linear Regression

Linear regression analysis is commonly used to examine the relationship between one continuous dependent variable and a set of independent variables. Simple linear regression involves examining the linear relationship between the dependent variable and a single independent variable. Conversely, multiple linear regression entails analyzing the impact of multiple independent variables on a dependent variable in the linear relationships.

Ordinary Least Squares (OLS) Model

It is important to note that various types of linear regression models exist, but we will focus on the Ordinary Least Squares (OLS) regression model in this chapter because it is the most widely used. OLS is a statistical estimation technique for determining a regression equation that best represents the relationship between the dependent and independent variables. This method calculates the slope and intercept by minimizing the sum of squared differences between observed and predicted values. Other statistical estimation methods, such as maximum likelihood, are available for establishing a regression model.

Inmate Self-Reported Survey

In the previous chapters, I have discussed various data collection methods (e.g., Uniform Crime Report or National Crime Victimization Survey). Police departments and residents in the community can be great sources of data related to crime, but one source of the data we have not covered yet is inmates. Many inmates are in jails or prisons because they are arrested, prosecuted, and convicted for their accused crimes. If we survey inmates, they may provide useful information that can help us understand crime from offenders' perspectives. This is part of the reason why inmates have been used for various academic articles. For this chapter, I will use the data from an inmate self-reported survey conducted in Korea (Choi & Dulisse, 2021). Specifically, we will first perform a simple linear regression to investigate the relationship between low self-control and risky lifestyles among inmates. Following that, we will conduct a multiple linear regression analysis, considering both low self-control and age as predictors, while evaluating their impact on risky lifestyle, which serves as the dependent variable.

Let's first load the data. You will download the data from [the shared Google Drive folder containing the Inmate Survey.sav data](#). The next steps should be familiar to you at this stage.

```
library(haven)

Inmate_Survey <- read_sav("C:/Users/75JCH0I/OneDrive - West Chester
University of PA/WCU Research/R/data/Inmate Survey.sav")

View(Inmate_Survey)
```

A total of 986 inmates from 20 geographically distinct prisons participated in this survey. Risky lifestyles (RL) were assessed using four items that gauge involvement in unstructured criminogenic activities within the prison environment: (a) possession of prohibited items, (b) breaking away from designated areas, (c) participation in gambling, and (d) involvement in illegal transactions of prohibited products. Participants rated each item on a scale ranging from 0 (never) to 4 (more than 10 times). The scores for these items were summed to obtain a composite measure of risky lifestyles, with higher scores indicating greater involvement. This set of items demonstrates strong internal consistency, prompting students to recall their understanding of reliability testing. Age (AGE) is a continuous variable representing the participants' age. Low self-control (LSC) was assessed based on six items: "I prefer to do things physically rather than verbally," "When encountering difficult or complicated tasks, I usually give up," "I lose my temper easily," "I enjoy doing things that are a little exciting," "I often tease others," and "I prioritize immediate pleasure." A composite measure of low self-control was created by summing the scores on these six items, with higher scores indicating lower levels of self-control.

Assumptions of Linear Regression

In the previous chapter, I emphasized the importance of checking multiple assumptions when conducting statistical analyses, as violating these assumptions can significantly impact linear regression results and lead to biased estimates of coefficients. To ensure the validity of our analysis, we need to consider several additional assumptions. Some of these may already be familiar to you, as they were also necessary for correlation analysis.

- Each observation in our dataset should be independent of the others.
- The outcome variable we are analyzing should be continuous.
- The relationship between the outcome variable and each continuous predictor should be linear.
- The variance of the outcome variable should be constant across all levels of the predictors, with points evenly distributed around the regression line.
- The residuals (the differences between observed and predicted values) should be independent of each other.
- The residuals should follow a normal distribution.
- There should be no strong correlations among the predictor variables, as this can cause numerical instability in the estimation of coefficients.

There are various methods available in R to assess these assumptions. However, discussing them in detail would exceed the scope of our current analysis. For the purposes of demonstration, we will proceed with the analysis, assuming that these assumptions have been met.

A Scatterplot of Low Self-Control and Risky Lifestyles

We can create a scatterplot of low self-control and risky lifestyles to explore the relationship between these two variables.

```
library(tidyverse)

Inmate_Survey %>%

  ggplot(aes(x = LSC, y = RL, color = "Points")) +
  geom_point(aes(size = "id"), color = "purple", alpha = 0.5) +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se =
    FALSE) +
  theme_minimal() +
  labs(y = "Risky Lifestyles", x = "LSC", color = "", shape = "") +
  scale_size_manual(values = 2, name = "")
```

The graph above should give us a general idea regarding a bivariate relationship between the two variables. The `geom_smooth()` function with `method = "lm"` is used to add a linear regression line to the plot, which represents the best-fit straight line through the data points. The `aes(color = "Linear fit line")` part specifies that the color of this linear fit line will be labeled as "Linear fit line" in the legend, making it distinguishable from other elements in the plot. Setting `se= FALSE` means that the standard error bands around the linear regression line will not be displayed on the plot. These bands are typically shown by default to indicate the uncertainty or variability of the regression line, but in this case, they are disabled. This line goes up from left to right, showing a

positive relationship between low self-control and risky lifestyles. Those with low self-control were more likely to engage in risky lifestyles, which makes sense.

Checking a Correlation Coefficient

You may want to confirm if there is a positive correlation between low self-control and risky lifestyles. You can conduct a correlation analysis such as the one we covered in the previous chapter.

```
library(tidyverse)

Inmate_Survey %>%
  summarize(correlation_lsc_rl = cor(LSC, RL, use = "pairwise.complete.obs"),
    sample_size = n())
```

There were missing values in our dataset. The `pairwise.complete.obs` argument allows us to compute the correlation using complete pairs of observations, effectively handling missing values pairwise.

Conducting Simple Linear Regression Analysis

We'll now proceed to calculate the slope and intercept using the Ordinary Least Squares (OLS) method. OLS is employed to minimize the sum of squared differences between the observed and predicted values of the dependent variable by minimizing the overall distance between the data points and the regression line. The y-intercept represents the value of risky lifestyle when low self-control is zero. Meanwhile, the slope denotes the change in risky lifestyle for every one-unit change in low self-control.

```
rl_by_lsc <- lm(formula = RL ~ LSC,
  data = Inmate_Survey, na.action = na.exclude)
summary(object = rl_by_lsc)
```

The linear regression model `rl_by_lsc` predicts the dependent variable RL (risky lifestyle) based on the independent variable LSC (low self-control) using the `lm()` function in R. The `na.action = na.exclude` argument ensures that observations with missing values are included in the analysis rather than being removed.

Based on the results, we can write down the regression equation for our model:

- Risky lifestyles = $-0.34 + 0.09 \times \text{low self-control}$

This means that if low self-control increases by one unit in an inmate, risky lifestyles would typically change by 0.09227.

NHST Steps for Simple Linear Regression Model

We may want to make inferences about the population (all inmates within the 20 prisons in South Korea where the current sample was drawn from) using the data we have. That is when we conduct Null Hypothesis Significance Testing (NHST), which was covered previously. Specifically, we may want to assess the statistical significance of the slope in simple linear regression. If the slope (i.e., the unstandardized coefficient of low self-control or the rate of change in risky lifestyle for a one-unit change in low self-control) is not equal to zero, it implies that there is a statistically significant relationship between low self-control and risky lifestyles.

Step 1: Formulate the Null and Alternative Hypotheses.

- H_0 : The unstandardized coefficient of low self-control is equal to zero.
- H_A : The unstandardized coefficient of low self-control is not equal to zero.

Step 2: Calculate the Test Statistic.

The test statistic for the significance of the unstandardized coefficient in OLS regression is the t-statistic we used previously (aka the Wald test).

```
rl_by_lsc <- lm(formula = RL ~ LSC,
  data = Inmate_Survey, na.action = na.exclude)
summary(object = rl_by_lsc)
```

The results indicate that the unstandardized coefficient of low self-control is 0.09, and the t-value is 6.023.

Step 3: Determine the Probability (P-Value) of Obtaining a Test

Statistic at Least as Extreme as the Observed Value, Assuming no Relationship Exists. We can see that the p-value of $< 2.45e-09$ for the unstandardized coefficient of low self-control is 0.09.

Steps 4 & 5: If the P-value is Very Small, Typically Less Than

5%, Reject the Null Hypothesis, but if the P-Value is Not Small, Typically 5% or Greater, Retain the Null Hypothesis. The p-value of < 0.05 in our simple linear regression model suggests that there is a very slim probability that the t-statistic for the unstandardized coefficient of low self-control would be as large as observed if the null hypothesis were true. In short, the null hypothesis was rejected in favor of our alternative hypothesis that the unstandardized coefficient of low self-control is not equal to zero.

Reporting the Results From the Simple Linear Regression Model

We found that low self-control reported by inmates is a statistically significant predictor of risky lifestyles ($b = 0.09$; $p < .05$) within our sample. Specifically, for every one-unit increase in low self-control among inmates, the predicted increase in risky lifestyle is 0.09 units.

Model Significance for Simple Linear Regression

You might have noticed another p-value toward the bottom of the output, adjacent to the F-statistic for the linear regression. This p-value corresponds to a test statistic that evaluates the improvement of the regression line's fit to the data points compared to the mean value of our dependent variable (risky lifestyles). The F-statistic serves as the test statistic for linear regression, assessing how well the regression line fits compared to the mean value of risky lifestyles. The model fit can be tested by following the NHST steps that we used above.

Step 1: Formulate the Null and Alternative Hypotheses.

- H_0 : A model including low self-control is not better at explaining risky lifestyles than a baseline model using the mean value of risky lifestyles.
- H_A : A model including low self-control is better at explaining risky lifestyles than a baseline model using the mean value of risky lifestyles.

Step 2: Calculate the Test Statistic.

From the provided output above, you can identify the F-value as $F(1, 941) = 36.28$.

Step 3: Determine the Probability (P-Value) of Obtaining a Test Statistic at Least as Extreme as the Observed Value, Assuming no Relationship Exists.

The probability of observing an F-value as large as 36.28, or even larger, if the null hypothesis were true, is very low ($p < 0.05$).

Steps 4 & 5: If the P-Value is Very Small, Typically Less Than 5%, Reject the Null Hypothesis, but if the P-Value is Not Small, Typically 5% or Greater, Retain the null Hypothesis.

Given the small p-value, we can reject the null hypothesis in favor of the alternative hypothesis that a model including low self-control is better at explaining risky lifestyles than a baseline model using the mean value of risky lifestyles.

Reporting the Model Significance for the Simple Linear Regression Model

You can add the results regarding the model significance when reporting the results from simple linear regression. Our model significantly outperformed the baseline model (which used the mean of risky lifestyles) in explaining risky lifestyles ($F(1, 941) = 36.28$; $p < .05$).

Conducting Multiple Linear Regression

Multiple linear regression involves incorporating multiple independent variables to predict the dependent variable. In the context of predicting risky lifestyles among inmates, it's clear that factors beyond just low self-control may play a role. For instance, age could be a significant demographic factor, as younger individuals might be more inclined to engage in risky behaviors compared to older inmates.

Therefore, multiple linear regression is better suited for real-life scenarios where multiple factors influence dependent variables. All you need to do is to tweak the R codes that we used to perform a simple linear regression.

```
rl_by_lsc_age<-lm(formula = RL ~ LSC + AGE,  
data = Inmate_Survey, na.action = na.exclude)  
summary(object = rl_by_lsc_age)
```

As you can see, even after including age in our regression model, low self-control remained statistically significant. Low self-control was positively and significantly associated with risky lifestyles ($b = 0.09$; $t = 5.51$; $p < .05$). Age was also a significant predictor of risky lifestyles ($b = -0.01$; $t = -2.30$; $p < .05$). Age was negatively and significantly associated with risky lifestyles.

Model Fit for Linear Regression

In the outputs for both simple linear regression and multiple linear regression, you may have observed multiple R-squared and adjusted R-squared values located just above the F-statistic. These statistics serve to evaluate the overall goodness of fit of the regression model. R-squared (aka the coefficient of determination) is calculated by determining the proportion of variance in the dependent variable that can be explained by the independent variables incorporated in the model. It ranges from 0 to 1, with 0 signifying that the independent variables account for none of the variance in the dependent variable, and 1 indicating that they explain all of the variance.

In our multiple linear regression, for instance, the R-squared value is 0.04252. To determine the percentage of variance explained by the model, multiply this value by 100. Therefore, 4.25% of the variance in risky lifestyles is explained by both low self-control and age. Now, what is adjusted R-squared? As additional variables are added to the model, the R-squared value tends to increase. Adjusted R-squared serves to counteract this tendency by slightly penalizing the R-squared value for each additional variable introduced into the model. This adjustment ensures that the measure appropriately accounts for the complexity of the model and prevents overestimation of its explanatory power.

Conclusion

This chapter introduced the concepts of simple and multiple linear regression, demonstrating how one or more independent variables can be used to predict a single dependent variable. I aimed to give those interested in becoming crime analysts an overview of basic statistics and how R can be employed to conduct statistical analyses. However, it is important to note that this book represents just the starting point of your exploration into statistics and the practical applications of the R programming language. There is much more to discover, and I encourage you to continue your journey toward a deeper comprehension of statistics and the versatile capabilities of R.

References

Choi, J., & Dulisse, B. (2021). Behind closed doors: The role of risky lifestyles and victimization experiences on fear of future victimization among South Korean inmates. *Journal of Interpersonal Violence*, 36(21-22), 10817 –10841. <https://doi.org/10.1177/0886260519888186>

This page titled [1.10: Linear Regression](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Jaeyong Choi](#) (The Pennsylvania Alliance for Design of Open Textbooks (PA-ADOPT)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.