Bios702 fall 2024 course notes

Contents

[Introduction 3](#_Toc174960392)

[CONSORT diagram 5](#_Toc174960393)

[Descriptive presentation of one variable 9](#_Toc174960394)

[Goals of the descriptive presentation 9](#_Toc174960395)

[Binary variables 10](#_Toc174960396)

[Unordered categorical variables 13](#_Toc174960397)

[Ordered categorical variables 16](#_Toc174960398)

[Continuous variables 18](#_Toc174960399)

[Structure of table 1 19](#_Toc174960400)

[1-sample inference 20](#_Toc174960401)

[Binary outcome 20](#_Toc174960402)

[Unordered categorical outcome 24](#_Toc174960403)

[Ordered categorical outcome 25](#_Toc174960404)

[Continuous outcome 26](#_Toc174960405)

[Models with 1 predictor and one outcome 27](#_Toc174960406)

[Introduction 27](#_Toc174960407)

[Continuous outcome and continuous predictor 28](#_Toc174960408)

[Introduction 28](#_Toc174960409)

[Visualization 29](#_Toc174960410)

[Notation 32](#_Toc174960411)

[Fitting the model 34](#_Toc174960412)

[Matrix version of the calculations 37](#_Toc174960413)

[Meaning of the estimated regression parameters 41](#_Toc174960414)

[Data transformations 42](#_Toc174960415)

[Extensions 45](#_Toc174960416)

[ANOVA table 46](#_Toc174960417)

[Generalizations and special cases 49](#_Toc174960418)

[Nested models 51](#_Toc174960419)

[Robustness and influence 52](#_Toc174960420)

[Different functional forms 54](#_Toc174960421)

[Interpreting the results 56](#_Toc174960422)

[Observer agreement 57](#_Toc174960423)

[Validation 59](#_Toc174960424)

[Continuous outcome and one categorical predictor 61](#_Toc174960425)

[2 categories 61](#_Toc174960426)

[More than 2 categories (1-way ANOVA) 64](#_Toc174960427)

[Continuous outcome and 2 predictors 69](#_Toc174960428)

[2 categorical predictors 69](#_Toc174960429)

[1 categorical predictor and 1 continuous predictor 75](#_Toc174960430)

[Binary outcome 79](#_Toc174960431)

[Introduction 79](#_Toc174960432)

[Transformations 80](#_Toc174960433)

[Logistic regression with 1 binary predictor 82](#_Toc174960434)

[Technical details 85](#_Toc174960435)

[Logistic regression with 1 continuous predictor 87](#_Toc174960436)

[Logistic regression with 2 categorical predictors 90](#_Toc174960437)

[Chi-square tests 94](#_Toc174960438)

[Introduction 94](#_Toc174960439)

[Chi-square test for a binary outcome and a binary predictor 95](#_Toc174960440)

[Chi-square test for unordered categorical variables 98](#_Toc174960441)

[Chi-square test for stratified 2x2 tables 99](#_Toc174960442)

[Other topics 101](#_Toc174960443)

[Ordinal categorical outcomes 101](#_Toc174960444)

[Diagnostic testing 105](#_Toc174960445)

[2x2 measures of observer agreement 107](#_Toc174960446)

# Introduction

This is a course in basic data analysis.

We will be using the visualize, analyze, interpret (VAI) framework as a way of organizing how those analyses are performed. In practice, these steps need not only proceed in a single direction: for example, after you have made a first pass through the data and produced a preliminary interpretation this might lead you to fine-tune your research questions and your analyses, this fine-tuning might require you to return to visualize the data in a more focused and sophisticated fashion, etc.

As part of the interpretation step, we'll emphasize how results can be effectively communicated. You'll have the opportunity to practice communication individually and in small groups, orally and in writing, etc.

Most of the content of this course will distinguish between predictor variables and outcome variables. After discussing how to work with single variables, we will consider models with continuous outcomes and categorical outcomes, the latter focusing on binary outcomes. Content that will be deferred until BIOS705 includes count outcomes, time-to-event outcomes, models with multiple predictors (with the exception of a few such models introduced in BIOS702), and models that don't distinguish between predictors and outcomes.

Most of you will already had some exposure to the material in this course, especially the introductory topics. Nevertheless, we will start from scratch. In all likelihood, we will be taking a somewhat different approach to the course content than what you've previously encountered, and don't anticipate that you will be bored.

These course notes are a high-level summary, and aren't intended to be comprehensive. They have varying levels of detail. The intention is to illustrate how things work and how they fit together. These course notes are also intended to provide background for the exercises.

I hope you enjoy the course!

# CONSORT diagram

The main topic in BIOS702 is simple linear regression (SLR). It has one predictor variable (regardless of its scale of measurement) and one continuously-scaled outcome variable. However, before any analyses that connect predictors with the outcome you should (1) describe the flow of study participants; and (2) perform descriptive analyses of each of the variables in your dataset. This module covers the first topic.

To describe the flow of study participants, we consider (1) why do you do it; and (2) how do you do it. "Why" is to assess internal and external validity. "How" is accomplished using a CONSORT diagram or its functional equivalent. (CONSORT diagrams were originally developed for randomized trials, and some people limit the use of that terminology to trials.)

Example: randomized trial

Please refer to the figure in the Word document "CONSORT diagram example", which illustrates a generic CONSORT diagram. It uses a template downloaded from https://www.researchgate.net/figure/Algorithm-of-study-flow-Revised-template-of-the-CONSORT-diagram-showing-the-flow-of\_fig1\_236598999.

The top part of the figure begins with everyone that was considered for enrollment in the study and ends with those who were actually included. This pertains to the study's external validity.

The bottom part of the figure describes the patients who were analyzed, including study attrition due to loss to follow-up (and, sometimes, missing values). This pertains to the study's internal validity.

When presenting a CONSORT diagram to an investigator, you should include a narrative summary and interpretation. An example narrative summary follows, which is based on the figure in "CONSORT diagram example". Our comments on the narrative summary are italicized.

1,000 people were assessed but only 100 were randomized. 800 of the 900 excluded patients failed to meet the inclusion criteria, which might suggest that a different data source (i.e., one with a greater percentage of eligible patients), or a different initial screening algorithm, should be used in the future. Of the remaining 200 candidates, 100 were randomized but 100 were not, 90 of which refused to participate. Thus, the refusal rate among potentially eligible patients was 45% (i.e., 90/200), which should be considered when assessing generalizability. *Whether 45% is "too high" depends on context, but should be commented upon as it reduces generalizability.*

Among the randomized patients, it appears that some method was used to ensure perfect balance between the groups. 95 of the 100 randomized patients received their intended intervention -- all 50 in the usual care group and 45 of 50 in the intervention group. It appears that a prompt in the electronic medical record was part of the intervention, and this prompt didn't work properly for 5 of the patients.

*In the example, the loss to follow-up was reassuringly trivial. You should especially check for cases when the loss to follow-up is notably greater in the intervention group, as it suggests that the intervention was so poorly tolerated that it caused some of the participants to drop out. This induces a potential bias in favor of the intervention, when those patients who dropped out were also doing poorly on the primary outcome variable.*

2 patients in the usual care group were excluded because they were missing the primary outcome variable, which is a reassuringly small number. 5 patients from the intervention group were excluded because they didn't receive the intervention. This is acceptable in some circumstances, but does violate the "intention to treat" principle that, in general, subjects should be analyzed according to their assigned intervention, regardless of whether or not they actually received it. *The structure of the CONSORT diagram has highlighted this departure from usual practice.*

Example: observational study

This is a study of patients with chronic pain. Pain is measured on a scale from 0 (no pain) to 10 (worst possible pain) at each clinic visit. Duke is considering a practice guideline where all patients with a pain scale of 5 or more receive an intervention. The intervention is time-consuming and costly, and is only worth performing if patients don't get better on their own. Accordingly, the investigators scanned the electronic medical record from January 1, 2023 to December 30, 2023, identified everyone with pain scores of 5 or more (i.e., the predictor variable, X), and recorded the pain score at their next visit (i.e., the outcome variable, Y). We found:

* 4,000,000 individual visit records were considered
* Of these, 1,000,000 had non-missing values for the pain scale
* All 1,000,000 values were in the correct range of 0-10
* These 1,000,000 values came from 200,000 distinct patients
* Of these 200,000 distinct patients, 10,000 had at least one pain score of 5-10
* Considering the first pain score that was >=5 for these 10,000 patients, 5,000 had a subsequent pain score.
* For the subsequent pain score, 1,000 values were missing.

A CONSORT diagram would reflect the above. A complicating factor is that there are two possible units of analysis: (1) the patient; and (2) the visit, and this would affect how the diagram is organized. Here, we are assuming that the algorithm is (1) find the first pain score for a patient (and exclude the rest); (2) exclude those patients without a subsequent pain score; and (3) cross-classify the first and second pain score, even if that subsequent score is missing.

The interpretation of the flow of study participants would focus on various points. Perhaps most importantly, and relevant to the external validity of a regression of Y on X, is that only 25% of the records had a pain score. Before proceeding further, you should ask the investigator to discover why -- perhaps there was an error in how the dataset was created, especially since it was reported to you that pain is measured at each visit. If the dataset is eventually used as is, the investigator should mention that only having pain scores for 25% of the dataset limits external validity.

It's never too soon to begin thinking about the analysis plan. An initial descriptive presentation would be a contingency table of Y versus X. For concreteness, the skeleton of such a table is presented below

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Y=missing | Y=0 | Y=1 | Y=2 | Y=3 | Y=4 | Y=5 | Y=6 | Y=7 | Y=8 | Y=9 | Y=10 |
| X=5 |  |  |  |  |  |  |  |  |  |  |  |  |
| X=6 |  |  |  |  |  |  |  |  |  |  |  |  |
| X=7 |  |  |  |  |  |  |  |  |  |  |  |  |
| X=8 |  |  |  |  |  |  |  |  |  |  |  |  |
| X=9 |  |  |  |  |  |  |  |  |  |  |  |  |
| X=10 |  |  |  |  |  |  |  |  |  |  |  |  |

One of the questions for discussion with the investigator concerns what to do if the pain scores on consecutive visits are 6, missing, missing, 4. Should these be analyzed as "6, missing" or "6,4"? There isn't a "statistically right" answer, and the choice of what to do depends on fine-tuning the scientific question -- in other words, does it pertain to "next visit" or does it pertain to "next visit with a non-missing pain score"?

# Descriptive presentation of one variable

## Goals of the descriptive presentation

Presentations on single variables typically have one of two goals. Goal 1 is simply to describe the overall study population -- for example, so that a reader can assess how similar the study population is to one that they are interested in. In the pain example, the reader might be interested in the average patient age, primary diagnosis, etc. This goal 1 is often satisfied by "Table 1" of a manuscript.

Goal 2 is to provide more detail on key variables used in subsequent modeling. This detail might support (for the analyst) and anticipate (for the audience) subsequent analytical decisions. As example, if there are relatively few patients with pain scores of 8 or more and they can be similarly interpreted as "extreme pain", then these patients might be grouped together into a "bin" of 8-10. Descriptive presentations for goal 2 typically contain more detail than those for goal 1. In a manuscript, this additional detail provides the reader with a contextual clue that the variables in question are more important than the others.

One of our references (Vickers *et al*) provides sound advice about whether information should be placed in figures, tables or text: *"Tables are predominately used to report numerical findings in greater detail than would be appropriate in the text. Figures are generally used to give an immediate impression of the data that is not easy to convey with words or numbers."* It also provides excellent general advice about how to create effective tables and figures.

As with any element of a statistical presentation, for descriptive presentations you will probably perform an extensive analysis and then pick and choose which results to show. We will describe some of the things that can be done, and also how you might decide on which results to present.

Our presentation is organized by scale of measurement: namely, binary, unordered categorical, ordered categorical and continuous.

## Binary variables

Binary variables contain very little information -- essentially, just the frequency of each of the two groups. As with any variable, the presence of missing values should also be considered.

Suppose that gender is collected as a binary variable. The output from R is:

* Female 50
* Male 50
* Missing 100

You might literally cut and paste this output into your statistical report. However, the default output from R isn't necessarily in an ideal format.

For example, you might add proportions, with missing values included in the denominator. This would be especially reasonable if missing values are to be treated as a separate category in the subsequent analysis. For example, gender might be treated as a predictor with 3 categories (i.e., female, male, missing). This is sometimes done in order to avoid excluding the subjects with missing values of gender.

* Female 50 (25%)
* Male 50 (25%)
* Missing 100 (50%)

Or, you might drop the missing category from the calculation of the proportions. This structure allows you to (1) highlight the number of missing values; and (2) describe the gender distribution among those with non-missing data. In general, your audience will be more interested that "50% of the subjects are female" than "25% of the observations in the dataset are from females".

* Female 50 (50%)
* Male 50 (50%)
* Missing 100

Or, you might move the information about missingness into the header. If the sample size is known, the number of missing values is implicit. This structure still provides information about both (1) the number of missing values; and (2) the gender distribution among those with non-missing data, and highlights the latter.

Gender (n=100)

* Female 50 (50%)
* Male 50 (50%)

In the header, you could either report the number of non-missing values (as above) or the number of missing values:

Gender (100 missing)

* Female 50 (50%)
* Male 50 (50%)

Or, because the frequencies must add to 100%, you might drop a row. This avoids presenting redundant information.

Gender (n=100)

* Female 50 (50%)

Or, you might embed the denominator into the body of the row. This moves the information about missingness to another location.

Gender

* Female 50/100 (50%)

Or, you might highlight the percentage by presenting it first. This recognizes that your audience will be more interested that the sample is "50% female" than "50 females were sampled".

Gender

* Female 50% (50/100)

As general recommendations:

* Unless "missing" will be used as one of the categories of an analysis variable, don't include it in the calculation of the proportion.
* Don't include the redundant second category without a good reason to do so. If the second category isn't obvious, you could label the row as "female, not male".
* Proportions are more fundamental information than raw numbers, and should be highlighted.
* Confidence intervals aren't usually needed for Table 1.

The above recommendations pertain to the descriptive "Table 1" that typically begins a manuscript. If the binary variable is one of the main analytical variables -- for example, the primary predictor or the primary outcome -- it can still be summarized as above. Especially if it is an outcome variable, a confidence interval can be added. For example: "50% (i.e., 50/100) of subjects were female: the 95% confidence interval is 40%-60%".

Although you will sometimes see this done in the literature, in general it is poor practice to report proportions using histograms, pie charts or other graphical methods, as there is too little information to warrant the use of a figure (i.e., the information can be adequately conveyed using numbers).

## Unordered categorical variables

An example of an unordered categorical variable is region of the country. The default output from R is:

* Northeast 50
* Northwest 70
* Southwest 100
* Southeast 3
* Missing 2

All of the above considerations for binary variables apply, and you won't typically have a separate line for missing. This is one way to treat the missing values.

Region (n=2 missing)

* Northeast 13.5% (27/200)
* Northwest 35.0% (70/200)
* Southwest 50.0% (100/200)
* Southeast 1.5% (3/200)

This is another:

Region (n=200, plus 2 missing)

* Northeast 13.5% (n=27)
* Northwest 35.0% (n=70)
* Southwest 50.0% (n=100)
* Southeast 1.5% (n=3)

The main decisions pertain to (1) ordering; and (2) grouping.

Rows might be ordered alphabetically, as above.

Rows might be ordered by frequency, starting with the most common:

Region (2 missing)

* Southwest 50.0% (100/200)
* Northwest 35.0% (70/200)
* Northeast 13.5% (27/200)
* Southeast 1.5% (3/200)

If there is a natural grouping, then this might be used to order the rows. For example, grouping by latitude:

Region (2 missing)

* North 48.5% (97/200)
* Northwest 35.0% (70/200)
* Northeast 13.5% (27/200)
* South 51.5% (103/200)
* Southwest 50.0% (100/200)
* Southeast 1.5% (3/200)

Rows with small categories can be problematic. Sometimes, there are lots of such rows, which will clutter Table 1 with unimportant information. Sometimes, there are so few cases in a category that reporting it would allow the study subjects to be identified. For example, if there is a single Pacific Islander in the dataset then including that demographic group in Table 1 allows that person to be identified, which represents a violation of confidentiality.

Sometimes, the best that can be done is to combine small groups into an "other" category, recognizing that the more heterogenous this category, the less interpretable it will be.

Race and ethnicity have additional considerations pertaining to how categories are defined. These are discussed in the supplemental readings.

## Ordered categorical variables

As an example of an ordered categorical variable, disability after stroke is sometimes classified using the modified Rankin scale:

* 0=no symptoms at all
* 1=no disability despite symptoms
* 2=slight disability (unable to carry out all previous activities but able to look after own affairs without assistance)
* 3=moderate disability (requiring some help but able to walk without assistance)
* 4=moderately severe disability (unable to walk without assistance)
* 5=severe disability (bedridden, requiring constant nursing care)
* 6=dead

In Table 1, reporting of an ordered categorical variable should be organized by content rather than frequency, for example:

Disability (0 missing)

* Rankin 0 10% (10/100)
* Rankin 1 15% (15/100)
* Rankin 2 25% (25/100)
* Rankin 3 20% (20/100)
* Rankin 4 10% (10/100)
* Rankin 5 10% (10/100)
* Rankin 6 (died) 10% (10/100)

This illustrates ordering the results by increasing level of disability.

If it is one of the main analytical variables, a stacked bar graph provides essentially the same information as the empirical cumulative distribution function in visual form. For example, in the stacked bar graph Rankin categories 0-2 could also be annotated as "independent functioning", categories 4-5 could be annotated as "significant disability", etc., which would serve to convey information that is more effectively transmitted visually than with numbers.

## Continuous variables

Continuous variables have (1) central tendency; (2) spread; and (3) shape. The initial data exploration should consider all three. Moreover, for critical variables you should typically report a figure that addresses shape in addition to the other two constructs. Sometimes these figures are smoothed, in order to highlight essential features of the shape.

For Table 1, shape need not be explicitly reported. However, shape can affect how you choose to summarize central tendency and spread. For example, if the data are approximately bell-shaped, then central tendency and spread can be summarized with the mean and the standard deviation, respectively. This information can also be placed in the text. As an example (that also includes a confidence interval): *"Figure 1 describes the distribution of systolic blood pressure (SBP). It appears to be relatively bell-shaped with a mean of 135 and a standard deviation of 20. The 95% confidence interval for the mean SBP is 130-140."*

On the other hand, if a variable is highly skewed, reporting various quantiles would serve to summarize central tendency (e.g., through the median), spread (e.g., through the interquartile range), and would also provide clues about shape (e.g., if additional quantiles are reported).

Regardless of the above, also reporting the minimum and maximum values should be considered, in part because it provides a check on data integrity.

Sometimes, investigators ask that continuous variable be binned. For example, the continuously scaled systolic blood pressure might be reported as systolic hypertension, if it falls in the bin of SBP>=140 or not systolic hypertension if it falls into the bin of SBP<140. Although there are times when this is reasonable, as a general statistical principle transforming a continuous variable into a categorical one loses information, and so isn't recommended. In the SBP example, if binning this variable is clinically important, and so helpful to report within Table 1, then its continuous version should be reported as well.

## Structure of table 1

In designing table 1, you should consider (1) its columns; and (2) its rows.

Typically, one of the columns contains data from the entire sample. Often, there are additional columns that correspond to the categories of the primary predictor. For example, in a randomized trial Table 1 would also have separate columns for each of the study groups. If the primary predictor is continuous, it might be binned. Adding these columns allows a quick assessment of balance -- for example, the degree to which the study groups are similar on other variables that might influence the outcome. Whether p-values should be reported as part of the assessment of balance is debated, although most statisticians recommend against doing so.

Sometimes, the columns of Table 1 correspond to categories of the outcome variable. For example, if the outcome is binary there would be one column summarizing the patients with a "good" outcome and another column summarizing the patients with a "poor" outcome. Although this is sometimes a reasonable thing to do, in general this table structure is discouraged, because it isn't consistent with the flow of causation (i.e., from predictor to outcome).

The rows of Table 1 are often arranged by category, with white space between the categories. One possible arrangement is demographic variables, clinical variables, etc. Outcome variables are sometimes placed in Table 1, and sometimes they are placed in other tables, other figures, or in the text of the statistical report. The general presentation principle is to arrange every element of the statistical report, Table 1 included, in a way that is easy for your audience to follow.

# 1-sample inference

## Binary outcome

Reading: Altman, on 1-sided versus 2-sided tests

Suppose that the primary outcome is binary and that there are no other predictors. As an example, a study of a new cancer drug might classify patient outcomes into "complete response" versus "not a complete response". In practice, this would be the result of binning, as there are multiple ways that "not a complete response" can be achieved (e.g., one of which is a "partial response").

As another example, a study of the safety of a new drug could classify the outcome of each patient as "at least one serious adverse event" or "no serious adverse events".

The visualization step is to simply report a proportion. In particular, a figure (e.g., a histogram plot) isn't helpful, as it violates Vickers' advice about when to create figures. In other words, the figure presents relatively little information, and that information could have been successfully transmitted through text.

The statistical theory around proportions is covered in BIOS701. Briefly, a statistical analysis should match the study's data generating mechanism. In the oncology example, we assume that each patient's outcome is independent of the others, and that each patient has approximately the same chance of achieving a complete response. Having an identical chance of achieving a complete response might not be realistic -- "approximately" is often satisfactory. If, on the other hand, there is one variable that is related to the chance of a complete response, you should use a chi-square / logistic regression approach with one predictor, as covered in this course. If there are many variables that are related to the chance of achieving a complete response, you should use a logistic regression with multiple predictors, as covered in BIOS705.

Let X be the number of patients with a complete response out of n patients, and so the observed proportion is X/n. Based on the data generating mechanism (i.e., a set of Bernoulli trials), X is binomial with sample size n and an unknown probability of success π (with π identical for each patient). Because n is fixed, X/n is a simple transformation of a binomial random variable, and so the statistical theory around X/n can be based on the binomial distribution as well.

The binomial distribution (or, for large samples, its approximation by a normal distribution) can be used to create a confidence interval for π. For example, if n=70 and X=14, the observed proportion is 0.20 (this is also the maximum likelihood estimator), and a 2-sided 95% confidence interval is (0.11-0.31). This confidence interval can be directly obtained using R.

In the oncology example, in the absence of a control group the observed proportion of patients with complete response might be compared against historical "controls" -- in other words, the proportion of patients with a complete response from a similar population. As discussed in BIOS701, if it is important to discover whether the new drug performs better, worse or similarly to the historical comparator, then a 2-sided analysis is appropriate. Checking whether the 2-sided 95% confidence interval contains the value of π under the null hypothesis is equivalent to 2-sided statistical test with a type 1 error rate of 0.05. This is termed "inverting the hypothesis test".

In some cases, a 1-sided test is preferred to a 2-sided one, the primary reason being that it has greater power against a 1-sided alternative hypothesis. Please see the reading for additional details. The R code only requires a tiny modification.

The topic of statistical power is covered more thoroughly in BIOS701. Power is the probability that the results will be statistically significant at as specific value of the parameter under the alternative hypothesis. Briefly, the steps in performing a power calculation via simulation are:

* For a single iteration of the study, create a simulated dataset, using what you know about the study design, data generating mechanism, and a specific value of the parameter(s) under the alternative hypothesis.
* Perform a statistical test, and record whether or not the results are statistically significant (e.g., with p<0.05).
* Repeat the process for a large number of iterations of the study
* Record the proportion of studies: this is your estimate of statistical power.

As an illustration of the structure of a power calculation, assume that the true proportion is π=0.20, and that the proportion of patients with complete response in the historical comparison group is π=0.15, and that there will be n=100 patients in your dataset.

To estimate the power, generate a large number of replicates of the study -- for example, 1000. For each replicate, we will simulate the results for 100 distinct individuals, as illustrated in the table below.

|  |  |
| --- | --- |
| Individual | Result (1=complete response, 0=not) |
| 1 | 0 |
| 2 | 1 |
| 3 | 0 |
| … |  |
| … |  |
| 99 | 0 |
| 100 | 0 |

The reason that the results differ across individuals is that they are based on a pseudorandom number that differs from individual to individual.

Next, summarize the results across individuals. For example, suppose that the number of patients with complete response is 21, as illustrated in the table below.

|  |  |
| --- | --- |
| Iteration | Y = # with complete response |
| 1 | 21 |

Here, in fact, we didn't need to simulate data at an individual level, but could simply have used a binomial distribution with n=100 and π=0.20, but in general you'll need to simulate individuals and then aggregate their results.

Next, we perform a 1-sample test for proportions, obtaining a 2-sided p-value of 0.09, which is not statistically significant using the benchmark of p=0.05. This adds 2 columns to the table:

|  |  |  |  |
| --- | --- | --- | --- |
| Iteration | Y | p-value | Statistically significant (1=yes, 0=no) |
| 1 | 21 | 0.09 | 0 |

Finally, we repeat the process for 1,000 replications. Suppose that Y=23 in iteration 2 and Y=19 in iteration 1000. The first 2 rows of the resulting table are:

|  |  |  |  |
| --- | --- | --- | --- |
| Iteration | Y | p-value | Statistically significant |
| 1 | 21 | 0.09 | 0 |
| 2 | 23 | 0.03 | 1 |
| … | … | … | … |
| 1000 | 19 | 0.26 | 0 |

The simulation-based estimate of the power is the proportion of results that are statistically significant.

## Unordered categorical outcome

When the outcome variable is measured on an unordered categorical scale, its distribution is multinomial. Using a table to visualize the data has been previously illustrated (i.e., as per Table 1), and figures aren't generally recommended (i.e., they provide relatively little information per unit of space, and essentially no information beyond what can be transmitted with a table).

Although inference on a single binomial proportion is somewhat common, inference on a set of multinomial proportions is rare. Accordingly, we won't cover it here. A common approach uses a chi-square goodness of fit test, which will be introduced later in a more common context.

## Ordered categorical outcome

Reading: t-PA trial report with stacked bar graphs

When the outcome variable is an ordered categorical one, its distribution is multinomial. For visualization, a stacked bar graph (or CDF plot) is a reasonable option, because it contains more information than is the case for an unordered categorical variable.

Although inference on a single binomial proportion is somewhat common, inference on a set of multinomial proportions is rare. Accordingly, we won't cover it here. A common approach uses a chi-square goodness of fit test, to be covered later.

When designing an analysis strategy, ordered categorical variables are challenging, in large part because the choices are (1) use a specialized technique, which might be complex and unfamiliar to your audience; or (2) perform a less efficient analysis, either by (a) ignoring the ordering; or (b) collapsing the categories into 2 groups; or (3) introducing validity issues by treating the variable as continuous and assigning numerical scores to each category. We'll defer discussion of these options until later -- for now it's enough to recognize that the problem exists. This is an especially important consideration when your primary outcome variable is an ordered categorical variable with few categories.

## Continuous outcome

One-sample inference on continuous outcome variables is somewhat common, and is covered in greater detail in BIOS701. It typically pertains to central tendency. Briefly, consideration of the study question, the data generating mechanism, the results of the visualization should provide sufficient information to decide whether the focus should be on the mean or the median. If, for example, interest is in the mean and the normality assumption is appropriate then the observed mean is the maximum likelihood estimate, and you should report that mean along with a confidence interval. If the goal is to compare against an external value, then hypothesis testing proceeds similarly to 1-sample testing for a proportion.

# Models with 1 predictor and one outcome

## Introduction

We will consider two types of outcomes, according to their scale of measurement. These are continuous and categorical. Categorical outcomes are further divided into 2 categories (binary), >2 unordered categories, and >2 ordered categories. We will classify predictor variables similarly. In other words, our organizing framework is number of variables and their scale of measurement.

With modest exceptions, all models in BIOS702 will have a single predictor. The exceptions are a limited number of 2-predictor models. Among others, these 2-predictor models allow us to introduce the important concept of statistical interaction in the simplest possible context.

All of the methods presented here fall within the purview of the framework of general linear models, which will be introduced in BIOS705.

Although everything can be placed into a general framework, we will also cover some specialized tests that essentially do the same thing: for example, a t-test that compares two group means is equivalent to performing SLR where the predictor is coded as a 1 or 0 to denote study group. In practice, it is helpful to know that this is also a t-test, in order to read the medical literature, effectively communicate with investigators, etc.

## Continuous outcome and continuous predictor

### Introduction

For continuous outcomes, we start with the most general case: namely, a continuously scaled predictor. We will use this opportunity to introduce a number of statistical concepts -- for example, robustness -- that many applied data analysis courses put off until later.

### Visualization

The visualization exercise is a scatterplot. You should be assessing a number of things:

* What is the shape of the relationship between the predictor and the outcome?
* Is the relationship between the predictor and the outcome linear?
* Is the variability of Y similar for all values of X?
* Are there any outliers?
* Are there any unusual values of X with high leverage?

When assessing the shape of the relationship between X and Y, you should start with whatever science that the investigator is able to share with you. For example, should Y rise or fall with increasing values of X? Should the value of Y plateau? This information allows you to draw the shape that you anticipate, based on theory. Often, but not always, a first approximation to this shape is a straight line.

To visually assess whether the relationship between Y and X is linear, have R append the best-fitting line to the scatterplot. For the moment, we need not worry about how R selects this line. Do the datapoints consistently fall near the line? Do the discrepancies (i.e., vertical distance from Y to this line) follow any particular pattern? If not, then a SLR is plausible.

Although a scatterplot is sufficient for performing the above assessments, the same information is contained in residual plots, which will better generalize to the case of multiple predictors than will the scatterplot. The residuals are Yp-Y, where Yp is the predicted value from the best-fitting line. If the functional form of the regression model is correctly chosen, residuals (which are observed) will behave similarly to the error terms in the model (which aren't observed). One of the model's assumptions is that the variance of the error term is identical for all values of X. Accordingly, a plot of the residuals versus X should have a similar degree of scatter for all values of X.

For the purposes of hypothesis testing, the SLR model will assume that the distribution of the error terms is normal. The most straightforward way to assess this is through a Q-Q plot of the residuals. If the errors are approximately normal, their quantiles will be similar to the quantiles of a normal distribution (with the appropriate mean and standard deviation). Accordingly, a plot of the residuals versus these normal quantiles should be a 45-degree line through the origin. You can use R to generate Q-Q plots.

Regarding shape, the scatterplot of Y versus X should closely follow the best-fitting line that R appends. Equivalently, a plot of the residuals versus X should have no pattern (i.e., it should look like a line with a slope of 0 and an intercept of 0). The same can be said for a plot of the residuals versus Yp, which will more easily generalize to the case with multiple predictors (i.e., because Yp is a scalar but X has multiple dimensions).

As an example, if the regression function is actually quadratic rather than linear, the scatterplot should show this pattern, as should the above residual plot. We will discuss how to perform a more formal assessment of whether the regression function is linear later.

Outliers are data points whose values of Y are inconsistent with the fitted line. In other words, outliers have residuals with large absolute values. We will discuss what to do with outliers later -- if the outlier is sufficiently extreme an initial step is to check the raw data if possible. Outliers degrade model performance -- in other words, the model will fit better without outliers than with them. Sometimes, outliers point to unusual individuals or subpopulations that are sufficiently important to consider separately.

A leverage point is an unusual value of X. For example, as illustrated in an exercise, if most of the observations are on patients aged 20-39, except for a single 90-year old, the 90-year old is a point with high leverage. This data point will have disproportionate influence on the fitted line -- in essence, the fitted line will pass at or near through that data point regardless of the line that would have otherwise been fitted to the remainder of the data. Leverage points can be identified visually through the scatterplot, and also through leverage statistics that essentially quantify how far X is from the center of its distribution.

The analysis of residual plots relies upon an important subtlety. We previously mentioned that "*if the functional form of the regression model is correctly chosen, residuals (which are observed) will behave similarly to the error terms in the model (which aren't observed)*". To make the implications of this statement explicit, consider two possibilities, based upon whether or not the residual plots behave "as desired" (i.e., are unrelated to X, are unrelated to Yp, have similar levels of variability for all values of X, are approximately normally distributed). Assume that the sample size is sufficiently large that the observed patterns in the residual plots can be taken seriously.

* If the residuals behave as desired, then our postulated model (i.e., linear, with variances that are normal and independent of X) is correct.
* If the residuals don't behave as desired, then our postulated model isn't correct. The departure from the desired pattern points to which assumptions are violated, and how serious are those violations.

### Notation

Setting X0 to 1 for all observations in order to help estimate the intercept, the predicted portion of the regression model is

Yp = β0 X0 + β1 X1

Here, we have replaced X with X1, which will simplify the notation when we eventually add more predictors to the model. For SLR, we will use X and X1 interchangeably.

In R, the data array only requires X1 and Y. X0 is created automatically (and can be suppressed if necessary).

Simplifying the above equation:

Yp = β0 + β1 X1

Adding the error term:

Y = β0 + β1 X1 + ε, where ε is normally distributed with mean 0 and standard deviation σ.

These error terms are independent of one another, and the standard deviation is the same for all values of X.

Another way of stating this is:

Y = Yp + ε, where Yp is the prediction and ε is the error term,

If individuals are denoted by "i", a more detailed notation is possible:

Yi = β0 + β1 X1i + εi. This notation emphasizes that the value of X1 differs from individual to individual, and also that each individual has their own error term.

One implication is that, for an individual with a specific value of X1, their predicted value of Y will be normal with mean β0 + β1 X1i and standard deviation σ. Here, Xi1 varies from individual to individual, whereas β0, β1, and σ do not.

### Fitting the model

"Fitting the model" means finding the "best" values of the parameters β0 and β1. How do to so is covered in BIOS701. Two approaches are the method of least squares and the method of maximum likelihood. For SLR, their results are identical.

The results of the model-fitting process are summarized in the ANOVA table.

The ANOVA table is based on 3 quantities that are calculated for each observation: (1) (Y-Ym)2, which is summed to produce the total sum of squares (SST); (2) (Y-Yp)2, which is summed to produce the error sum of squares (SSE); and (3) (Yp-Ym)2, which is summed to produce the model sum of squares (SSM). Because SST=SSM+SSE, maximizing model fit (i.e., maximizing SSM) is equivalent to minimizing residual variance (i.e., minimizing SSE).

Although deriving the distribution of test statistics requires an assumption of normality, producing SST, SSM and SSE does not. Indeed, the method of least squares only requires minimizing SSE. The method of least squares was originally applied to astronomy data without assuming normality -- this latter assumption was made later.

Without assuming normality, because Yp = β0 + β1X, then (Y-Yp)2 = (Y - (β0 + β1X))2, where the subscript "i" on the Xi and Yi has been suppressed. Here, X and Y are fixed quantities, having been observed in the dataset, and β0 and β1 are the variables. We want to find the values of β0 and β1 that minimize Σ (Y - (β0 + β1X))2. That's a straightforward calculus problem whose solution can be easily found online, and that solution is the usual formula for β0 and β1.

Now consider the problem of finding the maximum likelihood estimator of the mean of a normal distribution with a specified value of the standard deviation σ. The probability density function of a single observation is

1/ σ (2π)0.5 exp { - (Yi-µ)2 /(2σ2) }.

The likelihood function is the product of terms with the above format. For example, if n=3, σ=2 and the observed data are Y1=0, Y2=4 and Y3=5, the first element in the product is 1/ 2 (2π)0.5 exp { - (0-µ)2 /(2\*4) }.

The second element in the product is 1/ 2 (2π)0.5 exp { - (4-µ)2 /(2\*4) }.

The third element in the product is 1/ 2 (2π)0.5 exp { - (5-µ)2 /(2\*4) }.

The method of maximum likelihood asks us to find the value of µ that maximizes the above product.

To perform this maximization, we can apply three tricks from 701. First, the constants, including σ, can be ignored, because anything that maximizes the product of exp { - (Yi-µ)2 } also maximizes the product of 1/ σ (2π)0.5 exp { - (Yi-µ)2 /(2σ2) }.

Second, we take the log of the above product, since anything that maximizes the log of a quantity also maximizes that quantity.

Finally, when X is positive, minimizing X is the same as maximizing 1/X. This idea is applied to the exponential component of the likelihood, where exp{-X} = 1 / exp{X}. This is the point where estimating the MLE, which is a "maximization problem", will become minimizing the SSE, which is a "minimization problem".

Accordingly, we want to minimize the sum (Y1-µ)2 + (Y2-µ)2 + (Y3-µ)2 -- in other words, (0-µ)2 + (4-µ)2 + (5-µ)2 as a function of µ. It turns out that the answer is to use the mean of the observations, and so the maximum likelihood estimator of µ is 3.

The details of the associated algebra can be found online. The crucial insight, for our purposes, is that the menu for creating maximum likelihood estimates from normally distributed data requires minimizing a sum of squares -- indeed the SSE. Moreover, this happens because the probability density function for normal random variables includes a squared term.

Returning to simple linear regression, once the assumptions of normality and equal variances are made, the observations Yi become normally distributed with mean

β0 + β1Xi and standard deviation σ. In the example dataset from the previous module, the general form of the function to be maximized is

Σ log [ 1/ σ (2π)0.5 exp { - (Yi- (β0 + β1Xi))2 /(2σ2) } ].

After working through some tedious algebra that can be found online, a key step in this algebra turns out to be minimizing the SSE -- essentially as above.

Nothing in this derivation depends on X being continuously-scaled -- in fact, it applies in equal force when X an indicator variable denoting group membership.

For the collaborative statistician, the key message isn't in the details of the above algebra, but instead lies in the insight that the normality assumption implies that finding maximum likelihood estimators in the SLR model is functionally equivalent to applying the method of least squares. Accordingly, all of the general statistical theory about maximum likelihood estimates that you learn in 701 also applies to SLR. Among others: the excellent statistical properties of maximum likelihood estimators apply to the regression coefficients β0 and β1, which are derived using the method of least squares. This principle will also apply to linear regression with multiple predictors.

### Matrix version of the calculations

Our goal is to help you develop intuition about how the matrix version of simple linear regression works, using a small, toy dataset. Although we use a continuously-scaled X for purposes of illustration, nothing in this section relies on X being continuous.

Background: This assumes that you can perform matrix multiplication. Please review this topic if necessary.

Our initial dataset is listed below. It contains the observed values of the predictor, X1, and the outcome, Y, for n=3 individuals. (The reason we have denoted the predictor by X1 rather than X should soon be clear.)

|  |  |
| --- | --- |
| X1 | Y |
| 1 | 0 |
| 2 | 4 |
| 3 | 5 |

Running statistical software (e.g., the lm() function in R) generates a fitted model:

Yp = -2 + 2.5\*X1. This model provides the "best fit" to the observed data. Much of this module focuses on deriving this fitted model directly, using matrix manipulations.

Plugging the values of X1 into the predictive equation (i.e., fitted model) generates predicted values, denoted by Yp.

|  |  |
| --- | --- |
| X1 | Yp |
| 1 | -2 + 2.5\*1 = 0.5 |
| 2 | -2 + 2.5\*2 = 3.0 |
| 3 | -2 + 2.5\*3 = 5.5 |

This predictive equation works regardless of the value of X1. In particular, X1 need not be included in the dataset -- plugging in X1=4 yields Yp = -2 + 2.5\*4 = 8.0.

Here, X1=4 is outside the range of the data used to fit the model, and so the resulting predicted value should be interpreted with some skepticism. Indeed, the farther X1 is from the range of the data used to fit the model, the greater should be the degree of skepticism. That X1=4 is outside the range of the data used to fit the model is obvious from the scatterplot. However, when we move from one to multiple predictors, more effort will be required to describe the range of the data to which the model directly applies.

We now add a column of 1's, and label it X0 -- in other words, X0=1 for everyone -- and combine X0 and X1 into the matrix X. X0 will be used to generate the intercept, β0, and X1 (i.e., the observed values of the predictor) will be used to generate the slope, β1. X is printed below.

X

|  |  |
| --- | --- |
| X0 | X1 |
| 1 | 1 |
| 1 | 2 |
| 1 | 3 |

The transpose of X, denoted by X`, rotates X by 90 degrees:

X`

|  |  |  |  |
| --- | --- | --- | --- |
| From X0 | 1 | 1 | 1 |
| From X1 | 1 | 2 | 3 |

We can use matrix multiplication to create X`X:

X`X

|  |  |
| --- | --- |
| 1\*1 + 1\*1 + 1\*1 = 3 | 1\*1 + 1\*2 + 1\*3 = 6 |
| 1\*1 + 2\*1 + 3\*1 = 6 | 1\*1 + 2\*2 + 3\*3 = 14 |

The inverse of X`X, denoted (X`X)-1, is the matrix with the property (X`X)-1 \* X`X = I, where I denotes the identity matrix. It turns out to be:

(X`X)-1

|  |  |
| --- | --- |
| 7/3 | -1 |
| -1 | 1/2 |

The identity matrix has ones on the main diagonal and zeros elsewhere. Multiplying a matrix by the identity matrix leaves it unchanged. (For a scalar, the "identity matrix" is 1, since multiplying any scalar number by 1 leaves it unchanged.)

To demonstrate that this is the correct value of the inverse, we multiply (X`X)-1 \* X`X and check the result.

|  |  |
| --- | --- |
| (7/3\*3) + (-1\*6) = 1 | (7/3\*6) + (-1\*14) = 0 |
| (-1\*3) + (1/2\*6) = 0 | (-1\*6) + (1/2\*14) = 1 |

Creating the inverse of a matrix is the conceptual equivalent of "dividing by" that matrix. For example, consider a scalar with a value of X=5. Its inverse is X-1=1/5, since 1/5\*5=1.

The other matrix in the formula for the regression coefficients is X`Y:

X'Y

|  |
| --- |
| (1\*0) + (1\*4) + (1\*5) = 9 |
| (1\*0) + (2\*4) + (3\*5) = 23 |

To better understand what X`X and X`Y represent, begin with X`X. Its diagonal elements are the sums of squares of the values of Xi for each individual. For example, X`X [1,1] = X[1,1]2 + X[2,1]2 + X[3,1]2. Here, for example, the notation X[1,2] refers to the element of X contained in row 1 and column 2. These squared terms are related to variances.

The off-diagonal elements of X`X are the cross-products of the vectors X0 and X1. For example, X`X [1,2] =X[1,1]X[2,1] + X[2,1]X[2,2] + X[3,1]X[3,2]. These cross-product terms are related to covariances (i.e., between X0 and X1).

The elements of X`Y are the cross-products between X and Y. For example, X'Y [1,1] = X[1,1]Y[1] + X[2,1]Y[2] + X[3,1]Y3. If Y increases as X increases, then the elements of X`Y will be positive. Similarly, if Y decreases as X increases, the elements of X`Y will be negative. These are also related to covariances (between the predictors and the outcome) and also correlations (between the predictors and the outcome).

The matrix formula for the vector of the 2 regression coefficients (i.e., β0 and β1), denoted by β, is β = (X`X)-1 \* X`Y. R should generate the same estimates as below.

β

|  |
| --- |
| (7/3\*9) + (-1\*23) = -2 |
| (-1\*9) + (1/2\*23) = 2.5 |

### Meaning of the estimated regression parameters

Consider 2 individuals whose values of X1 differ by 1 unit -- X1=21 and X1=20 for example.

Since Yp = β0 + β1X1, their predicted values are:

When X1=21: Yp = β0 + 21 β1

When X1=20: Yp = β0 + 20 β1

By subtraction, the difference between these predicted values is β1. (i.e., since β0 cancels). Accordingly, β1 denotes the predicted difference in outcome for individuals with a 1-unit difference in X1.

The above holds regardless of the value of X1 for the comparator -- all that matters is that the they differ by 1 unit. For example, comparing X1=41 with X1=40:

When X1=41: Yp = β0 + 41 β1

When X1=40: Yp = β0 + 40 β1

After cancellation, we are left with β1 as before.

The above interpretation compares 2 separate individuals. It doesn't necessarily mean that an intervention that changes the value of X1 in a particular individual will change their expected outcome. That would have to be demonstrated by experiment.

### Data transformations

Here, we explore by illustration, what happens if the values of X are transformed. For example, we consider what happens if we shift the values of X to be centered on their mean. The new X matrix is now:

X

|  |  |
| --- | --- |
| X0 | X1 |
| 1 | 0 |
| 1 | 1 |
| 1 | 2 |

Visually, this is essentially the same thing as physically shifting the graph of Y versus X to the left. The slope shouldn't change, but the intercept will. We'll work through the above matrix calculations, and verify that this is the case.

X`

|  |  |  |  |
| --- | --- | --- | --- |
| From X0 | 1 | 1 | 1 |
| From X1 | 0 | 1 | 2 |

X`X

|  |  |
| --- | --- |
| 3 | 0 |
| 0 | 2 |

(X`X)-1

|  |  |
| --- | --- |
| 1/3 | 0 |
| 0 | 1/2 |

X'Y

|  |
| --- |
| 9 |
| 5 |

β

|  |
| --- |
| 3 |
| 2.5 |

Applying the new predictive equation Yp = 3 + 2.5\*X\*, where X\* is the transformed value of X1, we obtain the same predicted values as before.

|  |  |  |
| --- | --- | --- |
| X\* | Y | Yp |
| -1 | 0 | 0.5 |
| 0 | 4 | 3.0 |
| 1 | 5 | 5.5 |

Now, we consider what happens if we also shift the values of Y to be centered on their mean. The new Y vector is now:

Y

|  |
| --- |
| Y |
| -3 |
| 1 |
| 2 |

This is essentially the same thing as physically shifting the graph of Y versus X down the page. As before, the slope shouldn't change, but the intercept will.

X`

|  |  |  |  |
| --- | --- | --- | --- |
| From X0 | 1 | 1 | 1 |
| From X1 | 0 | 1 | 2 |

X`X

|  |  |
| --- | --- |
| 3 | 0 |
| 0 | 2 |

(X`X)-1

|  |  |
| --- | --- |
| 1/3 | 0 |
| 0 | 1/2 |

X'Y

|  |
| --- |
| 0 |
| 5 |

β

|  |
| --- |
| 0 |
| 2.5 |

Returning to the original data, and applying the predictive equation Yp = 0 + 2.5X\*, we obtain the same predictive values, once we add 3 back to Yp.

|  |  |  |
| --- | --- | --- |
| X | Y | Yp |
| -1 | -3 | -2.5 |
| 0 | 1 | 0 |
| 1 | 2 | 2.5 |

In summary, the impact of data transformations is consistent with intuition.

### Extensions

The above examples, can be extended in two ways. First, as the sample size increases the structure of the critical matrices -- namely, X`X and X`Y, is unchanged.

For example: with n=3, X`X is generated by multiplying a 2x3 matrix by a 3x2 matrix, in order to produce an output matrix that is 2x2. With n=10, X`X is generated by multiplying a 2x10 matrix by a 10x2 matrix, in order to produce an output matrix that is also 2x2. The interpretation of its elements as sums of squares and cross-products remains the same.

Second, as additional predictor variables are added to the model the matrices in question become larger, but they retain the same essential structure. For example: with 4 predictors (including the intercept), X'X is generated by multiplying a 4xn matrix by a nx4 matrix, to produce an output matrix that is 4x4. The interpretation of its elements as sums of squares and cross-products remains the same. Ultimately, a matrix of 4 estimated regression coefficients is produced.

To illustrate (a bit heuristically) that the matrix approach is doing the same thing as the algebraic formula for a regression coefficient, let "i" denote individual observations.

Then, β1 = Σ (Xi-Xm) (Yi-Ym) / Σ (Xi-Xm)2.

Here, the numerator is essentially X`Y and the denominator is essentially X`X. Dividing by X`X is essentially the same as multiplying by its inverse.

### ANOVA table

Returning to the original data, and applying the predictive equation

Yp = β0 + β1X1, yields the following:

|  |  |  |
| --- | --- | --- |
| X1 | Y | Yp |
| 1 | 0 | 0.5 |
| 2 | 4 | 3.0 |
| 3 | 5 | 5.5 |

We can denote the mean of Y by Ym, and add a column containing it. The value of Ym is the same for everyone.

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | Y | Ym | Yp |
| 1 | 0 | 3 | 0.5 |
| 2 | 4 | 3 | 3.0 |
| 3 | 5 | 3 | 5.5 |

We now add the components of the various sums of squares included in the ANOVA table. These calculations are based on Y: more specifically, they are based on 3 quantities: (1) the observed Y; (2) the predicted Yp; and (3) the mean Ym. X1 is no longer needed -- its only role was to help derive the best-fitting line, and thus to help derive Yp. The bottom row was obtained by summing the elements in its associated column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y | Ym | Yp | (Y-Ym)2 | (Y-Yp)2 | (Yp-Ym)2 |
| 0 | 3 | 0.5 | (0-3)2 = 9 | (0-0.5)2 = 0.25 | (0.5-3)2 = 6.25 |
| 4 | 3 | 3.0 | (4-3)2 = 1 | (4-3)2 = 1 | (3-3)2 = 0 |
| 5 | 3 | 5.5 | (5-3)2 = 4 | (5-5.5)2= 0.25 | (5.5-3)2 = 6.25 |
|  |  |  | SST = 14 | SSE = 1.5 | SSM = 12.5 |

SST (total sum of squares) quantifies the total amount of variation in the data. The numerator of the formula for the variance is the SST. SST is an absolute measure, whereas the variance is a relative measure (i.e., it is scaled by the number of observations).

SSE (error sum of squares) quantifies the amount of variation unexplained by the model.

SSM (model sum of squares) quantifies the amount of variation explained by the model.

R-squared, the proportion of variation in the outcome explained by the model, equals SSM/SST, which also equals 1 - (SSE/SST). R-squared has a range of 0 to 1. For SLR, R-squared is also the square of the Pearson correlation coefficient between X and Y.

The ANOVA table summarizes the sums of squares described above and provides the inputs to assessing the predictive ability of X. These assessments include a p-value, generated by a F-test, and a measure of overall model fit, the R-squared statistic.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Degrees of freedom | Sum of squares | Mean square | F-statistic | p-value |
| Model | DFM = 1 | SSM | MSM = SSM/1 | MSM/MSE |  |
| Error | DFE = n-2 | SSE | MSE = SSE / (n-2) |  |  |
| Total | DFT = n-1 | SST |  |  |  |

The total sum of squares is the numerator of the variance, and represents the "total amount of variation in the outcome variable". Loosely speaking, the degrees of freedom (denoted by DF) represents the number of "moving parts". For a sample of size n, because the formula for the variance requires calculating the sample mean, that counts as one moving part and so the total degrees of freedom is n-1 instead of n. Once the sample mean is estimated, even though it appears as if the regression model has 2 moving parts (i.e., β0 and β1) only one of them matters (as illustrated in BIOS701), namely β1, so the degrees of freedom for the model is 1. The degrees of freedom for error is obtained by subtraction. This is a loose description, and you are welcome to find a more precise treatment elsewhere.

The degrees of freedom allocated to the model should equal the number of predictor variables, ignoring X0. According to the rules for creating indicator variables for categorical predictors, each categorical predictor contributes k-1 degrees of freedom, where k is the number of categories.

Sums of squares quantify how much of the variation Y is attributable to the model and how much is attributable to error. As such, they are absolute quantities. To turn them into relative quantities, divide by their respective degrees of freedom. The result is termed a mean square (denoted by MS).

As covered in BIOS701, the F-statistic is MSM/MSE. The p-value is obtained by comparing the result to a central F-distribution with parameters DFM and DFE.

The R-squared statistic is SSM/SST.

### Generalizations and special cases

Nothing here depends on X being continuously scaled (i.e., since the above discussion was "about Y, not about X").

Adding multiple predictor variables to the regression model has no impact on the structure of the ANOVA table. The reason is that the only role of the multiple predictors is to generate the predicted value of Yp, which is a scalar, at which point these predictors can be dropped. As in the case of SLR, the ANOVA table is based on the same 3 quantities: (1) the observed Y; (2) the predicted Yp; and (3) the mean Ym.

Statistical theory tells us that our model is "optimized to the data". In other words, it is the "best fitting regression line". Equivalently, SSR is a maximum and SSE is a minimum.

Now, we'll consider some special cases. When the slope of the regression line is 0, SSM=0 and SSE=SST. To demonstrate this, when β1=0, Yp will be β0 regardless of the value of X. Moreover, the best-fitting value of β0 is Ym -- in other words, the default prediction for Y is its mean, which should be intuitive.

When the model fits perfectly, SSE=0 and SSM=SST.

As the slope of the regression line increases, SSM increases and SSE decreases. When SSM increases, the R-squared statistic, SSM/SST, increases. The following statements are equivalent:

* X is a statistically significant predictor of Y
* The R-squared statistic is significantly greater than 0
* The amount of variation explained by the model, SSR, is statistically greater than 0.
* The regression coefficient β1 is statistically different from 0
* The model explains at least some of the variation in the outcome.

In SLR, these tests provide identical results:

* A t-test, using β1 divided by its standard error
* A chi-square test with 1 degree of freedom, using the square of the above z-statistic
* An F-test, using MSM/MSE, where MSM denotes the mean square for the model and MSE denotes the mean square error.

These distributional results are discussed in BIOS701. They are all ultimately derived from the distribution of weighted sums of normal random variables, and especially from weighted sums of squares of normal random variables, and so the above statistical tests require the additional model assumption of independent, identically distributed, normal errors.

For collaborative statisticians, the action items around this body of theory are:

* Recognize that the above tests are equivalent, which can help clarify communications about modeling results.
* Recognize that the above standard tests require that the assumptions of normality and equal variances -- when these assumptions are importantly violated other approaches should be considered.

### Nested models

Two models are nested if the predictors in the smaller model -- the "reduced model" -- are a proper subset of the predictors in the larger model -- the "full model".

Equivalently, the reduced model can be understood as having the same set of predictors as the full model, but with their parameter estimates all constrained to equal 0 -- since in that case they have no impact on the outcome.

A partial F-test compares the fit of full and reduced models. It takes the general form:

F = [(SSMf - SSMr) / (DFf-DFr)] / MSEf. The degrees of freedom for this F-test are (DFf-DFr) and DFEf.

The partial F-test is an application of a general rubric, applicable to essentially any set of nested models regardless of the scale of measurement of the outcome variable:

* Fit the full model, record a quantity that measures its fit
* Fit the reduced model, record a quantity that measures its fit
* Combine these 2 quantities to obtain something that quantifies how much better the full model fits than the reduced model (i.e., a test statistic)
* Compare this test statistic with its distribution under the null hypothesis

To apply this general rubric to testing whether X matters in SLR, replace SSMr with 0, DFr with 0, and obtain the usual F-test.

Nothing in this section depends on X being continuously scaled.

### Robustness and influence

Discovering that X is statistically significant doesn't necessarily imply that you should automatically conclude that it is important. Its regression coefficient might be small when compared with a benchmark (i.e., a benchmark for "clinical impact", "scientific impact", "public health impact", etc.) Moreover, this result might be "statistically unstable".

A technical term for statistical stability is "robustness". For example: "robustness to violations of linearity", "robustness to violations of equal variances across different values of X", etc. As another example: "robust to small changes in the dataset, especially around influential data points".

Robustness has a number of criteria, not all of which need to be explored in detail for every dataset. Some elements that contribute to the robustness of a SLR model include:

* Minimal impact of outliers -- i.e., values of Y that "fit poorly" (i.e., for which Y and Yp strongly differ)
* Minimal impact of high-leverage points -- i.e., extreme values of X
* Similar conclusions across multiple modeling approaches

The first two criteria essentially address the question of whether, if the dataset were a little bit different, how much would the conclusions change. Outliers and leverage points are sometimes placed into the category of "influential" data points.

Outliers can be assessed using studentized residuals (among others), which are calculated at the level of the individual observation. Once these studentized residuals are created, the dataset can be sorted by their decreasing absolute values, with the focus on the most extreme ones. A poorly fitting data point, in and of itself, isn't necessarily a reason to delete it from the dataset -- for example, doing so would overstate the fit of the model. However, it is worth checking, both to verify that the data were correctly recorded, and also to try to identify any characteristics that makes that individual different from the others. Later, the variables that define these characteristics are natural candidates for inclusion in a multi-predictor model.

Leverage can be assessed using leverage statistics such as Cook's distance, which are also calculated at the level of the individual observation. As for outliers, the dataset can be sorted by decreasing absolute value, with the largest values compared against a predetermined threshold. At the very least, the largest values should be visually assessed. For example, in a dataset where X denotes age, if all the values of X fall within 20-39 except for one individual with X=90, that individual would have high leverage. A reasonable strategy could be to change the research question to pertain to a target population aged 20-39 -- that is, because of the limitations of the dataset -- and drop the observation with X=90 even though it represents "valid data".

The above approaches are applied to individual data points. However, sometimes the problem isn't with a single data point but, instead, is with a small cluster of such points. An approach that helps identify this problem is to use bootstrapping to create similar datasets to the one currently being analyzed. If the results (e.g., the magnitudes of the regression coefficients, the R-squared statistic) are consistent across the bootstrapped samples then they are "robust to small changes in the dataset". Bootstrapping makes the notion of "small changes in the dataset" more general than exploring the impact of dropping individual data points, one at a time.

A detailed discussion of robustness to different modeling approaches will be deferred until BIOS705. To give a simple illustration of the notion, suppose that the scatterplot suggests that Y generally increases as X increases, but the relationship might not be linear, the equal variances assumption might not hold, overall fit is degraded by the presence of some extreme outliers, etc. If the investigator is willing to settle for the weaker scientific hypothesis that "as X increases, Y monotonically increases", then both X and Y can be transformed into ranks, and a SLR fit to these ranks. This "non-parametric" approach, which is functionally equivalent to calculating a rank-based correlation coefficient, wouldn't assume that the resulting error terms have a normal distribution.

### Different functional forms

Review of the scatterplot (or the residual plot) might suggest that a straight line is inadequate to fit the data, and another functional form should be considered. The general menu for proceeding is to fit nested models: the reduced model is the usual SLR, whereas the full model has a more complicated functional form that includes a linear term. Comparing these two models using a partial F-test provides a statistical test of whether the more complex model "fits better" than the SLR model. Here, "fits better" means "the full model explains more of the variation in the outcome than the reduced model, after considering the impact of chance". By construction, the full model will fit at least as well as the reduced model, and probably better in the sense of having a larger SSM -- the question is whether this improved fit is more than would be expected due to chance alone.

For example, if the scatterplot suggests a quadratic function might be appropriate, then create a new variable X2, where X2 = X12. (These variables are highly correlated, but that isn't a problem.) The two nested models to be compared are:

Full model: Yp = β0 + β1X1 + β2X2

Reduced model: Yp = β0 + β1X1

X1 is included in both models, in order to create the nesting. Equivalently: if the quadratic term isn't important, then the true value of β2 will be 0, the full and reduced models will generate identical predictions, and their fits will be identical.

The more complicated the model, the more likely that it will be overfit -- in other words, that the parameter estimates will be optimized to the random noise in the dataset rather than the actual signal -- and thus perform poorly going forward. Accordingly, in this example X2 should not only be statistically significant, but the model should have notably better fit (e.g., notably larger R-squared). Another way of describing this phenomenon is that, because more complicated models are likely to be overfit, simpler models will tend to perform better when fit to new datasets. Simpler models are also easier to interpret. For most problems, your analytical approach should favor simplicity.

In the above example, the alternative functional form to be considered was evident from the scatterplot. If not, a spline or similar flexible function can be fit. The idea is the same, and the only differences is that X2, X3, etc. have a different definition. As before, the full model should fit notably better than the reduced model before being seriously considered.

### Interpreting the results

To interpret the regression coefficients (i.e., β1 should always be interpreted, β0 might or might not be scientifically interesting), report the point estimate, the confidence interval, a p-value, and provide an assessment of clinical significance. For example, "β1 = 4.0 (95% C.I. = 2.5-5.5), p<0.01. All points in the confidence interval exceed the clinically important value of 2.0."

To illustrate the implications of the model, you might plug in interesting values of X, calculate Yp, and then ask the investigator to provide a clinical or scientific interpretation of Yp.

Overall model fit should be reported using an R-squared statistic. If a model is statistically significant but fits poorly then this latter point is also worth emphasizing.

Models should be considered to be potentially overfit unless there is a demonstration to the contrary. Options include performing a validation or, if this isn't realistic, including the lack of validation in the limitations section of your statistical report.

Figures can often assist in supporting your interpretation of the results.

### Observer agreement

Reading: Altman article on observer agreement

The purpose of an observer agreement study is to assess the quality of the study's instrumentation -- more specifically, how consistently Y can be measured. Accordingly, the design involves taking multiple measures of the same thing (often an outcome variable) and assessing them for consistency (broadly defined).

The multiple readings could be multiple applications of a measuring device -- in which case the more consistent (and the less variability) the better. Here, we distinguish between consistency, the focus here, and calibration against a known external standard. If the thing being measured is SBP for example, we don't know the "true" SBP, and all we can do is assess the consistency of multiple readings of our measuring device.

The multiple readings could also compare different measuring devices, different users of the device, application of the device at different time points (in which case the SBP is assumed to be stable over time).

A comprehensive treatment of this topic involves fitting a "mixed model", which deferred until BIOS705. The mixed model for observer agreement studies generalizes to cases with more than 2 raters, more than 2 readings, etc. Here, we assume that there is only 1 reading for each of 2 raters. We'll focus on visualization, and also on some simple summary measures that are straightforward for investigators to interpret. Later, we will discuss observer agreement studies for binary outcomes, and observer agreement for continuous outcomes can provide a comparison.

For any particular patient, denote the observed value from rater 1 as X and the observed value from rater 2 as Y. This notation doesn't imply that X is a predictor and Y is an outcome.

One visualization is a scatterplot of Y against X, with a 45-degree line through the origin superimposed. If all the data points fall exactly on the reference line then agreement is perfect.

If agreement is less than perfect, we can report the magnitude of agreement and also the pattern of disagreement. Fitting a SLR to the above scatterplot, an overall measure of agreement is the R-squared statistic associated with that line.

The pattern of agreement considers calibration and accuracy. Miscalibration is a consistent tendency for the data points to fall above (or below) the reference line. A simple measure of miscalibration is the proportion of points falling above the reference line. Accuracy pertains to the average vertical distance between X and Y. A simple measure of accuracy is the mean of the absolute value of Y-X. In terms of more general statistical concepts: calibration maps to bias and accuracy maps to precision.

Observer agreement studies also assess discrimination, which measures whether large (small) values of X are at least associated with large (small) values of Y. Discrimination is a relatively weak criterion, and can be assessed by comparing the slope of the SLR of Y on X with 0. A statistically significant non-zero slope implies that some degree of discrimination is present, not that the level of agreement is excellent.

### Validation

Model validation begins with the same data array as an observer agreement study. However, rather than having the results from 2 observers, we have the observed outcome Y and the predicted outcome Yp, where Yp is derived by plugging the values of X in the test dataset into the SLR model created in the training dataset. The analogy to an observer agreement study isn't perfect -- in an observer agreement study it is often the case that who is assigned to which observer is arbitrary, whereas in a model validation exercise we're asking "how close is the observed Y to the predicted Yp", not the other way around.

In a validation study, the original dataset is sometimes called the "training set" and the new dataset is called the "test set" (i.e., as in "we will use the test set to assess the performance of the model derived from the training set"). Because the test set is different from the training set, the performance of the model won't be exaggerated due to overfitting.

If the test set comes from a different population than the training set, then "external validation" is being performed. The more different are the populations, the greater the generalizability of the results and the stronger conclusions that can be drawn. If an external dataset is unavailable, then the available dataset can still be divided into training and test datasets. This results in an "internal validation" that assesses overfitting but not generalizability. As might be imagined, there are various ways to divide the data into training and test datasets, and model validation is covered in greater detail in BIOS705.

The initial visualization -- namely, the scatterplot of Y versus Yp -- is the same as for an observer agreement study. Moreover, the things you're checking are also the same, although we're more precise about naming them:

* Calibration asks whether, for example, Yp is consistently higher (or lower) than Y.
* Discrimination asks whether the model's predictions can at least distinguish between those with high and low values of Yp.
* Accuracy asks how close, in absolute value, Yp is to Y.

To assess calibration, calculate the proportion of observations for which Yp>Y, and compare with a null hypothesis value of 0.50. Alternatively, calculate the difference scores Y-Yp, and compare the mean difference score to 0.

To assess discrimination, the fitted line in the scatterplot should have a positive slope -- the larger the better. Since the ideal result has β1=1, you can formally test this hypothesis

To assess accuracy, create the absolute values of (Yp-Y) and generate summary statistics.

## Continuous outcome and one categorical predictor

### 2 categories

For a categorical predictor with 2 categories, the general approach is to apply SLR with an indicator variable as its predictor. A t-test does the same thing.

Visualization is accomplished by creating separate boxplots for the 2 groups and checking the assumptions of normality and equal variance. When the predictor variable is categorical the assumption of linearity doesn't apply.

The above description -- namely, plot the values of Y for the 2 groups and check for normality and equal variances -- is typical but can introduce a misconception going forward. What we actually do is (1) fit a regression model, where the best-fitting value of Yp for each individual turns out to be the mean of Y for their group; and (2) apply the criteria of normality and equal variance to the residuals Y-Yp. Assessing residuals in this fashion is what generalizes. In particular, there is no need to check whether the distribution of Y as a whole is normal, as this is not one of the assumptions of the model. In this special case the distribution of the residuals is identical to that of the Ys: it is merely shifted to be centered around 0.

As with SLR, normality can be best assessed with a Q-Q plot. Equality of variances can be assessed by direct comparison. Although a formal test for equality of variances exists, the results are robust to everything but large differences -- for example, standard deviations that differ by a factor of 2 or more. In other words, small but statistically significant differences in the variances don't usually matter. (Since the variance is simply the square of the standard deviation, we refer to them interchangeably.)

In practice, violation of the assumption of equal variances often happens because Y is on the wrong scale. For example, with laboratory data log-transforming Y often fixes the problem.

Label the groups as A or B, and also set X0=1 to create the intercept and set X1=1 if group=A and 0 if group=B. X1 is an "indicator variable" -- in other words, it points to (indicates) the study group to which the individual is assigned. The fitted model can be written as:

Yp = β0 X0 + β1 X1, and also as

Yp = β0 + β1 X1, because the value of X0 is always 1.

To obtain the predicted value for anyone in group A, set X1 to 1 and solve:

Yp = β0 + β1 \*1, and so Yp = β0 + β1

To obtain the predicted value for anyone in group B, set X1 to 0 and solve:

Yp = β0 + β1 \*0, and so Yp = β0

Accordingly, the meaning of the parameters is:

* The mean of group B, and also the best prediction for anyone in group B, is β0.
* The difference between the mean of group A and the mean of group B is β1. The best prediction for anyone in group A is β0 + β1.

Testing β1 = 0 is equivalent to testing the null hypothesis that the true (but unknown) group means are identical. When the two groups represent an intervention and a control, this is equivalent to testing the null hypothesis that the intervention doesn't work.

The formula for the t-test is identical to the formula for β1 in the SLR model (with X1 as defined above).

When defining the indicator variables, we only used information about the predictors, not the outcome. Accordingly, indicator variables are defined identically, regardless of the outcome variable's scale of measurement.

### More than 2 categories (1-way ANOVA)

When the single categorical predictor has k categories, with k>2, we move from the t-test to the analysis of variance (ANOVA). The t-test is a special case of the ANOVA, with k=2. We will assume that the categories in question are unordered.

As terminology, "ANOVA" implies that the predictor(s) in question are all categorical. Moreover, "1-way ANOVA" implies that there is only 1 predictor. These are all special cases of linear regression.

ANOVA induces some additional considerations. The technical consideration is how to define the indicator variables. The strategic consideration is that, while for a t-test there is only a single possible comparison -- namely, between groups A and B -- in the ANOVA there are multiple such comparisons. As the analyst, you will need to decide which comparisons to make and how to address the issue of "multiple comparisons" (discussed below).

Visualization is a straightforward extension of the 2-group case -- prepare separate boxplots (or similar summaries) separately for each group. Explore the assumption of approximate normality within each group. Then, assess whether the standard deviations are relatively similar -- for example, within a factor of 2.

To illustrate how to create indicator variables, assume that there are 3 groups: A, B and C. One way to define the indicator variables is:

* X0=1 for everyone (this is implicit)
* X1=1 for group A, 0 otherwise
* X2=1 for group B, 0 otherwise

The SLR model can be written as:

Yp = β0 + β1 X1 + β2 X2

Following the same procedure as before:

For group A: Yp = β0 + β1\*1 + β2\*0, and so Yp = β0 + β1.

For group B: Yp = β0 + β1\*0 + β2\*1, and so Yp = β0 + β2.

For group C: Yp = β0 + β1\*0+ β2\*0, and so Yp = β0.

Accordingly, the meaning of the parameters is:

* The mean of group C, and also the best prediction for anyone in group B, is β0.
* The difference between the mean of group A and the mean of group C is β1. The best prediction for anyone in group A is β0 + β1.
* The difference between the mean of group B and the mean of group C is β2. The best prediction for anyone in group A is β0 + β2.

It doesn't matter how the indicator variables are defined, so long as you can map the regression coefficients back to the underlying group means. In the above example, group C was the "reference group". In practice, if there is a natural group to which the others should be compared it should become the reference group.

Testing that both β1 = 0 and β2 = 0 is equivalent to testing the null hypothesis that the true (but unknown) group means are identical. This test has 2 degrees of freedom. Most statisticians recommend starting with a test of the "overall null hypothesis" (sometimes termed the "strong null hypothesis") that the true group means are identical to one another. Although there are cases when you should proceed otherwise, when this joint test is "negative" (i.e., not statistically significant), your default procedure should be to stop and not perform additional analyses.

Like any test in SLR, the above joint test compares two statistical models. The full model allows β0, β1 and β2 to be unconstrained. The reduced model sets β1 and β2 to equal 0, and only β0 is allowed to vary. In the reduced model, β0 is the overall mean of Y. In the full model, β0, β1 and β2 are derived from the mean values of Y within each of the groups. The ANOVA table is created in the usual way, and the degrees of freedom associated with the model is 2.

Although we won't follow up on all the implications of this idea, the joint test can be stated in a more general form using "contrasts":

0 \* β0 + 1 \* β1 + 0 \* β2 = 0, and

0 \* β0 + 0 \* β1 + 1 \* β2 = 0.

As should be apparent, contrasts are simply linear combinations of the parameters in the SLR model.

A joint test is usually followed up with tests of additional contrasts. In addition to the p-value, you should typically report a confidence interval and an interpretation of the magnitude of the effect (i.e., in comparison with the minimum clinically important difference, if available). For simplicity, we'll limit coverage to hypothesis testing.

Tests of prespecified hypotheses provide stronger evidence than tests that were generated after the fact. One reason is scientific -- prespecified hypotheses have clear scientific justification. The other reason is statistical -- considering multiple hypotheses, and especially considering hypotheses that are suggested by the data, risks drawing falsely positive conclusions.

The risk of falsely positive conclusions is especially strong when considering the "experiment-wise type 1 error rate" -- that is, the probability that at least one of the statistically significant results will be a false positive. The type 1 error rate increases (1) as the number of tests increase; and (2) when you focus on the relationships that appeared to be strongest in the data. Exceptionally strong observed effects usually reflect some truth and some good luck, and by focusing on the strongest apparent relationships you risk differentially selecting those that were driven by good luck.

There are multiple ways to perform formal statistical adjustment for multiple testing. The simplest, albeit not necessarily the most efficient, is the Bonferroni correction, which sets the threshold for statistical significance to be .05 divided by the number of tests. For example, with 5 tests, you would require a p-value of .05/5 = .01 to declare statistical significance. The confidence intervals associated with the relevant contrasts are also widened, so that the results of the hypothesis testing are consistent with the coverage of the confidence intervals.

In practice, you can always look up the details about how to perform adjustments for multiple testing. What is more important is that you get the hypotheses right (e.g., specified ahead of time, scientifically reasonable), that you specify the analysis plan ahead of time, and that this analysis plan appropriately considers the dangers associated with falsely positive conclusions.

As a simple illustration of how hypotheses might be prespecified, suppose that group A represents one intervention, group B represents another intervention, and group C represents placebo. Some relevant comparisons include:

* Any difference between the 3 groups
* Comparing the 2 interventions
* Comparing A with placebo
* Comparing B with placebo
* Comparing the average of A and B with placebo

The first comparison has 2 degrees of freedom, the others have 1.

We first define the indicator variables (recognizing that the placebo is the natural choice for the reference category) using the same coding scheme as before:

* X0=1 for everyone (this is implicit)
* X1=1 for group A, 0 otherwise
* X2=1 for group B, 0 otherwise

Then, we (1) translate the null hypotheses into relationships among the group means; and then (2) translate the relationships among the group means into parameter values. These can also be translated into contrasts.

The final step in translating the hypotheses from cell means into model parameters and contrasts is to recognize that:

µC = β0

µA = β0 + β1

µB = β0 + β2

|  |  |  |  |
| --- | --- | --- | --- |
| Hypothesis | Group means | Parameters | Contrasts |
| Any difference between the groups | µA = µB = µC | β1=0 and β2=0 | 1β0 + 0β1 + 0β2 = 0,  0β0 + 1β1 + 0β2 = 0 |
| Comparing the 2 interventions | µA = µB | β1-β2=0 | 0β0 + 1β1 + (-1)β2 = 0 |
| Comparing A with placebo | µA = µC | β1=0 | 0β0 + 1β1 + 0β2 = 0 |
| Comparing B with placebo | µB = µC | β2=0 | β0 + 0β1 + 1β2 = 0 |
| Comparing the average of A and B with placebo | (µA + µB) / 2 = µC | .5β1+.5β2=0 | 0β0 + (-.5)β1 + (-.5)β2 = 0 |

Within the ANOVA, the comparison between groups A and B will yield a different result that a t-test which directly compares those groups. The difference is that the t-test only uses the patients in groups A and B, whereas the ANOVA utilizes the patients in group C to estimate the mean square error. The numerator of the test statistic is the same, but the denominator is different.

## Continuous outcome and 2 predictors

### 2 categorical predictors

Although most of multi-predictor modeling is deferred until BIOS705, we briefly consider two special cases. In both cases, there is a predictor of primary interest, measured on a categorical scale. There is also a single predictor of secondary interest -- either categorical (leading to the 2-way analysis of variance) or continuous (leading to the analysis of covariance). In both cases, we are trying to "adjust" for the impact of the secondary predictor as we focus on the predictor that is primary.

Let Y denote a continuous outcome variable and X1 and X2 denote 2 predictors of interest. For example, X1 might denote study group (drug A or drug B) and X2 might denote gender (female or male). The table below describes the "imbalance" in the study: males tend to receive drug A and females tend to receive drug B.

Sample sizes

|  |  |  |
| --- | --- | --- |
|  | Drug A | Drug B |
| Females | 50 | 150 |
| Males | 150 | 50 |

As illustrated in the next table, drug is associated with a 10-unit difference in outcomes, regardless of gender. However, outcomes tend to be higher for females. We'd typically report a standard deviation in addition to the mean.

Group means

|  |  |  |
| --- | --- | --- |
|  | Drug A | Drug B |
| Females | 100 | 90 |
| Males | 70 | 60 |

A 1-predictor analysis for drug -- that is, ignoring gender -- will report weighted averages of the male and female groups:

* Drug A: 1/200 (50\*100 + 150\*70) = 77.5
* Drug B: 1/200 (150\*90 + 50\*60) = 82.5.

The 1-predictor analysis draws the wrong conclusion: namely, that drug B is associated with a 5-unit decrease in outcomes, when in fact it should be a 10-unit increase. The cause of this phenomenon (termed "Simpson's paradox") is the confounding between drug and gender.

Briefly, a variable (here: gender) "confounds" the relationship between two other variables (here: between drug and outcome) when (1) it is related to the predictor (here: drug); and (2) it is related to the outcome.

Instead of a 1-predictor assessment of the relationship between drug and outcome, what should instead be analyzed is the table below, where an additional column has been added, to report the proper estimate of the impact of drug for each value of gender.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Drug A | Drug B | Difference |
| Females | 100 | 90 | +10 |
| Males | 70 | 60 | +10 |

Here, the impact of the drug (i.e., outcomes are 10 units higher for drug A) is the same regardless of gender. Equivalently, there is no "interaction" between drug and gender, and a "main effects" model without an interaction term applies.

Briefly: an interaction occurs when the relationship between X and Y depends on the value of Z -- here: an interaction would be present if the relationship between drug and outcome depends on gender.

We could also add a row that illustrates the impact of gender:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Drug A | Drug B | Difference |
| Females | 100 | 90 | +10 |
| Males | 70 | 60 | +10 |
| Difference | +30 | +30 |  |

To translate these ideas into the context of a 2-way ANOVA.

Define the intercept in the usual way:

* X0=1 for everyone

Define the indicator variable for drug in the usual way:

* X1=1 for drug A
* X1=0 otherwise

Define the indicator variable for gender in the usual way:

* X2=1 for females
* X1=0 otherwise

Define the indicator variable corresponding to the interaction:

* X3=X1\*X2

The 2-way ANOVA model with interaction is

Yp = β0 X0 + β1 X1 + β2 X2 + β3 X3

We can derive interpretation of the model parameters in the usual way:

Drug = A, gender = female: Yp = β0 1 + β1 1+ β2 1 + β3 1

Drug = A, gender = male: Yp = β0 1 + β1 1+ β2 0 + β3 0

Drug = B, gender = female: Yp = β0 1 + β1 0+ β2 1 + β3 0

Drug = B, gender = male: Yp = β0 1 + β1 0+ β2 X2 + β3 0

Accordingly:

Drug = A, gender = female: Yp = β0 + β1 + β2 + β3

Drug = A, gender = male: Yp = β0 + β1

Drug = B, gender = female: Yp = β0 + β2

Drug = B, gender = male: Yp = β0

The best estimates of the above parameters are the 4 cell means, and so:

60 = the mean outcome for the reference group, drug = B, gender = male

30 = the increase in the mean outcome, moving from the reference group to drug = B, gender = female

10 = the increase in the mean outcome, moving from the reference group to drug = A, gender = male

0 = the additional adjustment, after applying the 30 and the 10, to generate the mean outcome for drug = A, gender = female.

Here, β0=60, β1=10, β2=30, and β3=0.

As illustrated, the impact of the drug is identical for females and males, there is no interaction, and β3=0. Testing β3=0 is equivalent to testing for an interaction, and the test is performed in the usual way. That is, we can fit full and reduced models and perform a partial F-test. Equivalently, since this test has a single degree of freedom, we can simply look at the Wald-test-base p-value in the R output, which uses the estimated value of β3 and its standard error.

Since the interaction test is not statistically significant, we can refit the model without X3. The same parameter estimates are obtained. Importantly, β1=10 correctly estimates the impact of drug, after accounting for the imbalance in gender. Indeed, fitting a model without interaction (termed a "main effects model") allows us to discuss the impact of drug separately from the impact of gender. We can also say that β2=30 correctly estimates the impact of gender, after accounting for the drug that the patient received. This is not essentially different from how the results from a 1-parameter model are presented.

On the other hand, the table below illustrates an interaction between drug and gender.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Drug A | Drug B | Difference |
| Females | 120 | 90 | +30 |
| Males | 70 | 60 | +10 |
| Difference | +50 | +30 |  |

Now, the impact of drug is that drug A is associated with a 30-unit increase in mean outcome for females, but only a 10-unit increase in outcomes in males. The parameter estimates become β0=60, β1=10, β2=30, and β3=20. If the difference between a 30-unit increase and a 10-unit increase is clinically important (i.e., in addition to being statistically significant), then the interaction term should be retained, and the interpretation of the results is more nuanced than is the case for a main effects model. In other words, instead of describing a single impact of drug, we should separately describe the impact of drug on males and on females. This will be more complicated for your audience to absorb, and a general preference for simplicity suggests that trivial interactions should be ignored, even if they are statistically significant.

The table below illustrates an interaction that can't be ignored: drug A leads to lower values of the outcome for females but higher values for males. The model parameters are β0=60, β1=10, β2=30, and β3=-20.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Drug A | Drug B | Difference |
| Females | 80 | 90 | -10 |
| Males | 70 | 60 | +10 |
| Difference | +10 | +30 |  |

### 1 categorical predictor and 1 continuous predictor

When the secondary predictor (sometimes called the "covariate") is continuous, the structure of the analysis plan is essentially the same as for a 2-way ANOVA, namely:

* Fit a model with interaction. If that interaction is statistically and clinically significant, report the results separately by subgroup.
* Otherwise, fit a main effects model. Interpret the parameter associated with the primary predictor as being adjusted for the covariate.

Visualization proceeds by creating a scatterplot of X2 versus Y, with different plotting symbols for the 2 groups, and appending best-fitting lines for each of the groups. Apart from the usual things to check for simple linear regression, you should visually assess whether or not the 2 regression lines appear to be approximately parallel, as assumed by the first ANCOVA model below.

Now, assume that X1 denotes drug and X2 denotes age. Each of the predictor variables is defined as for single-predictor models, and the interaction term is defined as the product of X1 and X2. Thus:

* X0=1 for everyone
* X1=1 for drug A, 0 for drug B
* X2=age (a continuous predictor)
* X3=X1\*X2.

The model is:

Yp = β0 + β1 X1 + β2 X2 + β3X3

Plugging in 0s and 1s in the usual way and simplifying,

For group A: Yp = β0 + β1 + β2 X2 + β3X2 = (β0 + β1) + (β2+β3)X2

For group B: Yp = β0 + β2X2

The slopes of these two regression lines differ by β3. When β3=0, the slopes of the two lines are identical and the simpler "main effects" ANCOVA model holds.

As for the 2-way ANOVA, in ANCOVA the presence of an interaction changes the target of inference -- that is, from a single parameter summarizing the constant impact of study group, to how the impact of study group differs by age. When the slopes don't differ by much even though they are statistically significant, we say that "the interaction is statistically significant but not clinically significant" and drop it from the model.

In the main effects model, interpretation focus on β1, which represents the constant impact of drug, regardless of age. However, in the model with interaction the impact of drug depends on age -- which in turn is an implication of the two regression lines having different slopes.

For example, when age=20, the difference in predicted values is β1 + (β2+β3) \* 20, whereas when age=30, difference in predicted values is β1 + (β2+β3) \* 30. The parameter β3 quantifies how much the impact of drug changes for 1-unit differences in age.

As for the main effects ANCOVA model, the statistical significance of the interaction term can be assessed in two equivalent ways:

* You can perform a t-test on β3 / se (β3).
* You can perform a partial F-test comparing the fit of a full model with X1, X2 and X3 against a reduced model with X1 and X2 only.

To verify that X3 represents an interaction, we can apply the general definition of an interaction is "the magnitude of the impact of X on Y depends on the value of Z". Here, the magnitude impact of X1 on Y, as quantified by the slope of its regression line, depends on the value of X2.

As with ANOVA, when the categorical predictor has more than 2 groups then you will need more than one indicator variable to define group (loosely speaking, to create "X1"). Because the interaction term is a product of "X1" and X2, you will need more than one indicator variable to define the interaction. The indicator variables representing the interaction should be tested as a group, not individually (a mistake commonly made by investigators).

If the interaction term can be dropped, we fit a main effects ANOVA model:

Yp = β0 + β1 X1 + β2 X2

Consider patients with identical values of X2 (i.e., of identical ages).

For group A:

Yp = β0 + β1\*1+ β2 X2 and so Yp = β0 + β1 + β2 X2.

For group B:

Yp = β0 + β1\*0+ β2 X2 and so Yp = β0 + β2 X2.

After cancellation, the difference between these two predicted values is β1.

Thus, β1 quantifies the impact of study group, for patients with identical ages. By "comparing likes with likes", the assessment of study group "accounts (controls) for possible imbalances in age". In that sense, ANCOVA "adjusts for differences in case mix". ANCOVA can be applied to observational studies to statistically adjust for differences in case mix that are typical of such studies. As discussed in BIOS703, the inferences in observational studies isn't as strong as that in randomized trials, because observational studies rely on you including all the potential confounders (i.e., variables that are strong predictors of outcome, and imbalanced between the study groups, whereas randomization works (on average) to balance all covariates, whether or not they are measured.

The formal statistical test of whether "X1 matters" can be performed in two equivalent ways.

* You can perform a t-test on β1 / se (β1), where "se" denotes its standard error.
* You can perform a partial F-test comparing the fit of a full model with X1 and X2 against a reduced model with X2 only.

## Binary outcome

### Introduction

For binary outcomes, we will first present the general modeling approach via logistic regression and then consider various chi-square tests as special cases. For simplicity of exposition, all the chi-square tests will be considered together, regardless of whether the outcome variable has 2 categories or more than 2.

We will allude to, although we won't fully develop, the framework of general linear models. Because the interpretation of the parameters is simpler for categorical predictors than for continuous ones, we will begin there.

### Transformations

Denote Y as 1 if an event occurs and 0 if it does not. We are typically interested in the probability that the event occurs, namely Pr{Y=1}. Denote this probability by π.

The most natural way to describe the probability that the event will occur is directly -- that is, by using π. However, it's equivalent to replace the probability π with the "odds", namely π / (1-π). This is an invertible 1-1 transformation: if you know the probability you can calculate the odds, and vice versa.

Indeed, you could take the logarithm (to base e) of the odds. The log-odds is called the "logit". Because taking the log is also an invertible transformation, the probability, the odds and the logit provide equivalent information.

The table below illustrates values of these transformations for various values of π.

|  |  |  |
| --- | --- | --- |
| π | Odds(π) | Logit(π) |
| .001 | .001 | -6.91 |
| .01 | .010 | -4.60 |
| .05 | .053 | -2.94 |
| .10 | .111 | -2.20 |
| .25 | .333 | -1.10 |
| .50 | 1.000 | 0 |
| .75 | 3.000 | 1.10 |
| .90 | 9.000 | 2.20 |
| .95 | 19.000 | 2.94 |
| .99 | 99.000 | 4.60 |
| .999 | 999.000 | 6.91 |

Apart from illustrating that the transformations in question are 1-1 and invertible, the table also shows:

* When the probability is small, it is approximated by the odds
* To transform from the probability of a success (i.e., that the event occurs) to the probability of a failure (i.e., that the event doesn't occur), use 1-π.
* When π=0.50, the odds are 1 (usually stated as 1:1). When π exceeds 0.50, the odds are greater than 1:1. When π is less than 50%, the odds are less than 1:1.
* The probability scale ranges from 0 to 1.
* To move from the odds of success to the odds of failure, take 1/odds.
* The odds scale ranges from 0 to infinity.
* When π=0.50, the logit is 0. When π exceeds 0.50, the logit is greater than 0. When π is less than 50%, the logit is less than 0.
* To move from the logit of success to the logit of failure, take -(logit).
* The logit scale allows any possible value.

Of course, you can get back from the odds and logit scales to the probability scale:

* Odds to probability: π = odds / (1 + odds)
* Logit to probability: π = log {odds / (1 + odds)}

All of this is mathematically true, but doesn't answer the question of why someone would use odds and/or logits when probabilities seem simpler. The short answer is that they form the basis of various statistical techniques -- those techniques are mathematically simpler and more natural when these alternative scales are used.

For example, as discussed in BIOS701, the basic paradigm underpinning Bayesian inference is "prior information, modified by the data, because posterior information" -- or, "prior odds times likelihood is proportional to posterior odds".

Logistic regression is linear on the logit scale. In other words, for a single predictor X

Logit (π) = β0 + β1 X1.

### Logistic regression with 1 binary predictor

With a single binary predictor and a single binary outcome, visualization is accomplished by creating a 2x2 table. By convention, the rows denote the predictor and the columns denote the outcome. Literally creating the table isn't necessary as it is usually sufficient to simply report the event rates for the 2 study groups.

Recalling that the definition of indicator variables is unrelated to the scale of measurement of the outcome variable, let:

* X0=1 for everyone
* X1=1 for group A and 0 for group B

Then: logit (π) = log (odds(π)) = β0 + β1 X1.

Taking exponents on both sides: odds(π) = exp {β0 + β1 X1}

For group A: odds(π) = exp {β0 + β1 \* 1} = exp {β0 + β1}

For group B: odds(π) = exp {β0 + β1 \* 0} = exp {β0}.

Applying the laws of exponents, for group A: odds(π) = exp {β0} \* exp {β1}

Accordingly, the odds ratio -- which quantifies the impact of study group -- is exp {β1}. A more complete description of this odds ratio is the "ratio of (1) the odds that the outcome will occur in group A; and (2) the odds that the outcome will occur in group B".

Here, β0 serves the same purpose as an intercept -- it is used to derive the probability that the event will occur (equivalently, the odds that the event will occur) for group B.

When exp {β1} = 1 then group has no impact. But exp {β1} = 1 when β1 = 0. Accordingly, the null value of β1 is 0.

As with SLR, the full model imposes no constraints of β0 and β1, whereas the reduced model constrains β1 to equal 0.

When β1=0, β0 is derived from the relative frequency that the event occurred in the overall sample -- that is, ignoring study group.

The overall structure of the inference is similar to that of SLR with a binary predictor (i.e., a t-test). However, the analogy isn't perfect. The reason is that logistic regression doesn't assume that the errors in Y are normal. (The only possible values of Y are 0 and 1.) Accordingly, the comparison between the full and reduced models doesn't use an F-distribution. The next section covers details of hypothesis testing.

For now, interpretation is based on the magnitude of the odds ratio quantifying the impact of study group -- that is exp {β1}. As before, what to report is a point estimate, confidence interval, p-value, and a comparison of the confidence interval with the benchmarks for clinical significance. How to create the confidence interval is discussed in the next section.

In practice, much of the biostatistician's task at this point is to help the investigator better understand what the odds ratio that we have calculated means. By observation, many investigators default to interpreting it as a risk ratio and, indeed, much of the biomedical literature discusses odds ratios as if they are risk ratios. However, odds ratios only approach risk ratios when the probability of the outcome is small (the reason is that the denominator in the odds is (1-π), which approaches 1 when π approaches 0 --- the same holds true for odds ratios). Often, what works is to provide practical grounding for those ratios by restating the observed event rates in the two groups.

Also, it should be emphasized that an odds ratio is a relative measure rather than an absolute one such as a risk difference. If the event rates in question are tiny, a large odds ratio might nevertheless have a trivial practical impact. As above, restating the event rates can help.

### Technical details

Logistic regression is most naturally described using the logit scale:

Logit (π) = log (odds(π)) = β0 + β1 X1.

For example, it is linear on the logit scale, but not on the original probability scale (indeed, the plot of Y versus continuous X on the probability scale will have a sigmoid shape).

With sufficiently large sample sizes, on the logit scale the parameter estimates are approximately normally distributed. Symmetric 95% confidence intervals for β1 are:

β1 +- 1.96 s.e.( β1)

These confidence intervals are symmetric on the logit scale, but not on other scales. To create a confidence interval for the odds ratio, take the above confidence interval and then exponentiate its endpoints.

Logistic regression doesn't make the distinction between the normal distribution and the t-distribution as does SLR. There are two reasons: (1) logistic regression is based on large-sample assumptions, and so the notion that the z-distribution is a large-sample approximation to the t-distribution doesn't apply; and, even more importantly (2) both the t- and z-distributions are ultimately derived from the assumption of normally-distributed errors, which doesn't apply to logistic regression.

There are "exact" versions of logistic regression that can be applied when more precise inference is needed for small samples. These aren't covered here.

In SLR, nested models (full and reduced) are compared using F-tests. In logistic regression, this is accomplished using chi-square tests. The appropriate degrees of freedom for the chi-square test is the difference in degrees of freedom in the full and reduced models.

In SLR, when assessing whether a model parameter with one degree of freedom differs from 0, the partial F-test comparing full and reduced models gives identical results to a t-test based on the regression coefficient and its standard error which, it turns out, is also equivalent to a score test. (In other words, the likelihood ratio test, the Wald test and the score test are identical.) In logistic regression these tests give different results, although they are asymptotically equivalent. When the results are substantively different you should check for problems.

As covered in BIOS701, the likelihood function to be maximized is based on binomial likelihoods. In other words, we need to maximize the product of the πi for the individuals with the event and the (1-πi) for individuals without the event, with πi derived from the probability that the event will occur in the logistic regression model for individual "i".

Except in simple cases, logistic regression models don't have closed-form solutions for the parameter estimates. Instead, iterative methods are used. You usually don't have to worry about this, although problems sometimes arise. For example, with a single binary predictor, if one of the groups has a 0% or 100% event rate then the algorithm doesn't converge due to "complete separation". As an example of an algorithm that converges but does poorly, a continuous predictor could have all the events associated with X >= 10 and all the non-events associated with X <= 10. The slight overlap between the population with and without events allows for parameters to be estimated, but they will have huge standard errors, which should alert you to the problem. When working with multiple predictors essentially the same thing can happen, but it is harder to spot due to the difficulty in visualizing multidimensional spaces.

To assess overall model fit, various measures can be considered. One is a generalization of the R-squared statistic from SLR. Perhaps the most common, though, is the c-statistic. To calculate it, take all pairs of observations where one member of the pair has the event and the other does not. Calculate Yp for both, and check whether the results are in the anticipated direction -- that is, with a higher predicted probability of response for the individual that had the event. The c-statistic is the proportion of pairs whose outcomes are in the anticipated direction. Values of .70 are often considered adequate with .75 considered good.

### Logistic regression with 1 continuous predictor

When the predictor is continuously scaled, the main differences between logistic regression and SLR follow from the fact that in logistic regression the only possible values of the outcome are 0 or 1. This has multiple implications, and the workaround in most cases involves binning.

For example, to engage in visualization a scatterplot of Y versus X won't be informative, because even though X is measured on a continuous scale Y is not. However, similar values of X can be binned. For example, consider the 10 similar values of age below and their outcomes:

|  |  |
| --- | --- |
| X=age | Y |
| 39.0 | 0 |
| 39.2 | 0 |
| 39.4 | 1 |
| 39.6 | 0 |
| 39.8 | 1 |
| 40.2 | 0 |
| 40.4 | 0 |
| 40.6 | 0 |
| 40.8 | 1 |
| 41.0 | 0 |
|  |  |
| Mean X = 40.0 | Mean Y = 0.30 |

We can take the means of both X and Y -- they are 40.0 and 0.30, respectively. These values can then become the input into a "binned scatterplot". Indeed, because the logistic regression model is linear on the logit scale, the point to be plotted is not (40,0.30), but instead is (40, logit(0.30)).

The same binning idea can be applied when assessing the fit of a logistic regression model. For each individual in the dataset, the logistic regression model generates a predicted value Yp, which is constrained to fall between 0 and 1. The table below illustrates a similarly structured calculation, now with X being replaced with Yp.

|  |  |
| --- | --- |
| Yp = predicted value of π | Y |
| 0.20 | 0 |
| 0.24 | 0 |
| 0.25 | 1 |
| 0.25 | 0 |
| 0.29 | 1 |
| 0.31 | 0 |
| 0.33 | 0 |
| 0.37 | 0 |
| 0.38 | 1 |
| 0.40 | 0 |
| 0.40 |  |
| Mean Yp = 0.314 | Mean Y = 0.30 |

In this example, within this bin the observed outcome rate of 0.30 was close to the predicted outcome rate of 0.314, which is encouraging.

The same data structure is used for model validation, where the predicted event probability is derived using the parameters of an externally-generated model.

An example of the data structure that is eventually generated is provided below. A formal test for adequacy of fit is provided by the Hosmer-Lemeshow chi-square test described below. In addition to this test, you should produce a narrative summary that describes where predicted and observed values are similar or dissimilar, whether small predicted outcome probabilities also have small observed outcome probabilities (i.e., discrimination), and whether observed and predicted probabilities are close in absolute value (i.e., calibration). A figure that summarizes the contents of this "calibration table" can be an effective way to summarize the results.

|  |  |  |
| --- | --- | --- |
| Bin | Expected number of events | Observed number of events |
| 1 (lowest predicted probability of outcome) | 1.3 | 0 |
| 2 | 2.5 | 1 |
| 3 | 5.4 | 4 |
| 4 | 7.9 | 8 |
| 5 | 10.0 | 12 |
| 6 | 11.3 | 14 |
| 7 | 12.7 | 13 |
| 8 | 16.8 | 16 |
| 9 | 20.3 | 19 |
| 10 | 25.6 | 18 |

The Hosmer-Lemeshow chi-square test takes the sum of

(observed-expected)2 / expected, and compares the results to a chi-square distribution with degrees of freedom equal to the number of rows minus 2. The expected number of events are generated through a binning process as above. Bin size is a tradeoff between wanting a large number of bins, but with reasonably large sample sizes within each bin.

Another difference between logistic regression and SLR is that, while in SLR the slope of the regression of Y on X has intuitive meaning for investigators, the slope of the regression of logit(Y) on X does not. Instead of trying to explain this, it is usually more effective to take selected values of X and report the predicted probably of the outcome for those values.

### Logistic regression with 2 categorical predictors

In a later module we will consider chi-squared tests for 2 categorical predictors and a binary outcome. There, one predictor is primary -- for example, drug A versus drug B -- and the other is secondary. For example, in the analysis of a randomized trial we might ask whether the efficacy of the intervention is consistent across patient subgroup -- for example, whether it is consistent for females and males, in which case the secondary predictor is gender. For another example, in a meta-analysis we might analyze multiple randomized trials of drug A versus drug B -- in that case, the secondary predictor is study. In a meta-analysis, we are not only interested in determining how consistent the efficacy of drug A is across studies, but also in providing a summary estimate of its efficacy overall. This secondary predictor is sometimes termed a "stratification factor" or "stratification variable".

Here, we illustrate that the above analyses can be executed within the framework of logistic regression. The results will be conceptually similar and asymptotically equivalent to the more specialized chi-square approach often found in the biomedical literature. For simplicity, we will assume that there are only 2 strata.

The visualization can be summarized by a pair of 2x2 tables:

Stratum 1: odds ratio = 2.25

|  |  |  |
| --- | --- | --- |
|  | Y=1 | Y=0 |
| Drug A | 10 | 90 |
| Drug B | 20 | 80 |

Stratum 2: odds ratio = 2.33

|  |  |  |
| --- | --- | --- |
|  | Y=1 | Y=0 |
| Drug A | 30 | 70 |
| Drug B | 50 | 50 |

Here, the overall event rates differ by stratum but the odds ratios are similar.

No differently from the case of continuously scaled outcomes, we can create indicator variables.

Define the intercept in the usual way:

* X0=1 for everyone

Define the indicator variable for drug in the usual way:

* X1=1 for drug A
* X1=0 otherwise

Define the indicator variable for stratum in the usual way:

* X2=1 for stratum A
* X1=0 otherwise

Define the indicator variable corresponding to the interaction in the usual way:

* X3=X1\*X2

The logistic regression model with interaction is:

Logit(π) = β0 X0 + β1 X1 + β2 X2 + β3 X3

The "main effects" logistic regression model without interaction is:

Logit(π) = β0 X0 + β1 X1 + β2 X2

This is exactly the same as 2-way ANOVA models, with Yp replaced by logit(π).

The parameter estimates from the logistic regression model with interaction are:

|  |  |  |
| --- | --- | --- |
|  | Estimate | Standard error |
| Intercept | -1.1077 | 0.1278 |
| Stratum | -0.6841 | 0.1278 |
| Drug | -0.4146 | 0.1278 |
| Stratum\*Drug | 0.0091 | 0.1278 |

Plugging the parameter estimates into the formula for the predicted probabilities:

|  |  |  |
| --- | --- | --- |
| Stratum | Drug | Predicted probability |
| 1 | A | 0.10 |
| 1 | B | 0.20 |
| 2 | A | 0.30 |
| 2 | B | 0.50 |

Because the model is "saturated" -- in other words, there are 4 outcome probabilities, and the model has 4 degrees of freedom -- the model perfectly predicts the observed outcome frequencies. As for the 2-way ANOVA, the role of the interaction term is to "adjust" the predicted value in the final cell to match the observed data.

The interaction term is not statistically significant (p=0.94). Accordingly, we refit the model without the interaction term. In general, it is best not to interpret main effects in the presence of interactions.

We obtain:

|  |  |  |
| --- | --- | --- |
|  | Estimate | Standard error |
| Intercept | -1.1091 | 0.1263 |
| Stratum | -0.6860 | 0.1249 |
| Drug | -0.4176 | 0.1207 |

Plugging the parameter estimates into the formula for the predicted probabilities:

|  |  |  |
| --- | --- | --- |
| Stratum | Drug | Predicted probability |
| 1 | A | 0.09682 |
| 1 | B | 0.20140 |
| 2 | A | 0.30139 |
| 2 | B | 0.49861 |

The predicted probabilities have changed subtly, because the model forces the predicted probabilities for the 2 strata to have the same odds ratios. This odds ratio is 2.31 (1.44-3.70, p<0.001).

In general, model results such as predicted probabilities and regression coefficients should be reported with fewer significant figures. (In other words: don't simply cut and paste from R output with lots of significant figures!) We did so here in order to emphasize the subtle nature of the differences between the two models.

## Chi-square tests

### Introduction

When a categorical outcome variable has more than 2 categories its distribution is multinomial (i.e., and thus an extension of the Bernoulli / binomial distributions for 2 categories). Predictive modeling for unordered categorical outcomes is unusual, and won't be considered here.

Because it is relatively common in the biomedical literature, we discuss chi-square testing. By analogy to 2-way ANOVA, we also describe how chi-square tests for a single categorical predictor variable can be extended to cover 2 categorical predictors, where this second categorical predictor is a stratification variable. Finally, we describe some chi-square tests that don't distinguish between a predictor and an outcome.

We also illustrate that chi-square tests are a special case of the logistic regression model.

### Chi-square test for a binary outcome and a binary predictor

Consider a 2x2 table cross-classifying a binary predictor X with a binary outcome Y. Y=1 when the event occurs and Y=0 when it doesn't occur. X=1 for group A and X=0 for group B.

|  |  |  |
| --- | --- | --- |
|  | Y=1 | Y=0 |
| X=1 | A | B |
| X=0 | C | D |

We can add the marginal totals:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y=1 | Y=0 |  |
| X=1 | A | B | A+B |
| X=0 | C | D | C+D |
|  | A+C | B+D | A+B+C+D = N |

The observed event rate for group A is A/(A+B). This is the maximum likelihood estimator for the true (but unknown) event rate in group A, denoted by πA.

The observed event rate for group B is C/(C+D). This is the maximum likelihood estimator for the true (but unknown) event rate in group B, denoted by πB.

The null hypothesis is typically that πA=πB -- in other words, that the true event rates are identical and any observed differences can be attributed to chance. For example, if group A is an intervention group and group B is a comparator, the null hypothesis asserts that the intervention doesn't work (and so their outcomes should be identical).

The chi-square test proceeds by (1) determining what values of A, B, C and D can be expected under the assumption that the null hypothesis is true; then (2) comparing the observed values with what was expected. This comparison uses the "distance" between the observed and expected data, where this distance is 0 when the observed and expected data are identical, and is positive otherwise. As covered in BIOS701, under the null hypothesis, one possible distance measure is a "test statistic" that sums, over the 4 cells,

(observed-expected)2 / expected.

To calculate what should be "observed" in each cell under the null hypothesis, apply the overall event rate to the total number of patients in the group. For example, for the cell (X=1,Y=1), this observed value is (A+C)/N \* (A+B).

The test statistic has a chi-square distribution with 1 degree of freedom. Small values provide evidence in support of the null hypothesis, large values of the test statistic provide evidence against it.

Returning to the above 2x2 table, as covered in BIOS701, with large samples a normality assumption can be applied to the estimated event probabilities. Sometimes, 0.5 is subtracted from the numerator of the chi-square statistic to achieve a closer approximation to normality. Usually, all of the versions of the chi-square test provide similar results -- although it is important to specify which one you plan to use ahead of time.

As covered in BIOS703, depending on the context you can select various effect measures to summarize the strength of the relationship between X and Y. Denoting the estimated event rates for groups A and B by π1 and π2, respectively:

* The risk ratio is π1 / π2.
* The odds ratio is [π1/(1-π1)] / [π2/(1-π2)]
* The risk difference is π1-π2.

The event rate, odds and log(odds) (called the "logit") provide equivalent information, as previously illustrated.

Visualization proceeds by creating the above 2x2 table. Analysis proceeds by reporting the effect measure, along with a confidence interval and p-value for a test of the null hypothesis. Interpretation is based on a substantive assessment of the size of the effect measure -- for example, relative to a benchmark for the minimal clinically important difference.

### Chi-square test for unordered categorical variables

The same menu can be applied to larger tables. In other words:

* Create "expected" counts by multiplying the marginal totals -- these represent what would be expected under the null hypothesis.
* Calculate a test statistic by summing (observed-expected)2 / expected over all the cells.
* Compare this test statistic with a chi-square distribution having (R-1)(C-1) degrees of freedom, where R denotes the number of rows and C denotes the number of columns.

This menu assumes that the expected number of counts within the cells are reasonably large. If not, exact methods (which aren't described here) can be considered.

### Chi-square test for stratified 2x2 tables

Suppose that our primary interest is in the relationship between study group (intervention, control) versus a binary outcome. The data might come from a randomized trial, and the primary visualization is through a 2x2 table. Among many other choices, the effect measure might be an odds ratio.

The primary analysis generates a single odds ratio that estimates the average impact of the intervention. The investigators are often interested in how consistent this impact is across subgroup -- for example, is the odds ratio similar for males and females, similar for those under 65 and those who are 65 and over, etc. Each of these subgroups is a "stratum", and to visualize the data we form 2x2 tables for each stratum separately.

Stratum 1: Females

|  |  |  |
| --- | --- | --- |
|  | Event | No event |
| Intervention | 30 (30%) | 70 |
| Control | 20 (20%) | 80 |

Stratum 2: Males

|  |  |  |
| --- | --- | --- |
|  | Event | No event |
| Intervention | 50 (50%) | 50 |
| Control | 40 (40%) | 60 |

We want to know:

* What are the odds ratios within each stratum?
* How similar are these odds ratios across strata?
* What is the overall odds ratio, after accounting for stratum?

A Cochrane-Mantel-Haentzel chi-square analysis, illustrated below, provides the answers.

For stratum 1, the odds ratio is 1.71 (0.89-3.29), p=0.10.

For stratum 2, the odds ratio is 1.50 (0.86-2.63), p=0.15.

The intervention isn't statistically significant within either stratum. However, the odds ratios appear similar (Breslow-Day test for interaction p=0.76).

Accordingly, a summary statistic can be calculated:

The summary odds ratio is 1.59 (1.04-2.43), p=0.03.

This confidence interval is tighter than the stratum-specific estimates, and has become statistically significant.

The CMH approach is also applicable to meta-analysis, which is a way of combining results across studies having similar designs. In that case, each study is a stratum.

Especially within the context of meta-analysis, the CMH approach to calculating an overall odds ratio is sometimes described as a weighted average of stratum-specific odds ratios (sometimes after undergoing a transformation into log(odds) ratios. We won't follow up on this idea.

## Other topics

### Ordinal categorical outcomes

It's unusual to encounter an ordered categorical outcome. Accordingly, we'll describe the most common statistical technique only in outline. These are:

* A chi-square test that ignores ordinality
* Translating the ordinal outcome into continuous scores
* Performing ordinal logistic regression

As an example, the table below summarizes the results of a randomized trial of a new stroke drug, using the Modified Rankin Scale as the outcome. Low values of the scale represent better outcomes. By eye, it appears as if the distribution of outcomes is shifted, at least somewhat, in favor of the intervention.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Rankin | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Utility | 1.00 | 0.95 | 0.85 | 0.50 | 0.30 | 0.05 | 0.00 |
| I | 5 | 10 | 20 | 20 | 20 | 15 | 10 |
| C | 5 | 5 | 10 | 25 | 25 | 15 | 15 |

One approach is to perform a general chi-square test with 6 degrees of freedom. The test statistic is 7.11, and the p-value is 0.31. This test is less powerful than desired, because it doesn't use any information about the ordinality of the outcome.

Another approach is to assign scores to each of the Rankin categories, treat the result as continuously scaled, and apply a t-test. For example, if we used integer scores of 0-6, the mean for the intervention group is 3.65, the mean for the control group is 3.25, and the p-value for the t-test is 0.09. This is a more powerful test that takes the ordinality of the outcomes into account, but has 2 conceptual difficulties: (1) the outcome variable only has a few categories, and isn't truly continuous (although this fact is often overlooked); and (2) the scaling is arbitrary.

Instead of using 0-6, another scoring option is to replace the Rankin values with a quality of life measure. One option is a "utility" score, where patients rank the quality of life associated with a Rankin category from 0 (worst) to 1 (best). Here, a Rankin score of 5 is considered to be barely better than death. With this scoring system, the mean for the intervention group is 0.52, the mean for the control group is 0.35, and the p-value for the t-test is <0.01.

Unless the choice of scoring system is obvious, you should assess the robustness of the conclusions to changes in the scoring. Here, the direction of the results is consistent, although their statistical significance is not.

Another approach utilizes the ordinality of the outcome variable but without assigning arbitrary scores. The idea is to make successive cut-points, thus generating a set of 2x2 tables:

|  |  |  |
| --- | --- | --- |
|  | Rankin 0 | Rankin 1-6 |
| I | 5 | 95 |
| C | 5 | 95 |

|  |  |  |
| --- | --- | --- |
|  | Rankin 0-1 | Rankin 2-6 |
| I | 15 | 85 |
| C | 10 | 90 |

|  |  |  |
| --- | --- | --- |
|  | Rankin 0-2 | Rankin 3-6 |
| I | 35 | 65 |
| C | 20 | 80 |

|  |  |  |
| --- | --- | --- |
|  | Rankin 0-3 | Rankin 4-6 |
| I | 55 | 45 |
| C | 45 | 55 |

|  |  |  |
| --- | --- | --- |
|  | Rankin 0-4 | Rankin 5-6 |
| I | 75 | 25 |
| C | 70 | 30 |

|  |  |  |
| --- | --- | --- |
|  | Rankin 0-5 | Rankin 6 |
| I | 90 | 10 |
| C | 85 | 15 |

These tables can form the strata in a stratified CMH analysis as in the previous section, especially if the odds ratios are relatively similar across strata.

For stratum 1, the odds ratio is 1.00 (0.89-3.29).

For stratum 2, the odds ratio is 1.59 (0.68-3.73).

For stratum 3, the odds ratio is 2.15 (1.14-4.08).

For stratum 4, the odds ratio is 1.49 (0.85-2.61).

For stratum 5, the odds ratio is 1.29 (0.69-2.40).

For stratum 6, the odds ratio is 1.59 (0.68-3.73).

These odds ratios are similar: Breslow-Day test for interaction chi-square statistic = 1.85, with 5 degrees of freedom, p=0.87.

The summary odds ratio is 1.55 (1.16-2.08), chi-square statistic = 8.70, with 1 degree of freedom, p<.01.

The key assumption is that the impact of the intervention is relatively constant, regardless of how it is dichotomized. This approach utilizes the ordinality of the outcome, but without imposing any specific scoring system. However, the same patients appear in more than one stratum, thus artificially inflating the sample size.

Ordinal logistic regression essentially creates the same strata as the CMH approach, and relies on the same assumption that the odds ratios are consistent across strata. It processes this information similarly albeit not identically. Instead of a single intercept coefficient and a single slope coefficient (which is translated into an odds ratio by exponentiation), it produces multiple intercept coefficients (here, 6 such coefficients, corresponding to the 6 possible 2x2 tables) and a single slope coefficient. After exponentiating this slope coefficient, the summary odds ratio is 1.57, with confidence interval 0.96-2.57, p=0.07. The odds ratio is virtually identical to that generated by the CMH analysis, and the confidence interval is wider, because it recognizes that the strata in question aren't independent.

### Diagnostic testing

Diagnostic testing studies compare the results of a diagnostic test (the predictor) with the actual disease state (the outcome). Assuming that the diagnostic test either generates a binary prediction (i.e., disease presence predicted, disease absence predicted), or can do so after applying a threshold value (e.g., all continuous test values above a threshold are mapped into a prediction of disease), the results can be summarized as a 2x2 table:

|  |  |  |
| --- | --- | --- |
|  | Disease present | Disease absent |
| Prediction: disease | A | C |
| Prediction: no disease | B | D |

As covered in BIOS703, a number of measures can be reported, each of which can generate a confidence interval in addition to a point estimate. These include:

Sensitivity: A / (A+B)

Specificity: D / (C+D)

Prevalence: (A+B) / (A+B+C+D)

Positive predictive value: A / (A+C)

Negative predictive value: D / (B+D).

Here, the actual diagnosis is the "gold standard", whereas the prediction is the "test standard". In other words, we want to see how well the test standard predicts the gold standard. Accordingly, we can denote the test standard prediction as X and the gold standard prediction as Y. The gold standard need not be absolute: for example, X might denote the diagnosis of a novice physician and Y denote the diagnosis of an experienced specialist (i.e., which is probably correct, but not certainly so).

As another variation on this design, suppose that the data can be summarized as a continuous variable X. This might be based on a machine readout (e.g., SBP=145) or a statistical model (e.g., predicted probability of outcome = 0.21). For simplicity of exposition, we will focus on the latter. We can dichotomize the results: for example, "predicted probability of outcome > 0.10 generates a prediction that the disease is present, otherwise predict that the disease is absent" -- the resulting data structure is the same.

The cut-point for determining when the predicted probability of the event is "high enough" to declare that the prediction is "the event will occur" rather than "the event won't occur" can be varied. Different cut-points generate different 2x2 tables, and thus different estimates of sensitivity and specificity. The result is a "receiver operating characteristics" (ROC) curve. Typically, the ROC is plotted as sensitivity versus 1-specificity. The "optimal" cut-point, which is data-dependent and thus at risk for overfitting, is the one that is closest to the upper-left corner of the plot. (More specifically, fit 45-degree tangent lines to the ROC curve. The optimal cut-point is the one whether the curve intersects the left-most tangent line.)

The area under the ROC curve equals the c-statistic, one measure of the fit of your logistic regression model.

### 2x2 measures of observer agreement

In observer agreement studies with a binary outcome, 2 observers are asked to provide a diagnosis of disease (i.e., present or absent). If one observer is an expert and the other is not, then the previous framework for diagnostic testing should be used. However, if both observers are equally trained, the distinction between predictor and outcome should be dropped, and the 2x2 table is now as below:

|  |  |  |
| --- | --- | --- |
|  | Rater B: disease present | Rater B: disease absent |
| Rater A: disease present | A | B |
| Rater B: disease absent | C | D |

The goal is to summarize the magnitude of agreement.

Simple agreement is (A+D) / (A+B+C+D).

Second opinion for a positive rating is 2A / (2A+B+C)

Second opinion for a negative rating is 2D / (B+C+2D)

The kappa statistic is (observed agreement - expected agreement) / (1-expected agreement), where expected agreement is calculated in the same way as the expected cell counts in a chi-square test.

The kappa statistic is sometimes terms "chance-adjusted agreement", because it appreciates that some level of absolute agreement is expected, even if the raters are guessing (and, so, their ratings are uncorrelated). This interpretation has been criticized and, in practice, what investigators actually want to know pertains to how likely a second opinion is likely to change the diagnosis. We won't cover kappa here, beyond just giving you a heads-up that it might be something you encounter in practice.

These techniques can be extended to more than 2 observers and more than 2 categories, but these extensions aren't covered here.