Interpreting Data for a Continuous Dependent Variable

Example 1. Consider a toy example where we are interested in whether studying more increases your likelihood of getting a better grade. We have some intuition that there is a positive effect between the hours one diligent student spends studying and their high grades because studying helps people reinforce material so they are likely to do better on an exam. This leads us to create the following hypothesis:

Hypothesis

Hyp: The more hours a student spends studying, the higher their grade on the exam will be.

Note that we can test whether this hypothesis is true by collecting data on how many hours a student studies for the midterm exam and comparing that to their actual grade. If there is a strong positive relationship between the two, then we expect to see more hours studying corresponding to a higher grade.

Let’s pretend we collect data from two different sections about (1) how many hours students spent studying before the exam and (2) their final exam score.

Evidence Part 1

Data for Section 1: Midterm Exam Grade by Hours Spent Studying

Discussion: The results from section one are rather unclear. We can see that the hours spent studying for the exam seem to be uncorrelated (or unrelated) to their final score on the midterm exam. That is, we say that the correlation between the independent variable (hours spent studying) and the outcome variable (grade) are unrelated to each other. Similarly, we can see from the best fit line here that the slope is also zero. That is, a one hour increase in studying for the midterm exam increases a student’s grade by zero points.
Evidence Part 2

Discussion: The results from section two seem to support our hypothesis. Why? First, the association between hours spent studying and the grade on the exam is positive - that is studying more tend to correlate with you having a higher grade on the exam. Second, the slope estimate from the best fit line is larger. It says that a one-hour increase in studying for the midterm exam increases a student’s grade by ten points.
Example 2. Suppose we don’t know what a student’s actual score on the exam was, but only whether or not they passed the test. We call this outcome variable a binary variable because there are only two possible categories this outcome may take (pass or fail). How can we test whether the hours spent studying also affects whether a student passes? We will continue to rely on data from section 2 in order to ask this question in another way.

Reading a Contingency Table

For the purposes of the data analysis assignment, we are providing you with a contingency table or cross-tabulation of the data to see how many countries fall into each category. To understand the cross-tabs, we will build up by considering variation in the outcome variable and then variation in the outcome variable and independent variable. Let’s take a look at some descriptive statistics or raw counts of how students did defining a pass on the test to be a 71 or higher.

<table>
<thead>
<tr>
<th>Fail</th>
<th>Pass</th>
<th>Total Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>48</td>
<td>100</td>
</tr>
</tbody>
</table>

This shows us that 52 students failed and 48 students passed. What happens when we consider how long they studied? Let’s make the following hypothesis:

To simplify matters we will consider whether students studied above or below the median.

```
median(hours.spent.studying)  #median number hours students study
```

```
## [1] 2.382958
```

This says that 50% of students studied for 2.39 hours of less and 50% of students studied for 2.4 hours or more. This leads us to make the following hypothesis.

Hyp: Students who study more than 2.4 hours are more likely to pass.

We can see whether they are important differences in whether students who studied more are also more likely to have passed by looking at the cross-tabs here:

<table>
<thead>
<tr>
<th></th>
<th>Fail</th>
<th>Pass</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studied Less than 2.4 Hours</td>
<td>44</td>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td>Studied More than 2.4 Hours</td>
<td>8</td>
<td>42</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>48</td>
<td>100</td>
</tr>
</tbody>
</table>

What does this table mean?

- 44 students failed the test and studied less than 2.4 hours for it.
- 6 students passed the test and studied less than 2.4 hours for it.
Analyzing a Contingency Table

The raw counts here provide suggestive evidence for our hypothesis that students who study more are more likely to pass the test, but we would like to formally look at this by considering the marginal probabilities to see whether this is significant.

That is, what is:

Probability Student Passes Test if Study Less than 2.4 Hours versus Probability Student Passes Test if Study More than 2.4 Hours

We can find the probability a student passes the test if they study less than 2.4 hours by considering:

\[ \frac{\text{No. of Students who Pass and Study Less than 2.4 Hours}}{\text{No. of Students who Study Less than 2.4 Hours}} = \frac{6}{50} = 0.12 \]

Substantively, this means only 12% of students who study for less than 2.4 hours pass. Yikes!

We want to compare this with:

\[ \frac{\text{No. of Students who Pass and Study More than 2.4 Hours}}{\text{No. of Students who Study More than 2.4 Hours}} = \frac{42}{50} = 0.84 = 84\% \]

This means that 84% of students who study for more than 2.4 hours will pass. Yay!

The difference between the probability that a student passes a test if they study less than 2.4 hours versus the probability that a student passes a test if they study more than 2.4 hours is quite large (.84 - .12 = .72). We want to look for two things: (1) is the relationship in the direction we thought? (that is, are students who study more more likely to pass?) and (2) is there a large difference between those who pass depending on how long they study?

(1) Is the relationship in the direction we thought? (that is, are students who study more more likely to pass?)

Answer: Yes. It is above 50% which means students who study more are more likely to pass than fail.

(2) Is there a large difference between those who pass depending on how long they study?

Answer: Yes. As a rule of thumb, we will say that if the difference between the two is at least 0.1 (or 10%) then the difference is unlikely to be random and we take this to mean that studying for more than 2.4 hours is associated with a large and positive effect of passing.

This means the data supports our hypothesis above.

---

\(^1\)We do this in case the counts for the two categories are not equal which could then be misleading support for our hypothesis.