The theme of the today's lab is data analysis. This refers to using statistical tools to gain a better understanding of data collected in an experiment. In particular, you will use least-squares linear regression and analysis of variance (ANOVA) to estimate parameters in linear models. You will also use a combination of linear regression and ANOVA called ANCOVA (analysis of covariance). You will also learn to interpret the results of these analyses in proportional reduction of error measures of goodness of fit.

We will not be discussing statistical *inference* here – determining to what degree the results of data analysis give us information about a larger population or theoretical distribution. That is a very important subject, but requires more mathematics than we have available to understand well. You'll learn more about it in the *Research Methods* course in Cognitive Science, and much more about it if you take the *Probability* and *Mathematical Statistics* courses in the Mathematics Department.

Any serious data analysis requires computing. We'll be using *Excel* today because it's easy to create *Excel* workbooks that help you learn a statistical concept and because *Excel* is widely available. You should be aware that data analysis for scientific purposes is ordinarily performed using a dedicated statistics programming language like *R* or a dedicated statistics package like *SPSS*. MATLAB has a *Statistics Toolbox* which is excellent – we'll be using that in lab next week

Least Squares Regression Models

We'll be considering two models that we could fit to numerical data using *least-squares regression* – a constant model and a linear model.

Constant Model

This model predicts that, in the absence of random noise ε, the numerical response variable *y* will always have the same constant value *b*:

$$y_k = b + \varepsilon_k,\ k = 1, 2, \ldots, N.$$

Linear Model

The constant model has no *explanatory variables* – that is, the variation in the response variable *y* was all assumed to be due to random noise. The simplest model that has a numerical explanatory variable is a linear model:

$$y_k = mx_k + b + \varepsilon_k,\ k = 1, 2, \ldots, N.$$

For both models, the quantities ε_k are called the **residuals** and represent samples from a *random noise distribution*. The number of observed (explanatory value, response value) pairs is *N*.

***As in all modeling, we need to decide how to measure error*.**

A common approach is to use **SSE – the sum of squared errors**:

$$SSE(\text{model parameters}) = \sum_{k=1}^{N} \left(y_k - \hat{y}_k \right)^2 .$$

Choosing parameter values that *minimize* SSE is called **least-squares regression**.

Open the Excel file *Cog Sci 242 F18 Class14 – Linear Regression and Correlation.xls*.

The first two sheets in this workbook, "Constant Model" and "Simple Linear Regression," have cells near the top of the sheet that are outlined in with a heavy bold outline into which you can type different values of the parameters for the model the sheet is concerned with.

1. Take some time now to try different parameter values for both models in order to gather evidence for the following facts:[1]

The sample mean $\bar{y}$ is the value of b that minimizes SSE in the model

$$y_k = b + \varepsilon_k, \; k = 1, 2, \ldots, N,$$

and

the "optimal slope" and "optimal intercept" values minimize SSE in the model

$$y_k = mx_k + b + \varepsilon_k, \; k = 1, 2, \ldots, N.$$

Questions and Notes

[1] These facts can be established mathematically. See the Appendix to this lab for proofs, if you are interested.

Measuring Linear Association Between Two Numerical Variables

Now turn to the "Scatterplots and Correlation" sheet in the *Linear Regression and Correlation* file. Examine the **scatterplots** there.

A measure of the **linear association** observed between the two variables x and y based on the observations (x_k, y_k), for $k = 1, 2, \ldots, N$, is the

Pearson's sample correlation coefficient, $r = \dfrac{1}{N-1}\sum_{k=1}^{N} z_{x_k} z_{y_k}$,

where $z_{x_k} = \dfrac{x_k - \bar{x}}{s_x}$ and $z_{y_k} = \dfrac{y_k - \bar{y}}{s_y}$ are the **z-scores** for the x_k and y_k values, respectively.

Note that the z-score for the observed value of a variable first replaces that value with its displacement from the sample mean, then expresses that displacement in units of the sample standard deviation. Hence a z-score is read as a "number of standard deviations above or below the mean."

A *geometric interpretation* of r is as the *slope* of the least-squares regression line *fit to the z-score coordinates* (z_{x_k}, z_{y_k}), $k = 1, 2, \ldots, N$, rather than to the original data.

However, that property of r does not give us much insight into <u>how well</u> the regression line fits the data.

2. Suppose the sample correlation coefficient between data sets A and B is $r_{AB} = 0.4$,

while the sample correlation coefficient between data sets A and C is $r_{AC} = 0.8$.

Is the positive association between A and B *stronger* than between A and C ? Why?

Is it "twice as strong"? What do you mean by this?

A Dart Game (illustrating the concept of proportional reduction of error)

Suppose I'm playing darts with a friend.

My first throw lands 4 inches from the center. My second throw lands 3 inches from the center – which is better!

3. *Proportionally*, how much has my second throw reduced my error?

My friend's first throw lands 5 inches from the center, but their second throw lands only 2.5 inches from the center.

4. *Proportionally*, how much has my friend's second throw reduced their error?

5. Can we say that my friend improved *twice as much* as me on the second throw? In what sense?

Error Criteria, Prediction Methods, and Proportional Reduction of Error

In a very influential paper[2] published in 1965, Herbert Costner argued for using measures of association tbat could be interpreted as a **proportional reduction in error**.

Suppose you have an **error criterion** for evaluating a method of predicting observed values of a variable *y*. An example is the

Sum of Squared Errors, SSE:

$$SSE(\text{ prediction method}) = \sum_{k=1}^{N} \left(y_k - predicted(y_k)\right)^2.$$

[2] Hebert L. Costner, "Criteria for Measures of Association," *American Sociological Review*, 30 (June 1965), pp. 341-353.

Suppose we were comparing two such methods, *Method 1* and *Method 2*, and we found that *Method 2* did a better job.

A good way of expressing *how much* better *Method 2* was than *Method 1* would be the

PRE = **proportional reduction of error** = $\dfrac{\text{Error(} \textit{Method 1} \text{) – Error(} \textit{Method 2} \text{)}}{\text{Error(} \textit{Method 1} \text{)}}$.

Remarkably, the **coefficient of determination**, r^2, is a proportional reduction of error.

Proof of this statement would require calculus and some algebra. However, on both the "Scatterplots and Correlation" and "Proportional Reduction of Error" sheets in the *Linear Regression and Correlation* Excel file you can compare the values of r^2 and of PRE, calculated according to their different definitions, and see that they have the same value!

For least squares regression we calculate PRE as follows:

Our error criterion is the sum of squared errors, SSE.

Method 1 for predicting the values $y_1, y_2, \ldots, y_n$ assumes we only have these values.

Using only this information, if we want to make SSE as small as possible we must use the sample mean, $\bar{y}$, as our predictor for y_k, for each $k = 1, 2, \ldots, N$.

Method 2 assumes we have observed (x_k, y_k), for $k = 1, 2, \ldots, N$.

Making the best use of the information that knowing x provides, one can show that SSE will be as small as possible if we use the least-squares regression line:

$\hat{y}_k = mx_k + b$ to predict y_k for each $k = 1, 2, \ldots, N$.

Method 2 , using *regression*, is better than *Method 1*, using $\bar{y}$. How much better?

In the sense of proportional reduction of error, this is given by

$$PRE = \frac{SSE(\bar{y}) - SSE(regression)}{SSE(\bar{y})} = r^2,$$ in the case of linear regression.

I'm not suggesting that you use this formula to calculate r^2.

(In general you would use software that calculates r to find r, then square that.)

Rather, this formula shows us how r^2 can be *interpreted*. In words:

The coefficient of determination is the proportional reduction of SSE when the least-squares regression line, rather than the sample mean $\bar{y}$, is used to predict the observed values of y.

Measuring Association Between an Independent Binary Categorical Variable and a Dependent Numerical Variable

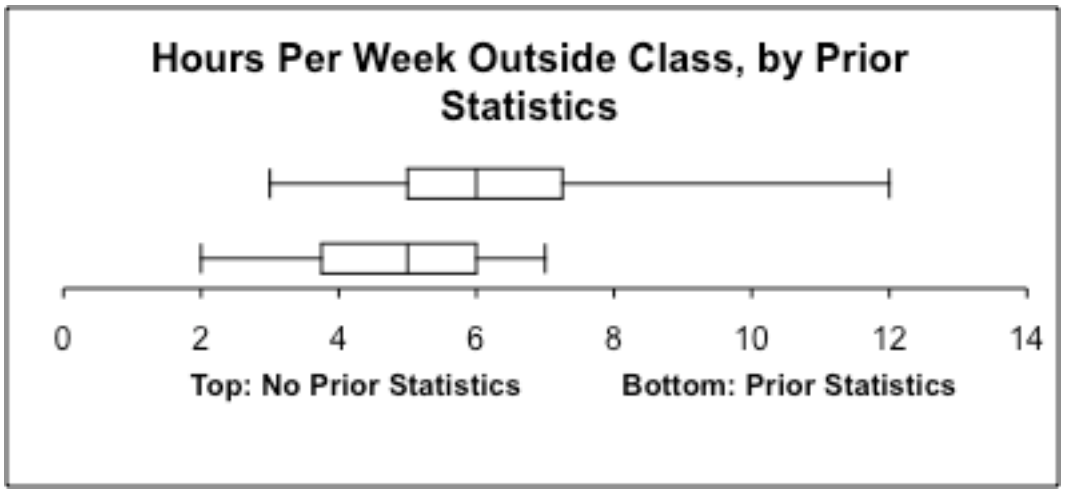
What are good statistical graphics and numerical measures for comparing two variables when one variable is binary categorical and one is numerical?

A good *graphical* tool is a comparative boxplot. (Use modified boxplots if outliers are present.) Separate the observations into two groups, based on the response to x, then compare the boxplots for each group. See the "Comparative Boxplots" sheet in the Excel file *Cog Sci 242 F18 Lab 6 - Correlation and ANOVA.xls.*

Example

These variables are from the survey given in statistics course. The categorical variable x and the numerical variable y are defined below:

$x = 0$ if the respondent had not taken a statistics course before;
$x = 1$ if the respondent had taken a statistics course before; and
y = estimated time spent on this course outside of class each week.



What about a *numerical* measure for how well prior statistics experience accounts for hours spent per week outside of class?

Remarkably, the coefficient of determination, r^2, can *still* be interpreted as a proportional reduction in error with a change of the second prediction method.

Specifically:

Our error criterion will be still be the error sum of squares, *SSE*.

Method 1 will still use for $\bar{y}$, as our predictor for y_k, for each $k = 1, 2, \ldots, N$.

Method 2 still assumes we have observed (x_k, y_k), for $k = 1, 2, \ldots, N$.

Now, however, the only possible values for x_k are 0 or 1, indicating which category we are observing.

We'll *re-index* the *y* observations to show which *value* of *x* they were observed with, effectively dividing them into two groups:

the observations $y_{0,j}$, j = 1, 2, …, N_0, were all observed together with $x = 0$,

&

the observations $y_{1,j}$, j = 1, 2, …, N_1, were all observed together with $x = 1$.

We'll predict that

y_k will have the value $\bar{y}_0$, the sample mean of the $y_{0,j}$, if $x_k = 0$, and

y_k will have the value $\bar{y}_1$, the sample mean of the $y_{1,j}$, if $x_k = 1$.

This prediction method is called Analysis of Variance (ANOVA), and one can show that it makes SSE as small as possible when the values of x_k are known.

Method 2 , using *ANOVA*, is better than *Method 1*, using $\bar{y}$. How much better?

In the sense of proportional reduction of error, this is given by

$$PRE = \frac{SSE(\bar{y}) - SSE(ANOVA)}{SSE(\bar{y})} = r^2$$ in the case of ANOVA as well.

(By the way, as a consequence of this you can use a least-squares regression program to calculate r^2 for ANOVA provided you code the x categories numerically, such as by 0 and 1.)

For the case of an independent binary categorical *x* and a dependent numerical *y*, the coefficient of determination r^2 is the proportional reduction in SSE when the ANOVA predictor, rather than the overall sample mean $\bar{y}$, is used to predict the observed values of *y*.

See the "Binary-Categorical" sheet in the Excel file *Cog Sci 242 F18 Correlation and ANOVA.xls* which shows that $r^2 = 0.1075$ using ANOVA to predict "Hours per Week Outside Class" by "Prior Statistics." This shows that "Prior Statistics" isn't a very good predictor of the hours per week students were spending on the course outside of class.

Check Your Understanding

Answer these questions concerning what we've learned so far about data analysis.

Correlation

Pearson's correlation coefficient *r* takes values in the range $-1 \le r \le 1$.

1. What would be true about a scatterplot if $r = 1$ for that scatterplot?

2. What would be true about a scatterplot if $r = 0$ for that scatterplot?

3. What would be true about a scatterplot if $r = -1$ for that scatterplot?

4. Suppose A, B, and C are three variables, and that $r_{AB} = 2r_{AC}$.

Does this mean that variable A is twice as useful in predicting the value of variable B as it is in predicting the value of variable C?

Suppose $r^2_{AB} = 2r^2_{AC}$.

Does this mean that variable A is twice as useful in predicting the value of variable B as it is in predicting the value of variable C?

Linear Regression

5. Suppose we are considering a random variable *y* as our response (dependent) variable.

Write the equation modeling *y* as a constant *b* plus a random error term.

Then write the equation for $\hat{y}$, the value of *y* predicted by this model when the constant *b* has been estimated by least-squares regression. (What is this constant?)

Now write the formula for the error-sum-of-squares *SSE*($\bar{y}$) for this model:

6. Suppose we are considering a random variable *y* as our output (dependent) variable, and a non-random numerical variable *x* as our input (independent) variable.

Write the equation modeling *y* as a linear function of *x* plus a random error term.

Then write the equation for $\hat{y}$, the value of *y* predicted by this model *when the slope and intercept parameters have been estimated by least-squares regression*. (Denote the estimated slope by $\hat{m}$ and the estimated intercept by $\hat{b}$.)

Now write the formula for the error-sum-of-squares *SSE*($\hat{m}x + \hat{b}$) for this model:

Analysis of Variance

7. Suppose we are considering a random variable *y* as our output (dependent) variable, and a non-random binary categorical variable *x* as our input (independent) variable. Assume the two categories are coded "0" and "1."

Write the equation modeling *y* as a function of *x* plus a random error term.

(It will be best to write this equation in a piecewise fashion.)

Write the equation for $\hat{y}$, the value of *y* predicted by this model when it value for each category has been estimated by least-squares. What is the predicted value for each category?

(Again, it will be best to write this equation in a piecewise fashion.)

Now write the formula for the error-sum-of-squares *SSE*(ANOVA) for this model:

8. What are the meanings of the following formulae, what are they called, and how could they be computed if you knew the value of the correlation coefficient r?

$$\frac{SSE(\bar{y}) - SSE(\hat{m}x + \hat{b})}{SSE(\bar{y})}$$

$$\frac{SSE(\bar{y}) - SSE(ANOVA)}{SSE(\bar{y})}$$

Note: For this lab, each of you will be analyzing your own individual search data that you collected in *Lab 5* using the CogsLab program. The purpose of this lab is to gain an understanding of the kind of information that different analyses give you. Next week we will use MATLAB to conduct these analyses more efficiently.

Get Your Data Into an Excel Workbook

Find the *html* file with your personal visual search data collected in *Lab 3* using *CogLab.* Paste the **Trial-by-Trial data** from the *html* file into the *Excel* workbook *Cog Sci 242 F18 Lab6 –Regression–ANOVA–ANCOVA.* Be sure to align with the column headings. Save the workbook to your G:\MATLAB directory.

You'll be replacing my own data from this experiment.

In the Excel workbook, **highlight the entire table** (but **not** the "outlier" analysis to the right of the data table)**,** then sort it by choosing

Data → *Sort*

… Sort by *Distractor Type*,
then by *Target*,
then by *Number of Distractors*,
then by *Response Time*.

I will be coaching you from the front of the room in how to use the *Excel* workbooks to complete this data analysis. Pay close attention so that you learn to use these files correctly.

More About Data Analysis

You may find that some of your observations are quite different from the others collected under the same experimental conditions. An unusual value like this is called an **outlier**.

The *Full Data* sheet in the workbook you are using today includes an *automatic* identification of outliers. They are using this definition:

> An observation is an **outlier** if it is
> either less than $Q1 - 1.5 \times IQR$
> or greater than $Q3 + 1.5 \times IQR$,
>
> where *Q1* is the first quartile, *Q3* is the third quartile, and *IQR* is the interquartile range.

In the *No Outliers* sheet, the values of the *Number of Distractors* and *Response Time* variables have been replaced by an asterix highlighted in yellow.

Data from either sheet can be copied and pasted into the *Number of Distractors* and *Response Time* columns of the sheet for the analysis you wish to perform. For the *No Outliers* data, you will then need to also manually delete each highlighted asterix in the analysis sheet. Note that each analysis sheet has an accompanying visual "chart" sheet that you can print as a PDF file if you wish.

Degrees of Freedom and the Adjusted Coefficient of Determination

If the observed responses y_k, $k = 1, \ldots, n$ were totally unrelated to each other, we could select their n values independently. In statistical jargon, we would say we had

"n degrees of freedom."

If, however, Model 0 holds:

$$y_k = \bar{y} + residual_k, \quad k = 1, \ldots, n,$$

then we could select only $n-1$ of these values independently (provided the mean value $\bar{y}$ was already fixed). This is because[3]

$$\sum_{k=1}^{n} y_k = \sum_{k=1}^{n} \bar{y} + \sum_{k=1}^{n} residual_k = n\bar{y} \text{ implies } y_1 + y_2 + \ldots + y_{n-1} + y_n = ny,$$

so $y_n = n\bar{y} - (y_1 + \ldots + y_{n-1})$ is determined once $y_1, \ldots, y_{n-1}$ have also been chosen.

[3] The residuals sum up to zero, by definition.

Similarly, if Model 1 holds:

$$y_k = mx_k + b + residual_k,\ k = 1, \ldots, n,$$

then we can only select $n-2$ of the y_k values independently once $\hat{m}$ and $\hat{b}$ are fixed. (To see this mathematically, we would need the formulas for $\hat{m}$ and $\hat{b}$.)

We can take degrees of freedom into account by evaluating a model according to the sum of squared error per degree of freedom, rather than just *SSE*.

Comparing the models based on "proportional reduction of SSE per degree of freedom" gives us the **adjusted coefficient of determination**:

$$\overline{R}^2 = \left(\frac{SSE(Model\ 0)}{n-1} - \frac{SSE(Model\ 1)}{n-2} \right) \Big/ \left(\frac{SSE(Model\ 0)}{n-1} \right)$$

$\overline{R}^2$ is the proportion by which SSE *per degree of freedom* is reduced if we use Model 1 rather than Model 0.

The adjusted coefficient of determination corrects for the fact that Model 1 would fit the data better than Model 0 simply because it has more parameters to fit, and hence is a better measure of the degree to which the independent factor accounts for the dependent response.

What You Will Do

Use the Regression Analysis sheet in this workbook to perform regression of *Response Time* on *Number of Distractors* for several experimental treatments. Perform your analyses both with and without outliers. After each analysis, record the **slope, R^2**, and **adjusted R^2** values in the table in the top right corner of the worksheet. Also study the chart produced by each analysis.

Are your analyses consistent with the predictions of Feature Integration Theory?

The Analysis of Covariance - ANCOVA[4]

Feature Integration Theory makes a variety of claims concerning factors and one response variable. The factors are:

Display Type	(categorical: "feature" or "conjunction"),
Target	(categorical: "present" or "absent")
Number of Distractors	(discrete numerical: 4, 16, or 64 in *Lab 6*)

The response variable is:

Response Time	(continuous numerical, measured in milliseconds)

For fixed values of the two categorical variables, you have performed *regression analysis* of your own CogLab data with *Response Time* as the dependent variable and *Number of Distractors* as the independent variable. This is an appropriate analysis for numerical dependent and independent variables.

You also performed *analysis of variance* – ANOVA – to compare the effect on *Response Time* of different levels of each categorical variable, *Display Type* and *Target*. For each comparison, the level of the other categorical variable was held fixed.

The third variable, *Number of Distractors*, which was not the focus of the ANOVA, was actually also treated as categorical in those analyses. If this seemed a bit odd to you, your statistical instincts are good! A better analysis than ANOVA for assessing the effect of categorical variables when you also have a *numerical* auxiliary or **concomitant variable** such as *Number of Distractors* is ANCOVA – analysis of covariance.

Like ANOVA, the basic idea of ANCOVA is to group observations based on "not distinguishing" versus "distinguishing" values of the categorical variable of interest, holding the values of the other categorical variables constant.

[4] In order to take advantage of certain theoretical results for statistical inference, standard ANCOVA often places certain constraints on the regression parameters. Since our goal here is data analysis rather than inference, we are not including those constraints.

Unlike ANOVA, in ANCOVA the estimates of the response variable within each group are not group means but instead are obtained by least-squares regression with the concomittant variables as independent variables.

We still measure error using SSE, and assess goodness of fit using R^2 as a PRE measure.

Because Feature Integration Theory makes claims about both categorical and numerical factors, ANCOVA is better suited than ANOVA to analyzing the data from this experiment for the purpose of evaluating the categorical claims of this theory.

What You Will Do

Run the ANOVA and ANCOVA analyses for several of the stated contrasts. Perform your analyses both with and without outliers. After each analysis, record the **R^2** and **adjusted R^2** values in the table in the top right corner of the worksheet. Also study the chart produced by each analysis.

Which of these two types of analyses is a better test of Feature Integration Theory, and why?

How well do the results of that analysis match the predictions of Feature Integration Theory?