10. LINEAR REGRESSION AND CORRELATION

The contingency table describes an association between two nominal (categorical) variables (e.g., "use of supplemental oxygen" and "mountaineer survival"). We have already used this approach to describe associations between continuous variables by breaking the continuous variables into categories ("size category of iguana" and "direction of change in iguana weight"). Regression and correlation are more powerful methods to describe and test associations between continuous variables.

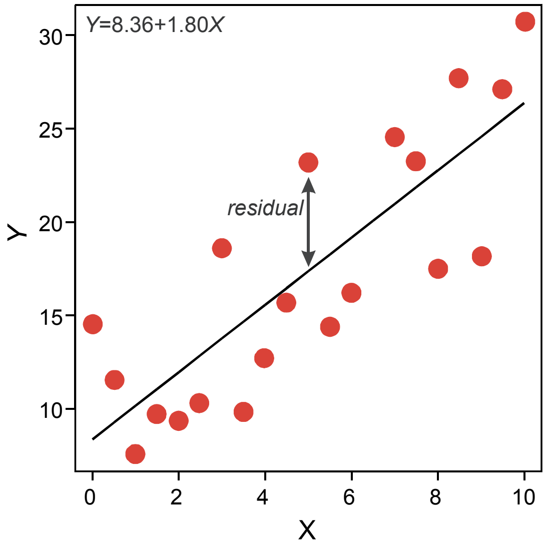
Regression and correlation are related but have different purposes. The goal of regression is to predict the value of one variable, *Y*, from measurements of the other variable, *X*, whereas correlation merely describes the strength of association between *X* and *Y*. When using regression, keep in mind that to predict *Y* from *X* in no way implies that *X* is the cause of *Y*. Demonstrating a cause and effect relationship requires careful experimental design with appropriate controls to rule out other causes.

Simple Linear Regression

Linear regression assumes that *Y*'s relationship to *X* is a straight line:

$$Y = \alpha + \beta X$$

α is the *Y*-intercept (the mean value of *Y* when *X* is zero), and β is the slope of the line (the amount that *Y* changes per unit change in *X*). α and β are population *parameters* that describe the true relationship of *Y* on *X*. The quantities *a* and *b* are the sample estimates of these two parameters.



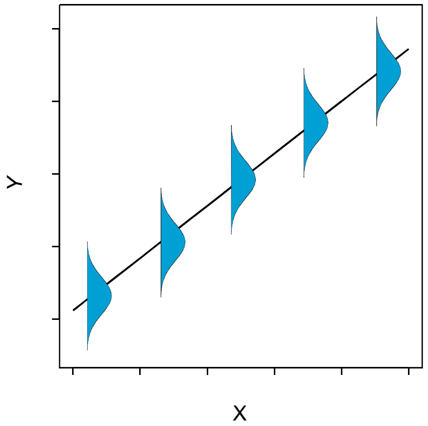
Individual *Y*-values will not lie directly on the line, but will be scattered above and below by a random amount. The difference between a *Y* observation and the predicted *Y* on the line is called the *residual*. Under the method of least squares, α and β are estimated from data by finding the values of *a* and *b* that minimize the sum of squared residuals.

In this lab, we emphasise graphical tools that help you evaluate the assumptions that underlie regression analysis. These methods rely on visual and statistical inspection of data. Your goal is to try to make the data fit the assumptions as closely as possible, and then decide whether the agreement between fact and assumption is close enough to proceed with the analysis. Be prepared to try several remedies and to choose the best among them.

Assumptions of Linear Regression

Linear regression rests on three special assumptions.

- For every value of *X*, there is a distribution of possible *Y* values whose mean falls on a straight line (i.e., the relationship is linear).

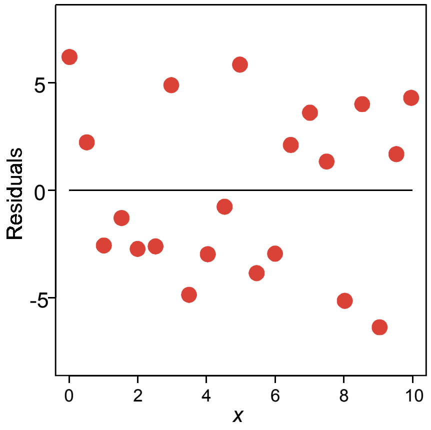


- The distribution of possible *Y* values at each *X* is normally distributed.
- The variance of the *Y* values is assumed to be the same at all values of *X*.

Added to these are the usual assumptions that observations must be independent of another. We will also assume that there is no measurement error in *X*.

When evaluating whether the assumptions are met, several tools are useful and should be considered a part of any regression analysis:

- Smoothing or Spline fitting. Draws a smooth function through the Y-observations that helps the eye determine whether the relationship is linear. Selecting a high value of lambda, the smoothing coefficient, will result in a straight line. Selecting a smaller lambda produces a more bumpy curve. Select the value of lambda that you think best describes the relationship between *Y* and *X*, and try to evaluate how closely this approximates a straight line fit.
- Plot of residuals. Fit a straight line and then plot the residuals in *Y* against *X*.

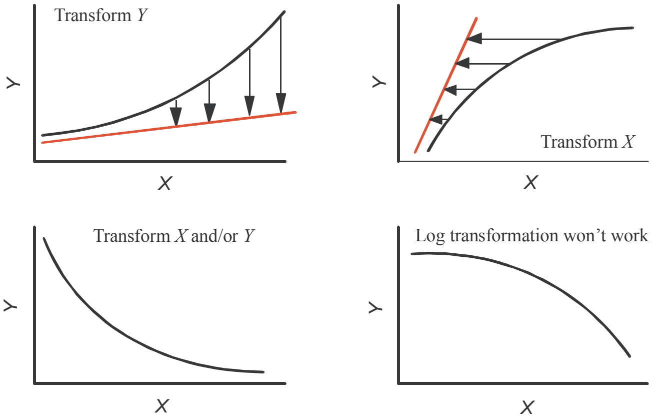


Examine the residual plot for indications that the variability of the residuals varies widely across the range of *X* values (indicating that the assumption of equal variances is violated), or that the residuals are highly unevenly distributed on both sides of the line over the range of *X* values (indicating that the relationship is probably not linear).

- Distribution of residuals using standard methods (e.g., histogram, boxplot) will help assess fit of the residuals to the normal distribution..

Transformations in Regression

To meet the assumptions of regression try transforming *X* and/or *Y*. The log transformation (or lox(*X*+1) when there are zeros) is by far the most commonly used transformation and we will use it primarily. If the relationship of *Y* on *X* is described by a power function (e.g. $Y = 3\,e^{1.4X}$) then transforming both *X* and *Y* should yield satisfactory results. For other relationships it may only be necessary to transform only one of the variables. Use whatever works best.



The square root and arcsine transformations are also frequently use (from the overview of these transformations in the previous ANOVA lab, you might be able to guess the types of data for which these other transformations might be suited). Transformations offer no hope, however, if the real curve contains either a distinct peak or a distinct valley. In these and other cases nonlinear regression methods are appropriate (we will not be covering this topic).

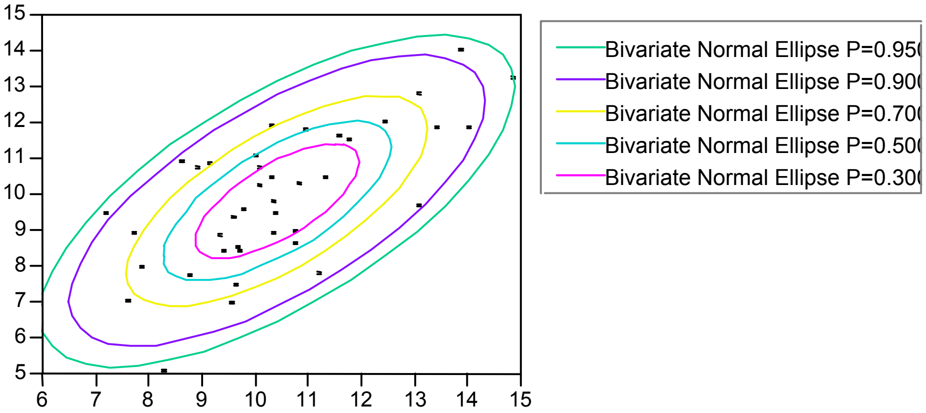
Other Topics in Linear Regression

Several other issues in linear regression will be covered in lecture, including

- The regression effect
- Standard error and 95% confidence interval for the slope
- Coefficient of determination (r^2)
- Confidence bands
- Prediction intervals
- t-test of slope
- ANOVA table and *F*-test
- Fixed and random effects
- Inverse prediction

Simple Linear Correlation

In correlation analysis, pairs of *X*,*Y* values are assumed to be drawn at random by the investigator from a population. Our goal is to determine whether the two variables are associated. The correlation coefficient, *r*, measures the strength of the association; *r* may vary between −1 and 1. Linear correlation assumes that the distribution of *X,Y* pairs in the population has a bivariate normal distribution. A bivariate normal distribution is a bell-shaped distribution in 3D. The figure below shows a sample of points from a bivariate normal distribution, and several contours of that distribution. These contours encircle 95%, 90%, … 30% of the observations in the population.



If this assumption of bivariate normality is violated, and cannot be corrected by transformation, then a nonparametric method, the Spearman rank correlation is used instead.

Using the Program

Linear Regression - Use **Fit Y by X** (or **Bivariate** in the **JMPIN Starter**) and designate your *X* and *Y* variables. The computer will display a scatterplot of the data. Click the red "▼" next to the "Bivariate" title bar to select → **Fit Line**, which fits a linear regression to the data. This also yields the estimated regression equation (slope and intercept), a summary of the fit (r^2, the fraction of variation in *Y* that is "explained" by *X*, an index of the strength of the fit) and the lack of fit (ignore for now). The result window will also display estimates, standard errors and significance tests of intercept and slope (based on the *t*-distribution), and an *F*-test of the fit of the whole model to the data (analogous to ANOVA).

Below the scatterplot you'll find another red "▼" next to "—Linear Fit". Click the symbol to select a series of other options:

→ **Confid Curves Fit**: Displays the confidence bands, i.e., the upper and lower bounds of 95% confidence intervals for the predicted mean *Y* value (i.e., for the location of the line itself) at each *X*.

→ **Confid Curves Indiv**: Displays the prediction intervals, i.e., the upper and lower bounds of 95% confidence intervals for the predicted value of a single new *Y* observation.

→ **Save residuals**: Creates a new column in the data table containing the residuals. Useful if you wish to generate a histogram and box plot for the residuals.

→ **Plot Residuals**: Generates a plot of *Y* residuals against *X*. Useful for checking equality of variance of residuals across the range of *X*-values.

→ **Remove Fit**: Deletes the line and associated results.

Spline - Use **Fit Y by X** (or **Bivariate** in the **JMPIN Starter**) and designate your *X* and *Y* variables. The computer will display a scatterplot of the data. Click the red "▼" next to the "Bivariate" title bar to select → **Fit Spline**. Start with a **lambda** of 1, and then go higher or lower if this value yields a spline fit with too little or too much wobble. The goal is to examine the general trend of the data.

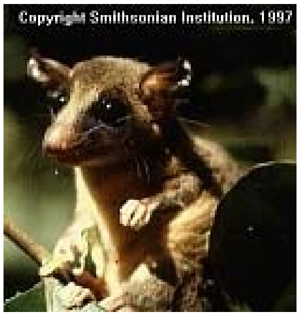
Transformations - Use the JMP IN calculator to create new variables, as you did for ANOVA.

Correlation – Use **Analyze→Multivariate** (or select the **Multivariate** tab in the JMPIN Starter). Select the two variables you want to correlate and place them both in the Y box. This will generate a table reporting the linear correlation *r* between all pairs of selected variables. A scatterplot for all pairs of variables also results, along with a 95% density ellipse for each pair. This ellipse should enclose approximately 95% of the observations if the two variables have a bivariate normal distribution. Click the red "▼" next to the "Multivariate" title bar to select other options:

→ **Pairwise correlations**: Reports the correlation coefficient for each pair of variables, the *P*-value for a test of the null hypothesis of no correlation, and a bar chart indicating the magnitude of each correlation.

→ **Nonparametric Correlations** → **Spearman's Rho**: Reports the Spearman's rank correlation coefficient and the *P*-value for a test of the null hypothesis of no correlation.

Problems



1. Open the data set **mammals.jmp**, which contains average brain and body masses for 62 species of land mammals. The data are from Allison and Cicchetti (1976, *Science* 194: 732–734).

   a) Use the **Multivariate** option to calculate the linear correlation between brain size and body size in this sample of mammals.

   b) Observe the scatter plot of observations and the 95% density ellipse. This ellipse represents a contour on the bivariate normal distribution that best fits the data (picture a bell in 3D). Do the data appear to conform well to the assumption of bivariate normality?

   c) If the data do not fit the assumption of bivariate normality, test the correlation between brain size and body size using a nonparametric method.

d) Explore transformations of these variables, to yield a scatterplot that fits a bivariate normal distribution better than the untransformed data.

e) With the transformed data, use linear regression to predict (transformed) brain mass from (transformed) body mass. Report also the r^2 (coefficient of determination, a measure of the strength of fit).

f) What are the main assumptions of linear regression? Do they appear to be met in this case? Show how you determined this.

g) Is the slope of the regression significantly different from zero?

h) Place the <u>confidence bands</u> on the scatterplot. What do these bands measure?

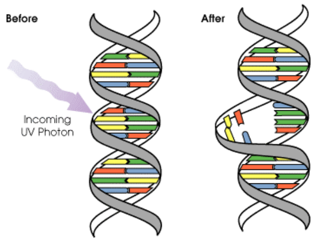
i) Remove the confidence bands and place the <u>prediction intervals</u> on the plot instead. Why are these wider than the confidence bands?

j) Which species of mammal has the largest brain, taking into account differences between species in body size? Show how you determined this. You should be pleased by the answer.

k) What other species have relatively large brains, again taking into account differences between species in body size? Which species has the smallest brain? Presumably it would be difficult to house train.

l) Do you think that the different mammal species in this data set represent a random sample? Do you think that the observations are independent?

2. Open the data file **mutation.jmp**. The file contains estimates of whole-genome, deleterious point mutation rates in a range of animal taxa. Each rate was estimated by comparing two closely related species in the numer of differences at “important” sites in shared genes relative to the number of differences at “silent” sites. Mutation rates are measured as the number of new, deleterious mutations expected per individual per generation (column #3). The estimates were obtained from Table 1 of Keightley and Eyre-Walker (2000, *Science* 290: 331-333).



a) Observe the range of values for different taxa of animals. Note especially the large values for humans and other primates! If this is not frightening, I don’t know what is.

b) Test whether whole-genome deleterious mutation rates, measured per generation, are correlated with generation time. Use the most powerful method available. Show how you met the assumptions of the method.

3. The data file **gagurine.jmp** contains data on the concentration of the chemical GAG in the urine of 314 children aged from zero to seventeen years. The aim of the study was to produce a chart to help a paediatrican assess whether an individual child's GAG concentration is "typical". Variables are age of child (in years) and GAG concentration (the units have been lost). The data were taken from Venables and Ripley's MASS library (original data from Prosser, cited in Venables and Ripley).

   a) Predict GAG concentration from the age of the child. Explain and justify your methods.

   b) To be useful to the pediatrician, who would like to use the relationship to decide if a given child has a typical GAG concentration, what other graphical information might your analysis provide?

4. Open the data file **kpmales.jmp**. The data are survival to 28 days of 7037 male human infants. The data were collected in English hospitals in the few years after the end of World War II. Gestation is rounded to the nearest 5 days, and birth weight is rounded to the nearest 0.5-pound. All infants having the same birth weight and gestation time are grouped. The number of infants in each gestation/birthweight group is indicated by frequency. Survival is the proportion of infants in each group that survived. The data are from Karn and Penrose (1951, *Ann. Eugen*. 16:147-164).

   a) Examine the relationship between survival (*Y*) and birthweight (*X*) using a spline (make sure the variable frequency is in the **frequency** box of the **Fit Y by X** window). Describe the shape of the relationship.

   b) The spline can be used as a nonlinear regression (a topic we won't have time to cover in this course). Using the spline curve, predict survival probability (to 28 days) of a male human infant born at 3 pounds (1.4 kg) (use your eye to make the predicton; never mind the spline equation)? Predict survival for a birth weight of 4 pounds, 7 ponds, and 10 pounds. What pattern is suggested by these data? Do you think that a transformation of these data might linearize the relationship between birth weight and survival?

   c) Compare the relationship between survival and birth weight to that between survival and gestation. How do the patterns differ?

   d) Are birth weight and gestation time correlated? Explain how you tested this, and provide a justification for your methods.