Statistical Thinking for Forensic Practitioners

Quiz on Part 7: Regression & Analysis of Variance

Fall 2020

1 Testing for significant correlation

You may want to use Excel for some of the calculations in this section. Slides 5-11 from the Part 7 lecture slides and the quiz/lab exercises on hypothesis testing from Part 6 will be useful references.

Research has demonstrated that the elemental composition can be used to differentiate between two panes of float glass (Example). To this end, it may be useful to determine whether the prevalence of one element is an indicator for the prevalence of another (i.e., an association exists between the concentrations of two elements). Forensic glass analysts were interested in determining whether the prevalence of the element Titanium (specifically the isotope Ti49) was associated with the prevalence of the element Iron (Fe57) in float glass manufactured at a local company. They collected a sample of $n = 1486$ glass fragments from the company and measured the fragments' elemental concentrations.

Let x_i denote the concentration of titanium and y_i the concentration of iron (measured in parts per million) in the ith glass fragment, $i = 1, ..., 1486$. Following are some summary statistics calculated from this data set.

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 44.408$$

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 155.682$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 23.411$$

1. The statistics provided above are enough to perform a hypothesis test to answer the research question. Perform this hypothesis test at the $\alpha = .05$ level. Your answer should include null and alternative hypotheses for a population parameter of interest, a test statistic, a determination of a critical value or p-value (or both), and a conclusion.

2. Suppose you meet an alchemist and show them the results of this hypothesis test. They exclaim: "This is proof that we *can* transmute one metal into another! Clearly having more iron in a pane of glass gives rise to more titanium." How would you respond to the alchemist?

2 Logistic regression example

We will practice forming and interpreting a logistic regression model.

A research team from Bowling Green State University in Ohio conducted a study in which they fit a logistic regression model to determine variables that impact whether a sexual assault kit (SAK) contains a probative DNA profile that is eligible for the Combined DNA Index System (CODIS) database (Source). As they say, this study was performed to "combat the influx of sexual assault kits (SAKs) that need to be tested." The researchers collected a sample of 2500 SAKs and considered predictors such as the age of the victim, the number of days between assault and SAK collection, and others. We will consider a model using age of the victim as the first predictor, x_{1i}, and the number of days between assault and SAK collection as the second predictor, x_{2i}.

1. Why might a logistic regression model be appropriate to analyze the variable of interest? (Hint: first identify the variable of interest. Then consider why the underlying probability distribution assumed for logistic regression might be an appropriate model.)

2. Define a logistic regression model for this situation. Your answer should include a definition of the response variable (i.e., what is y_i), the probability model assumed for each response, and an equation that defines an appropriate relationship between the parameter of interest and a function of the predictors x_{1i}, x_{2i}. (Note: refer to slide 87 in the lecture slides noting that we have one more predictor in this problem than the example described in lecture. Assume $y_i = 1$ when the variable of interest is in the affirmative.)

3. Interpret the parameter of interest.

4. The estimated coefficients are $b_0 = -2.11, b_1 = 0.3$, and $b_2 = -1.2$. Use these estimates to express the estimated parameter of interest based on the relationship you defined above.

5. Suppose we obtain a new SAK, #2501, where $x_{1,2501} = 26$ and $x_{2,2501} = 3$. Find and interpret an estimate for the parameter of interest.

6. Suppose we are interested in SAK #311 from our sample. In particular, we are interested in determining whether the probability this SAK contains a probative DNA profile eligible for CODIS is not equal to 0.50. Suppose an approximate 95% confidence interval for $\beta_0 + \beta_1 x_{1,311} + \beta_2 x_{2,311}$ is $[-.403, .983]$. Conduct a hypothesis test at the $\alpha = .05$ level to answer this question. (Note: refer to slide 94 and remember to include your hypotheses and conclusion in your answer.)

7. Suppose the researchers hypothesize that the effect of the number of days between an assault and SAK collection on the parameter of interest depends on the age of the victim. Propose an extended model that incorporates

this hypothesis. (Note: you only need to change the assumed relationship between the parameter of interest and predictors. Slides 64-67 may prove useful.)

8. The data for this study were collected from only within the state of Ohio. What model assumption(s) might be thrown into question due to the scope of the data collection?