Richard Gonzalez
Psych 613
Version 2.6 (Nov 2018)

LECTURE NOTES #6: Correlation and Regression

Reading assignment: Stay current with the reading

KNNL Chapters 1, 2, 3, 4, and 15; CCWA chapters 1 and 2

There are several ways to think about regression, and we will cover a few of them. Each perspective, or way of thinking about regression, lends itself to answering different research questions. Using different perspectives on regression will show us the generality of the technique, which will help us solve new types of data analysis problems that we may encounter in our research.

1. Describing bivariate data.

   The bivariate normal distribution generalizes the normal distribution. See Figure 6-1 for examples.

   Sometimes we want to find the "relationship"[1], or "association," between two variables. This can be done visually with a scatter plot. Examples of scatter plots are given in Figures 6-2 and 6-3 with n=20 and n=500, respectively.

   The correlation is a quantitative measure to assess the linear association between two variables. The correlation can be thought of as having two parts: one part that measures the association between variables and another part that acts like a normalizing constant. The first part is called the covariance. To understand better the concept of covariance recall the definition of sums of squares

$$\begin{aligned} S_Y &= \sum (Y - \overline{Y})^2 && (6\text{-}1) \\ &= \sum (Y - \overline{Y})(Y - \overline{Y}) && (6\text{-}2) \end{aligned}$$

   This is the sum of the product of the differences between the scores and the mean. The estimated variance of Y is

$$\frac{S_Y}{N-1} \qquad (6\text{-}3)$$

[1]Some people quibble about whether to refer to a correlation between two variables as a "relation" or a "relationship". I suppose proper usage would have a relation refer to two variables and a relationship refer to the bond between two people. But I'm not bothered such things and tend to use both.

Figure 6-1: Bivariate Density Function

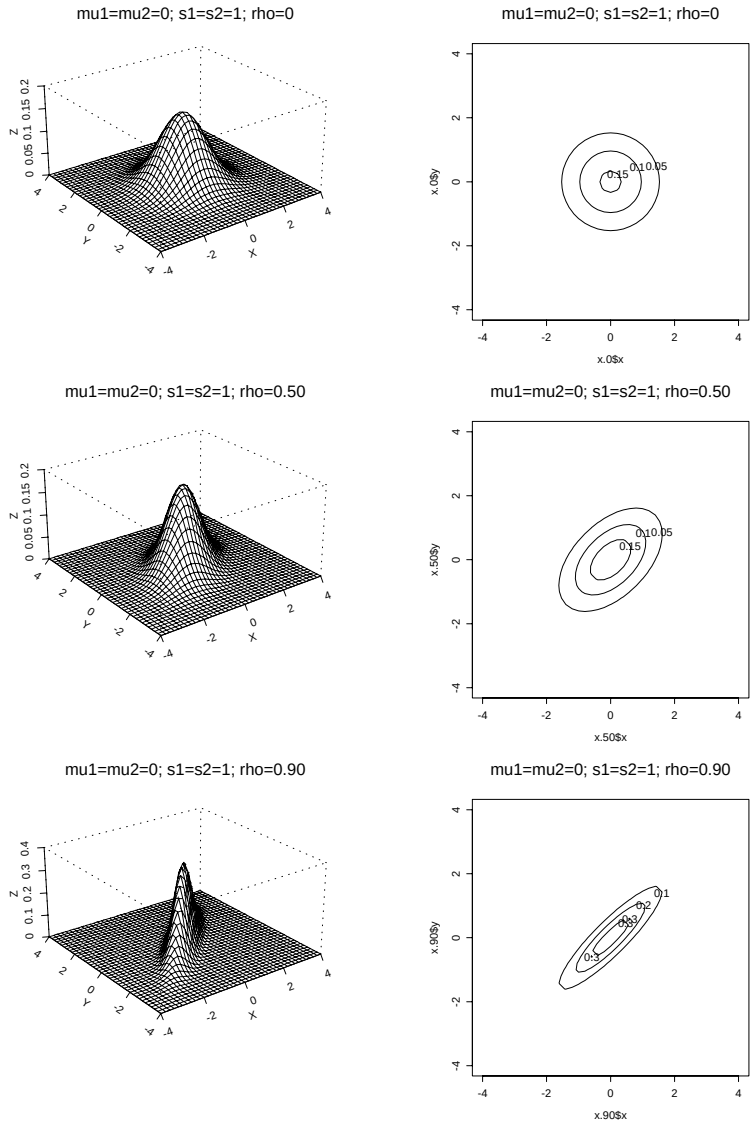


Figure 6-2: Varying ρ: Each scatter plot contains a sample with 20 points.

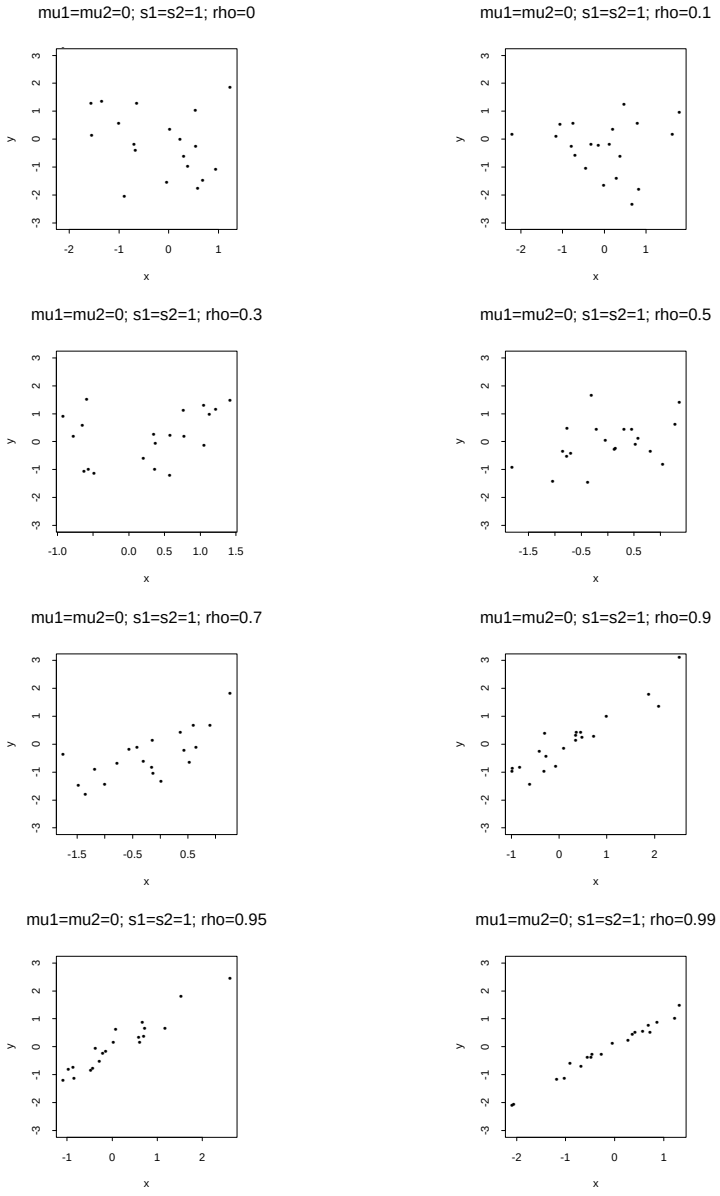
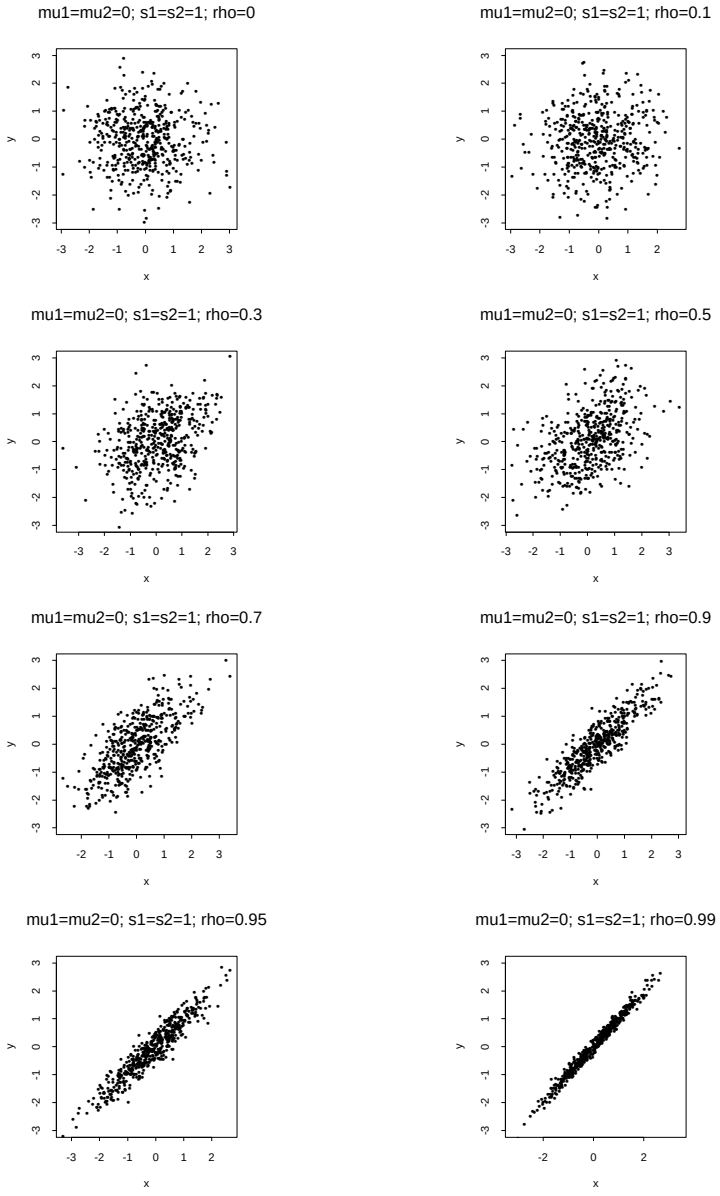


Figure 6-3: Varying ρ: Each scatter plot contains a sample with 500 points.



The covariance is similar to the variance except that it is defined over two variables (X and Y) rather than one (Y). We begin with the numerator of the covariance–it is the "sums of squares" of the two variables.

$$S_{xy} = \sum(X - \overline{X})(Y - \overline{Y}) \tag{6-4}$$

The (estimated) covariance is

$$\frac{S_{xy}}{N-1} \tag{6-5}$$

The interpretation of the covariance is very similar to that of the variance. The covariance is a measure of both the direction and the magnitude of the linear association between X and Y [2]. When will a covariance be positive? negative?

The covariance can be viewed intuitively as a sum of "matches" in terms of a subject being on the same side of the mean on each variable. That is, for a particular subject a match would mean that the subject is, say, greater than the mean on variable X *and* is also greater than the mean on variable Y. A "mismatch" is defined for a subject as the score on variable X is greater than the mean but the score on variable Y is less than the mean, or vice versa. For a particular subject i, a match leads to a positive product in Equation 6-4 whereas a mismatch leads to a negative product.

In these notes $S_{xx} = S_x$

We can think of the variance as the covariance of a variable with itself, denoted $S_{xx}/(N-1)$. The covariance of a variable with itself and the variance of that variable are identical. I will use $S_{xx}/(N-1)$ and $S_x/(N-1)$ interchangeably to denote the variance of X.

The covariance has the property that adding a constant to either variable does not change the covariance, i.e.,

$$\frac{S_{(x+c)y}}{N-1} = \frac{S_{xy}}{N-1} \tag{6-6}$$

and multiplying either variable by a constant changes the covariance by a multiple of that constant, i.e.,

$$\frac{cS_{xy}}{N-1} = \frac{S_{(cx)y}}{N-1} \tag{6-7}$$

The property that multiplying by a constant changes the covariance can make interpreting the covariance difficult because we would get a different covariance if we used one measurement as opposed to another (e.g., length in feet v. length in yards).

[2] Unlike a variance the covariance can be negative.

One simple trick fixes this scaling problem. Recall that the standard deviation also has these two properties (adding a constant doesn't change the standard deviation and multiplying by a constant changes the standard deviation by a multiple of that constant). So, the standard deviations can be used to "normalize" the covariance such that

definition: correlation

$$\frac{\text{covariance(X,Y)}}{\text{st.dev(X) st.dev(Y)}} \tag{6-8}$$

$$\frac{S_{xy}}{\sqrt{S_x S_y}} \tag{6-9}$$

Dividing by the standard deviation makes the scaling constant c cancel out. The N - 1 terms in both the numerator and denominator of Equation 6-8 also cancel. Thus, Equation 6-9 can be interpreted as a ratio of "sums of squares," equivalently as the ratio of the covariance to the product of the standard deviations.

We have just defined a useful concept. Equation 6-8 is the definition of the *correlation coefficient* (and so is Equation 6-9). The correlation is the covariance normalized by the standard deviations of the two variables and ranges from -1 to 1. The normalization removes the scaling issue mentioned in the previous paragraph about multiplying by a constant. The sample correlation is denoted r_{xy} (sometimes just r for short). [3]

The correlation squared, denoted r^2_{xy}, has an interesting interpretation: it is the proportion of the variability in one variable that can be accounted for by *a linear function* of the other variable. It turns out that regression has a structural model that is analogous to the structural model we saw for ANOVA. The structural model for the correlation is $Y = \beta_0 + \beta_1 X + \epsilon$, where β_0 is the intercept (analogous to the grand mean μ). β_1 is the slope (analogous to the treatment effect α in ANOVA), and ϵ is the usual error term.

Because we know that proportions add up to 100%, if the r^2_{xy} is the proportion of variance that is explained, then $(1-r^2_{xy})$ is the proportion of the variance that is not explained. That is,

$$r^2_{xy} + (1 - r^2_{xy}) \quad = \quad 1 \tag{6-10}$$

We will discuss what is meant by "variance accounted for" later, but for now we can make use of concepts from ANOVA. The term r^2_{xy} is telling you what your model is picking up (like sums of squares between groups in the ANOVA), the term $(1-r^2_{xy})$ is telling you what our model is not picking up (like sums of squares residual). [4]

[3] For the linear algebra buffs: the correlation is the dot product of two difference vectors normalized by their norm (length) and is equal to the cosine of the angle between the two vectors in N-dimensional space where N is the number of subjects.

[4] This property leads to an intuitive rationale for why the correlation is bounded between -1 and 1. Since

One must always be careful when interpreting a correlation coefficient because, among other things, it is quite sensitive to outliers. The effects of a single outlier can have dramatic effects. So, when interpreting a correlation one must always, always check the scatter plot for outliers.

Also, one needs to make sure that the scatterplot suggests that a linear relationship between the two variables is appropriate. Researchers have been misled because they compute a correlation that is essentially zero and conclude that the two variables are not "associated." The correlation of zero just means that (assuming no outliers are present) a *linear* "association" does not appear to be present. A quite dramatic curvilinear relationship might be present, and the correlation coefficient could be equal to zero. Just because one observes a correlation of zero does not mean that the two variables are not related.

2. Inferential tests on a correlation

We can test whether a correlation is significantly different from zero. We need to introduce some notation. The correlation coefficient r is a sample statistic that estimates the true population correlation, denoted ρ (in the analogous way that the sample mean estimates the population parameter μ). The underlying population distribution is assumed to be a bivariate normal.

The null hypothesis is

$$H_O: \qquad \rho = 0$$

t-test for a correlation

The test statistic is simple; recall the definition of the t-test

$$t \sim \frac{\text{estimate}}{\text{standard error of the estimate}} \tag{6-11}$$

I won't go into the details here, but the t-test for the correlation takes the usual form of estimate (in this case the correlation) divided by the standard error of the estimate

$$= \frac{r_{xy}}{\sqrt{1-r_{xy}^2}/\sqrt{N-2}} \tag{6-12}$$

r^2 = SSR/SST, which is the proportion of sum of squares regression to sum of squares total, we have r^2 being bounded between 0 and 1. The square root of r is then bounded by -1 and 1. A more rigorous proof involves covariance algebra, a topic we will cover in a later set of lecture notes. To show that the correlation between X and Y is bounded between -1 and 1, first take Z scores of X and Y so that the correlation equals the covariance. Second, set up a system of two equations, one as the sum of the two Z scores and the other as the difference. Using the covariance algebra rules one can show that the variance of a difference of two Z scores is equal to 2-2r and the variance of a sum of two Z scores is equal to 2+2r. Because variances are nonnegative we know that 2-2r$\geq$0 and 2+2r$\geq$0 and the bounds for the correlation r fall out automatically.

Figure 6-4: Hypothesis Testing Framework for Correlations

Null Hypothesis

- H_o: $\rho = 0$
- H_a: $\rho \neq 0$ (two-sided test)

where ρ is the population correlation coefficient.

Structural Model and Test Statistic

The structural model for the null hypothesis at $\rho = 0$ follows the usual t-test distribution

$$t \sim \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}}$$

For the correlation tested at $\rho = 0$, t observed is

$$t_{\text{observed}} = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}/\sqrt{N - 2}}$$

with df $= N - 2$, which is total sample size minus 2.

Critical Test Value We use the t table to find the critical value of t, denoted t_{critical} for the specific degrees of freedom, two-sided, and $\alpha = 0.05$.

Statistical decision If $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

This observed value is compared to the t-distribution having df = N - 2. This test can only be used when the null hypothesis is $\rho = 0$. Later in these lecture notes I present a test of the correlation that permits null hypotheses at other values instead of 0. The t-test for the correlation can be summarized using the hypothesis testing template introduced in Lecture Notes 1 (see Figure 6-4).

It turns out that it is even easier to test the null hypothesis that $\rho = 0$ in the context of a linear regression with a single predictor. You automatically get the test of significance for ρ when doing a simple linear regression. More on this later.

3. Fisher's r-to-Z transformation for confidence intervals and tests of null hypothesis not equal to 0

To test the null hypothesis that the population correlation ρ is different than zero or to build a confidence interval around an observed value of the correlation r, one needs to use a transformation developed by Fisher. The reason is that the sampling

distribution of the correlation is symmetric only when $\rho = 0$; it is asymmetric for all other values of ρ. So we need to use this transformation when constructing confidence intervals or testing null hypotheses other than zero.

I present Fisher's technique first for one correlation, then for two correlations from independent samples, and then for the general situation of computing a contrast over any number of correlations from independent samples.

(a) One correlation

Fisher r-to-Z

I will use the symbol Z_f to denote Fisher's r-to-Z transformation and the symbol Z to denote the "z-score" corresponding to a normal distribution. Fisher showed that if a single observed correlation r is transformed by the formula

$$Z_f = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \tag{6-13}$$

then the sampling distribution of Z_f is asymptotically normal with variance $\frac{1}{N-3}$, where N is the number of subjects on which the correlation is computed. By "log" I mean natural log (the ln button on your calculator). An interesting geometric interpretation of this transformation of the correlation is given by Bond and Richardson (2004, *Psychometrika, 69*, 291-03).

Because of this result we know that the ratio

$$Z = \frac{Z_f - Z_{\text{null}}}{\sqrt{\frac{1}{N-3}}} \tag{6-14}$$

is asymptotically normally distributed. This is a variant of the usual t-test formula, which people use as the asymptotic z form rather than the t form. In this formula, Z_{null} corresponds to the null hypothesis of r transformed to the Z_f scale. We can compare the computed value of Z against the table of the normal distribution. For example, for a two-tailed test at $\alpha = .05$, the critical value from the normal distribution table is 1.96. Hence, if the observed value of Z exceeds 1.96, then we reject the null hypothesis. However, if the observed value of Z is less than 1.96, then we fail to reject the null hypothesis.

Here is a numerical example. Suppose my null hypothesis is $\rho = .2$, I observe a sample r = .7 with $N = 40$ subjects, and want to perform a two-tailed test of significance. First transform both the observed correlation r and the hypothesized ρ into their respective Z_f values using Equation 6-13. The observed correlation r = .7 becomes 0.867 and the hypothesized value .2 becomes 0.203. Now apply Equation 6-14 and you have a test of significance:

$$Z = \frac{.867 - .203}{\sqrt{\frac{1}{37}}}$$

Figure 6-5: Hypothesis Testing Framework for One Correlation: Fisher's test

Null Hypothesis

- H_o: $\rho = k$
- H_a: $\rho \neq k$ (two-sided test)

where ρ is the population correlation coefficient and k is the value of the null hypothesis, which need not be 0.

Structural Model and Test Statistic

The null hypothesis follows the z test

$$z \sim \frac{\text{f}(r) \text{ - } \text{f}(k)}{\text{estimated st. dev. of the sampling distribution}}$$

where f(r) is the Fisher r-to-z transformation of r and f(k) is the Fisher r-to-z transformation of the null hypothesized value k.

We have z observed

$$z_{\text{obs}} = \frac{\text{f}(r) \text{ - } \text{f}(k)}{\sqrt{\frac{1}{N-3}}}$$

Critical Test Value We use the z table to find the critical value of z, denoted z_{critical} , which for two-tailed $\alpha = 0.05$ is equal to 1.96.

Statistical decision If $|z_{\text{observed}}| > z_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

$$= \quad 4.04$$

which is statistically significant because the observed Z exceeds the critical value 1.96 from the normal distribution.

A confidence interval can also be constructed around the correlation r by converting the correlation r into Z_f and using this formula:

$$Z_f \quad \pm \quad Z\sqrt{\frac{1}{N-3}} \tag{6-15}$$

where Z is the z-score corresponding to the normal distribution (e.g., if you want a 95% confidence interval, then Z will be 1.96). You may find it convenient to transform the endpoints back to the original scale in order to facilitate communication. For example, suppose you observe a correlation r = .7 with N = 40, this leads to a 95% confidence interval of

$$(Z_f - Z\sqrt{\frac{1}{N-3}}) \quad \text{and} \quad (Z_f + Z\sqrt{\frac{1}{N-3}}) \tag{6-16}$$

$$\tag{6-17}$$

which leads to the interval (.545, 1.19) on Z_f, or on the correlation scale the interval (.497, .83). This re-scaled interval is not symmetric around the observed correlation r of .7. The inverse of Equation 6-13 (i.e., the function that converts Z_f back to the correlation scale r) is

$$r \quad = \quad \frac{e^{2Z_f} - 1}{e^{2Z_f} + 1} \tag{6-18}$$

You would use this formula to convert the endpoints of the confidence interval back to the correlation scale.

(b) Extending the Fisher formulation to correlations from two independent samples

With two samples, we have a natural generalization. Suppose you want to test whether two population correlations ρ are identical. Your null hypothesis would be

$$\mathrm{H}_o : \rho_1 \quad = \quad \rho_2 \tag{6-19}$$

This null hypothesis can also be written in this (identical) manner, paralleling the form for the difference between two means: $\rho_1 - \rho_2 = 0$ (or more generally, we can test the null hypothesis that $\rho_1 - \rho_2 = k$, i.e., the difference in the two population correlations is equal to the value k).

Suppose you observe two correlations (one for each sample) $r_1 = .2$ and $r_2 = .8$ with $N_1 = 40$ subjects in sample 1 and $N_2 = 45$ subjects in sample 2. First,

transform each of the two correlations using Fisher's r-to-Z (i.e., Equation 6-13). Each of these Z_f will have a variance of $\frac{1}{N-3}$, and because the two correlations come from independent samples the variance for a difference between two Z_f's is equal to $\frac{1}{N_1-3} + \frac{1}{N_2-3}$.

testing two independent correlations

The computed Z value is given by

$$\frac{(Z_{f1} - Z_{f2}) - f(k)}{\sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}} \tag{6-20}$$

where $f(k)$ is the Fisher transformation of the value k as defined by $\rho_1 - \rho_2 = k$. As before, if the observed Z using Equation 6-20 exceeds 1.96, then you reject the null hypothesis.

Continuing with the example, suppose $r_1 = .2$ and $r_2 = .8$, $N_1 = 40$ subjects in sample 1, $N_2 = 45$ subjects, and k is equal to zero. First, apply Equation 6-13 to each of r_1, r_2, and k. When k =0, then the transformed value of k will also be 0. The values of Z_{f1} and Z_{f2} are, respectively, .203 and 1.099. Plugging everything into Equation 6-20, we get

$$\begin{aligned} Z &= \frac{.203 - 1.099}{\sqrt{\frac{1}{40-3} + \frac{1}{45-3}}} \\ &= -3.97 \end{aligned}$$

which is statistically significant using the two-tailed $\alpha = .05$ criterion because the observed Z exceeds the tabled value 1.96.

(c) Bonus: Contrasts on correlations from an arbitrary number of independent samples

We can generalize this notion from two independent correlations to an arbitrary number of T independent correlations. Convert each of the T correlations to their respective Z_f scale. Define any contrast you'd like on those T means, and compute an "$\hat{\text{I}}$" in the usual way

$$\hat{\text{I}} = \sum \lambda_i Z_{f_i} \tag{6-21}$$

where λ represents the contrast.

The variance of $\hat{\text{I}}$ is given by

$$\sum \left(\lambda_i^2 \frac{1}{N_i - 3} \right) \tag{6-22}$$

Gotta love those contrasts!

Finally, compute the value Z by taking the ratio of Equation 6-21 and the square root of Equation 6-22 as follows:

$$Z = \frac{\hat{\mathrm{I}}}{\sqrt{\sum \lambda_i^2 \frac{1}{N_i - 3}}} \tag{6-23}$$

You compare this computed value of Z to 1.96 for a two-tailed $\alpha = .05$. This formulation assumes that the null hypothesis for the contrast value is 0[5].

With this procedure you can test any contrast on T independent correlations. Convince yourself of this procedure by verifying that Equation 6-23 on two correlations (i.e., T = 2) with the contrast $\lambda = (1, -1)$ is equivalent to the case I presented for testing two correlations using Fisher's r-to-z (Equation 6-20).

Equation 6-22 does not involve pooling the different variances (i.e., sample sizes). This trick of defining an error in terms of a weighted sum of independent sample variances was proposed by Wald, and any procedure that defines the error term in this manner is known as the Wald test (another well-known Wald test is the test on proportions that we will cover later in the course). I point out the Wald test only because it differs from the other kinds of tests we use in this class (i.e., those based on the t and the F distribution).

4. Spearman's ρ: A nonparametric correlation coefficient.

Spearman's rank correlation

Spearman's rank correlation is identical to computing the usual Pearson correlation (Equation 6-8) on data that have been transformed into ranks. No need to learn a new formula, just compute the usual correlation on the ranks and you've done a Spearman's rho.

There are other types of nonparametric measures of association. One was developed by Kruskal and Goodman, called "gamma" (i.e., the Greek letter γ). There are also many variants of γ. The γ measure has an intuitive interpretation involving proportion of times that the ordering on the X variable is consistent with the ordering on the Y variable (i.e., proportion of times in the data set that if one subject had a higher X score than another, that same subject also has a higher Y score). If there is time next semester we may come back to these alternative nonparametric measures of association that pretty much just focus on the ordinal relation variables.

5. Fitting straight lines.

[5]For more general null hypotheses that the population contrast equals a specific value k, just put $(\hat{\mathrm{I}} - Z_f(k))$ in the numerator of Equation 6-23.

Recall from high school algebra that the equation of a line is

$$\mathrm{Y} = \mathrm{b} + \mathrm{mX} \tag{6-24}$$

where b is the y-intercept and m is the slope. Examples will be given in class. The "high school method" for finding the slope and intercept works great when the points fall along a straight line. However, with real data the points will rarely fall exactly on a straight line. There's bound to be some noise. We need a technique for estimating the slope and y-intercept of a line from noisy data–that is, from a scatterplot. This technique is called regression. We will begin with a simple regression equation and build up to some rather complicated regression equations.

As with most statistics, we define the population values. Denote β_0 as the population y-intercept and β_1 as the population slope. We gather a sample and compute the sample estimates such that

$$\mathrm{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X} + \epsilon \tag{6-25}$$

where $\hat{\beta}_0$ is the sample estimate of the y-intercept, $\hat{\beta}_1$ is the sample estimate of the slope and ϵ is the usual error term. Equation 6-25 is the structural model for a simple linear regression. The structural model posits a linear relationship between the X and Y variables.

Computational formulas[6] for the slope and intercept are:

$$\hat{\beta}_1 = \frac{\mathrm{S_{xy}}}{\mathrm{S_{xx}}} \tag{6-26}$$

$$\hat{\beta}_0 = \overline{\mathrm{Y}} - \hat{\beta}_1 \overline{\mathrm{X}} \tag{6-27}$$

6. Keeping track of units.

Psychologists perform linear regression on all sorts of variables. It is instructive to keep track of units when interpreting the intercept and the slope. For instance, suppose someone is regressing intelligence (the criterion variable) on some anxiety measure (the predictor variable). Intelligence is usually scaled to have a mean of 100 and a standard deviation of 15, whereas the anxiety measure may be on a 7-point scale, so will obviously be on a different scale. In this regression, the intercept will be in the units of the criterion variable and the slope will be telling you how to convert from the units of the predictor to the units of the criterion. So, in this example the slope can be interpreted as a conversion factor that is in intelligence per anxiety units.

[6]For those interested in derivations.... The derivation of the slope and intercept in the least squares sense is finding the value for the intercept and slope that minimize MSE. That is, we want to find values of the slope and intercept that minimize $(\sum(Y - (\beta_0 + \beta_1 X))^2)$. Take the partial derivative of MSE with respect to β_0 and to β_1, set them equal to 0, solve for the slope and intercept, and there you go.

It will be useful for you to work through this argument under different examples. For instance, what happens to the slope and intercept when the predictor is transformed by a linear function before it enters the regression (such as in converting degrees Fahrenheit to degrees Celsius).

7. Data, fits, and residuals

The fit of a regression equation is computed directly from the structural model (Equation 6-25 for the special case of simple, linear regression). That is, we take a particular subject's score on variable X, multiply by the slope, and add the y-intercept to get the predicted score $\hat{Y}$. The fitted value is $\hat{Y}$. Each subject has his or her own fitted value, which is a linear transformation of his or her own score on variable X.

There are two intuitive ways to understand how the fit works (i.e., how the line is determined). One is the eyeball method. Another is the rubber band method demonstrated in class[7].

A more rigorous method is to define an equation and solve for the relevant parameters. Regardless of the method used, the bottom line is that regression finds the "line of best fit" with all the details being in how "best" is defined.

Recall that the general schema for a model is data = fit + residual, or, by re-arranging terms, residual = data - fit. This last term (data - fit) will be very useful because the residual can be expressed as

$$\begin{aligned} \text{residual} &= \text{data} - \text{fit} && \text{(6-28)} \\ \epsilon &= Y - \hat{Y} && \text{(6-29)} \\ &= Y - (\hat{\beta}_0 + \hat{\beta}_1 X) && \text{(6-30)} \end{aligned}$$

One way to define "fit" is to find those $\hat{\beta}_0$ and $\hat{\beta}_1$ that make the residual as small as possible. Intuitively, we search all possible values for $\hat{\beta}_0$ and $\hat{\beta}_1$ until we found the smallest possible sum of squared residuals, i.e., SSResidual.

[7] This graphic illustrates that errors are defined in terms of vertical length from the regression line rather than shortest possible distance from the regression line. It turns out that the way to define regression was hotly debated. Gauss preferred to minimize squared vertical difference—the technique now in use—and Laplace argued for minimizing the sum of absolute values, or equivalently the sum of the minimum distances. One reason that Gauss' method prevailed is that it is not sensitive to linear transformations of the variables. That is, imagine a right triangle that has as one side the vertical difference between the datum and the regression line, and another side the shortest distance between the datum and the regression line (thus, forming a right triangle between the side and the regression line). A change of scale of the abscissa changes the shortest distance but not the vertical distance. Also, ease of tractability led to widespread use of the Gaussian formulation (squares are easier to work with than absolute values). An interesting side note: Gauss referred to ϵ as "the error to be feared"—maybe something got lost in the translation from Latin because all those little ϵs are not really that scary.

The reason we have to square the residuals is that across all subjects in a sample the sum of the ϵ's will always be zero. Squaring the ϵ's makes the residuals more useful as a quantity to minimize.[8] The two parameters in regression are chosen so as to minimize the following quantity

$$\sum \epsilon^2 = \sum (\mathrm{Y} - \hat{\mathrm{Y}})^2 \quad (6\text{-}31)$$

$$= \sum (\mathrm{Y} - (\hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}))^2 \quad (6\text{-}32)$$

One detail needs to be made clear. If there are N subjects there will be N values for the Y variable (the predicted variable, or the dependent variable), N different $\hat{\mathrm{Y}}$'s (the fitted values), N values of X (the predictor variable, or the independent variable), and N different ϵ's. On the other hand, there will only be one $\hat{\beta}_0$ and one $\hat{\beta}_1$ for the entire sample. These latter two values are estimated from the entire sample and are assumed to be the same for all subjects. More general versions of regression allow each subject or unit to have their own intercept and slopes; this is accomplished by treating subject as a random effect much like we did in repeated measures ANOVA.

8. Desiderata for $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (a) The line of best fit should pass through the point corresponding to the means of each of the two variables, i.e., the point $(\overline{\mathrm{X}}, \overline{\mathrm{Y}})$.

   (b) The sum of the squared residuals should be as small as possible.

9. Sum of Squares Decomposition

decomposition of sum of squares in a regression

Regression also corresponds to a decomposition of the sums of squares. The general framework is

$$\text{SS total} = \text{SS regression} + \text{SS residual} \quad (6\text{-}33)$$

The terms are defined as

$$\sum_{\mathrm{i}} (\mathrm{Y_i} - \overline{\mathrm{Y}})^2 = \sum_{\mathrm{i}} (\hat{\mathrm{Y}}_\mathrm{i} - \overline{\mathrm{Y}})^2 + \sum_{\mathrm{i}} (\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2 \quad (6\text{-}34)$$

In words, the sum of squares total is the variability between the observed scores and the overall mean ("grand mean") of the Y scores, the sum of squares regression is

[8] This is exactly the same trick used when computing the variance. The sum of the differences between each score from the sample mean is always zero so, instead of taking the sum of the differences, we take the sum of the squared differences.

the variability between the predicted values for Y and the overall mean (analogous to comparing the model to the grand mean), and the sum of squared residuals is the variability between the observed scores and the predicted values (how close the fitted values are to the observed values).

Note how similar this decomposition is to the one we saw for ANOVA. It turns out that ANOVA is a special case of regression. If one does regression in a particular way, one gets the identical results as ANOVA. But, regression allows you to do much more; it is more general, hence the term "general linear model." More on this later.

The population variance of the residuals, σ_ϵ^2, is estimated by

$$\hat{\sigma}^2_{\mathrm{Y.X}} = \frac{\sum(\mathrm{Y}-\hat{\mathrm{Y}})^2}{\mathrm{N}-2} \tag{6-35}$$

$$= \frac{\text{SS residual}}{\mathrm{N}-2} \tag{6-36}$$

where N is the number of subjects, i.e., number of (x, y) pairs. Why divide by N - 2? The general heuristic is that the denominator is the number of subjects minus the number of parameters that are fitted. In this simple case we are fitting two parameters, the slope and the intercept. The reason for caring about the number of parameters is because the fitted values, $\hat{\mathrm{Y}}$, depend on the number of parameters that are in the model. This heuristic also applies to t tests and ANOVA. For example, recall that in a two sample t test there are N - 2 degrees of freedom. The N - 2 comes from the fact that two parameters are being fitted: the grand mean μ and the treatment effect α_i.

Equation 6-35 is similar to the mean square error we saw in the ANOVA source table. So, if we have sum of squares and mean squared terms, then there's got to be a source table lurking around. The form of the regression source table is

SS	df	MS	F
SSR= $\sum(\hat{\mathrm{Y}}_\mathrm{i}-\overline{\mathrm{Y}})^2$	number of parameters - 1	SSR/df	MSR/MSE
SSE= $\sum(\mathrm{Y}_\mathrm{i}-\hat{\mathrm{Y}}_\mathrm{i})^2$	N - number of parameters	SSE/df	

The MSE is an estimate of the population residual variance σ_ϵ^2. Some people like to take the square root of the MSE as a measure of goodness of fit because, in a sense, the $\sqrt{\mathrm{MSE}}$ is the "average" residual so it gives a measure of the typical residual. I put the word average in quotes because it is the square root of the sum of squared residuals divided by the degrees of freedom for the error—it is thus an average—like term.

Important fact

The F test tells you whether the linear transformation of the X variable is a significant predictor of the Y variable. In the case of linear regression with one predictor, this F

test also tells you whether the slope is significantly different from zero and whether the correlation is significantly different from zero. All three tests (F for the regression, the t test for the slope, and t test for the correlation) are equivalent tests when there is only one predictor in the regression model—an important fact you should memorize.

dot notation

I introduced new notation in the left hand side of Equation 6-35: the left hand term has a dot in the subscript between Y and X. In general, what is on the left side of the dot refers to dependent variables (or variables that are predicted) and what is on the right side of the dot refers to independent variables (or predictor variables).

10. Interesting observations

The ratio $SS_{regression}/SS_{total} = R^2_{Y.X}$. The term $R_{Y.X}$ is interpreted as the overall correlation between the variables on the left side of the dot with the variables on the right side of the dot. Sometimes I'll just write R^2 for short, omitting the subscripts. In the case of simple linear regression with one predictor, there is one variable on each side of the dot, so R^2 equals the square of the correlation coefficient r. If the regression source table is available, a correlation can quickly be calculated by

$$r = \sqrt{\frac{SS_{regression}}{SS_{total}}} \tag{6-37}$$

To reiterate: the relation between R^2 and r only holds when there is one independent variable and one dependent variable. R^2 is more general than the correlation r because it can be defined for any number of independent (or predictor) variables[9].

R^2 is not a perfect measure. It is made up of two components: a part that is related to the regression (SS regression) and a part that is completely driven by the variance in the dependent variable (SST). So, two data sets may have different R^2s with identical SS regression (that is, in some sense identical fits of the regression) but different variances for the dependent variable. Some people (e.g., Gelman & Hall, 2007) argue

[9]CAVEAT. Here is an esoteric fact. If the regression is forced through the origin in the sense that the intercept is forced to be zero, then the correct way to compute R^2 is (SST-SSE)/SST, where SST is computed manually from the Y scores by $\Sigma(Y - \overline{Y})^2$. The reason has to do with the definition of SSR and whether or not the variables are mean-corrected. Of course, when the intercept term is included in the structural model, then SST-SSE=SSR so both forms are identical. However, they diverge when the intercept is not included, and to make matters worse, SPSS uses the incorrect R^2 formula when the intercept is omitted from the structural model (same issue for R as well). When using SPSS to force a regression through the intercept, one needs to compute R^2 manually as (SST-SSE)/SST. This SPSS error can lead to strange results in R^2. Rarely in psychology do we force the regression line to go through the origin because usually the scale has an arbitrary zero point. In economics, this makes more sense (e.g., when salary is \$0, then 0% of the labor force will work so the regression line between salary and percent work force who work should go through the point 0,0). There are just a few applications in Psychology where forcing the regression line through origin makes sense.

that it is misleading to evaluate the fit of a regression based on a measure such as R^2 that is sensitive to the variability of the dependent variable. An example would be if one data set has a restricted range in the dependent variable relative to the other data set but both data sets have the identical regression fit. The R^2s will differ in the two data sets because they have different dependent variable variances even though the line of best fit is (in the underlying data generating mechanism) identical. The square root of MSE is sometimes taken as a better measure of fit by some statisticians because it is an index of an "average residual" (this concept will come up in future lecture notes). There is also a related issue of restricted range in the predictor variable influencing the fit of the regression line.

11. Explaining Variance in Y

Earlier I mentioned that the square root of the MSE can be interpreted as the "average" residual. In this section I want to develop the MSE in a little more detail.

The residual variance $\hat{\sigma}^2_{Y.X}$ (aka MSE) refers to the variability of Y after X has been taken into account. Another way of saying this: the variance of Y with the linear effect of X partialled out.

An intuitive view. If you only had knowledge of the Y scores, what would be your best prediction about the average value of Y? Obviously, the sample mean $\overline{Y}$. You would want to place a confidence interval around $\overline{Y}$. To calculate this CI you use the sample variance of Y, $\hat{\sigma}^2_Y$ (which is an estimate of σ^2_Y).

Now, suppose you had knowledge of the corresponding X scores. To the extent that X and Y are related, the knowledge of X can be used to make better predictions of Y. For example, instead of predicting the overall sample mean of Y, you can predict the sample mean of Y corresponding to individuals who have a particular X score—this is a "correlational" statement because you are predicting Y given X. Naturally, this extra knowledge translates into narrower confidence intervals. Thus the MSE of the regression will be smaller than the variance of the Y scores when there is a correlation between X and Y. Specifically, MSE can be written in a form that illuminates the role of R^2

$$\begin{aligned} \text{MSE} &= \hat{\sigma}^2 \left(\frac{N-1}{N-2} \right) (1 - R^2) && \text{(6-38)} \\ &= \frac{\sum (Y - \overline{Y})^2 (1 - R^2)}{N-2} && \text{(6-39)} \\ &= \frac{S_Y}{N-2} (1 - R^2) && \text{(6-40)} \end{aligned}$$

Equation 6-40 makes transparent what is meant by "explaining variance in Y." The (1 - R^2) term makes MSE smaller as the correlation between the two variables increase.

So, the MSE is smaller than the variance of the Y scores whenever $R^2 > 0$. The better the predictor, in the sense of a higher R^2, the smaller the CI around the tailor made prediction.

12. Testing the null hypotheses that $\beta_1 = 0$ and $\beta_0 = 0$.

This section follows our usual hypothesis testing template. We will use the definition of t as the estimate divided by the standard error of the estimate, compute t_{obs} and compare the absolute value of t_{obs} to $t_{critical}$.

Recall that a t test can be formed when one divides an estimate by its standard error. As previously stated, the slope and intercept for simple linear regression can be estimated by

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad (6\text{-}41)$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \quad (6\text{-}42)$$

All we need are the standard error of the estimates. They are, respectively,

$$\text{st. err}(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum(X - \overline{X})^2}} \quad (6\text{-}43)$$

$$\text{st. err}(\hat{\beta}_0) = \sqrt{MSE}\sqrt{\frac{\sum X^2}{N\sum(X - \overline{X})^2}} \quad (6\text{-}44)$$

Form t-tests in the usual way as estimate/st.error of estimate. You compute one t test for the slope and another for the intercept. Both the slope and the intercept have df = N - 2. You can follow the same template we've been using throughout the semester from here on out. Just compare t observed to t critical. The t critical is from the table lookup using your desired α level and df = N - 2.

Confidence intervals around the slope and the intercept are easy to construct. The form of a confidence interval is

$$\text{estimate} \pm t_{\alpha/2,\ df} \text{ st. error of est.} \quad (6\text{-}45)$$

All you need to do is find the appropriate value in the t table (df = N - 2, or more generally, the degrees of freedom associated with the sum of squares residual) and then plug in the values for the estimate and the st. error of the estimate (equations given above).

Why do the denominators of the standard errors have $\Sigma(X - \overline{X})^2$? It would seem that the variability of the predictor X shouldn't matter? Actually it does. If X has a

Figure 6-6: Hypothesis Testing Framework for the Slope

Null Hypothesis

- H_o: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$ (two-sided test)

where β_1 is the population slope.

Structural Model and Test Statistic

The structural model for the null hypothesis at $\beta_1 = 0$ follows the usual t-test distribution

$$t \sim \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}}$$

For the slope tested at $\beta_1 = 0$, t observed is

$$t_{\text{observed}} = \frac{\hat{\beta}_1}{\sqrt{\frac{\text{MSE}}{\sum(\text{X}-\overline{\text{X}})^2}}}$$

with df $= N - 2$, which is total sample size minus 2.

Critical Test Value We use the t table to find the critical value of t, denoted t_{critical} for the specific degrees of freedom, two-sided, and $\alpha = 0.05$.

Statistical decision If $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

lot of variability, the slope will be more stable (have a smaller standard error) than if X has little variability. Draw some scatterplots where everything is the same except for the variability of X and convince yourself of this.

The slope is related to the correlation through a simple multiplication.

$$r = \beta_1 \frac{\text{sd}_x}{\text{sd}_y}$$

If you know the standard deviations and the slope, then you know the correlation. This is the reason why the t-test for the slope is identical to the t test for the correlations (both at the null value of 0)–if one is zero, then the other is also zero assuming non-zero standard deviations.

Both tests for the slope and intercept are summarized in the hypothesis testing template Figure 6-6 and Figure 6-7, respectively.

Figure 6-7: Hypothesis Testing Framework for the Intercept

Null Hypothesis

- H_o: $\beta_0 = 0$
- H_a: $\beta_0 \neq 0$ (two-sided test)

where β_0 is the population intercept.

Structural Model and Test Statistic

The structural model for the null hypothesis at $\beta_0 = 0$ follows the usual t-test distribution

For the intercept tested at $\beta_0 = 0$, t observed is

$$t_{\text{observed}} = \frac{\hat{\beta}_0}{\sqrt{\text{MSE}\frac{\sum \text{X}^2}{\text{N}\sum(\text{X}-\overline{\text{X}})^2}}}$$

with df $= N - 2$, which is total sample size minus 2.

Critical Test Value We use the t table to find the critical value of t, denoted t_{critical} for the specific degrees of freedom, two-sided, and $\alpha = 0.05$.

Statistical decision If $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

13. Statistical assumptions in linear regression.

As usual, when testing hypotheses or building confidence intervals there are statistical assumptions that must be made.

(a) Each observation is **independent** of all others (or equivalently, the residual terms, ϵ_i, are independent).

(b) At each value of X, the Y scores are **normally distributed** (or equivalently, at each value of X, the ϵ's are normally distributed).

(c) Across all values of X, the Y scores have the **same variance** (or equivalently, across all values of X, the ϵ's have the same variance).

14. Making predictions about Y given knowledge of X.

(a) Predicting individual points.

The regression equation can be used to make predictions *about individual points*. If you know a value of X, you can predict the corresponding value of Y (i.e., you can calculate the predicted value $\hat{Y}$ by multiplying by the slope and adding the intercept term). This yields the prediction $\hat{Y}$.

A confidence interval can be placed around the predicted value $\hat{Y}$. This confidence interval is typically called a "prediction interval." The formula for the standard error of the estimate is quite ugly.

$$\text{st. error}(\hat{Y}) = \sqrt{\text{MSE}}\sqrt{1 + \frac{1}{N} + \frac{(X - \overline{X})^2}{\sum(X - \overline{X})^2}} \tag{6-46}$$

N corresponds to the total number of subjects that entered into the calculation of the slope and intercept.

This value for the standard error is plugged into the usual form of the confidence interval

prediction interval around an individual prediction

$$\hat{Y} \pm t \text{ st. error}(\hat{Y}) \tag{6-47}$$

(b) Predicting average values

One can also predict the expected value of Y (i.e., the mean value of Y) given an X value. This contrasts with the above goal of predicting an individual Y for an X value. Clearly, there is more error in predicting an individual value than predicting a mean. This is reflected in the width of the confidence interval (i.e., in the size of the standard error)–a wider interval is assigned to predictions of individual points than to predictions of average values.

The standard error of the expected value of Y given an X value is

$$\text{st. error}(\hat{\text{E}}(\text{Y})) = \sqrt{\text{MSE}}\sqrt{\frac{1}{\text{N}} + \frac{(\text{X} - \overline{\text{X}})^2}{\sum(\text{X} - \overline{\text{X}})^2}} \quad (6\text{-}48)$$

confidence interval around predicted mean

The confidence interval for the predicted mean is

$$\hat{\text{E}}(\text{Y}) \pm t \text{ st. error}(\hat{\text{E}}(\text{Y})) \quad (6\text{-}49)$$

The phrase "confidence interval" is used to denote intervals around predicted average values and the phrase "prediction interval" is used to denote intervals around predictions of individual points.

(c) Predicting from new values of X (i.e., X scores that were not included in the original estimation of the slope and intercept)

The equations listed above also allow one to predict from new values of X. For example, suppose I perform a regression using 12 subjects. I then am ask about a 13th subject for whom an X value is available (and different from the previous 12 subjects' X scores) but no Y has been observed yet. I can predict this Y value and build a prediction interval around it by applying the above equations.

(d) SPSS stuff

SPSS calls st. error($\hat{\text{E}}(\text{Y})$) SEPRED. This is the standard error appropriate when predicting a mean. I don't think the term st. error($\hat{\text{Y}}$) (the standard error appropriate when predicting an individual score) is given directly in the output of the **REGRESSION** command. But, it is trivial to compute by hand as long as you already have SEPRED and MSE. Look at Equation 6-46. The only difference between Equations 6-46 and 6-48 is that the former has a 1 added under the radical. It is easy to show, with some high school algebra, that

$$\text{st. error}(\hat{\text{Y}}) = \sqrt{\text{MSE} + \text{SEPRED}^2} \quad (6\text{-}50)$$

Even though SPSS doesn't seem to print out the standard error of an individual prediction it does compute the confidence intervals of both mean and individual predictions. In the regression command just add the line /SAVE PRED SEPRED MCIN ICIN, which will add to the data set for each case the prediction, the sepred, and the confidence intervals (CIN) of both the mean prediction (M) and the individual prediction (I). Again, this information about each the CI for each case is added as new columns in the data set file not in the output.

SPSS is not very user-friendly at helping you predict from new values of X. The trick is to include the new values of X in your data and give the corresponding Y scores missing codes. The **REGRESSION** command will ignore any cases with missing values for computing slopes, intercepts, and p-values. Thus, inferential analyses will not be influenced by including these new values having missing codes on the Y. However, some of the output will print out that case with the missing data. For example, the SEPRED for a missing value will print because all that is needed for SEPRED is MSE and knowledge of the X values. Of course, some output will not be printed for that new value (e.g., the residual for the missing Y cannot be calculated so it will not be printed). An example is given Appendix 2.

Appendix 1: Relevant SPSS syntax

1. SPSS CORRELATION command

   The general syntax is

   ```
   correlation variables = list.
   ```

   You can specify additional subcommands such as asking for a two-tailed t-test

   ```
    /print = twotail sig
   ```

   and asking for some descriptive statistics on the variables.

   ```
   /statistics = all
   ```

   If you included these subcommands remember to put the period at the end of the last subcommand rather than the end of the CORRELATION line.

2. SPSS REGRESSION command

   This command has many nice features, especially the ability to analyze residuals to detect where the model is going wrong, giving you information you can use to improve the model.

   The general syntax is (the order of these lines appears to be important)

   ```
   REGRESSION VARIABLES = LIST
       /statistics R coeff anova ci cha
       /dependent = nameofdv
       /method=enter listofiv
       /casewise all plot(zresid) default sepred cook.
   ```

   All subcommands are self-explanatory except maybe "cha," which gives the R^2 change from method to method subcommand, as we will see later it is possible to have multiple

method subcommands in the same regression command. The "all" in casewise makes all cases print (omitting the word "all" prints only the outliers–those residuals plus or minus 3 standard deviation). In the examples I will do over the next few days you will notice the use of these different features.

One side-effect of listing all the variables on the first line is that missing data on any variable in "list" will automatically remove that subject's data from the analysis even if the regression does not name the variable with the missing data point in the /method or /dependent line.

Depending on the version of SPSS you use, it may be necessary to add the subcommand

```
/width=132
```

in the regression to permit the printing of all the options specified (one could also change the default line width to 132 through the SPSS Preferences dialog box). For instance, if you don't change the default line width you may not get "sepred" to show up in your output.

Another subcommand to REGRESSION that may be useful is

```
/residuals = defaults sepred
```

which produces a histogram and normal plot of the residuals (useful for checking normality).

Sometimes you may want to save the residuals and "sepred" to the data file for subsequent analysis. You can use this subcommand to save those numbers:

```
/save pred sepred residual
```

This will save the predictions, sepred and residual information into your worksheet as data. You could then use the GRAPH command on these variables to produce any kind of plot you want. For example, the residuals are automatically given the variable name `res_1` (check your data worksheet to see what SPSS named the variables). Then enter this command to plot the residuals against another variable, say, a variable in your dataset called "time"

```
GRAPH
 /scatterplot(bivar)= time with res_1
 /missing = listwise.
```

Plotting the residuals against a variable like time could provide information about the independence assumption (e.g., if the residuals show a correlation with time then they aren't independent). You could could also produce a histogram and normal plot manually on the saved residuals using the EXAMINE command.

The /save subcommand in REGRESSION can also store the mean and individual confidence intervals through the subcommand

```
 /save pred sepred residual mcin icin
```

You can also produce scatterplots directly in the regression command using the /SCATTERPLOT subcommand. If you want to plot residuals against one of the predictors, you can include this subcommand in the REGRESSION command

```
 /SCATTERPLOT (*resid, X1)
```

where X1 is the predictor you want to use in the scatterplot and the asterisk in front of resid tells SPSS that resid is not a variable in your dataset but a temporary variable created during the regression analysis. Using this method you would not need to save the residuals. This command coupled with a trick to plot against case number can be useful to check violations of independence (assuming your data are ordered by time). This will be covered in a later lecture notes.

3. SPSS Graphs

To get a matrix of scatterplots (all possible scatterplots between Y, X1, and X2):

```
GRAPH
 /scatterplot(matrix)= Y x1 x2.
```

To get the regression line printed you could edit the scatterplot produced by GRAPH or more directly use the IGRAPH command.

```
IGRAPH /VIEWNAME='Scatterplot'
 /X1 = VAR(test2)
 /Y = VAR(test1)
 /FITLINE METHOD = REGRESSION LINEAR LINE = TOTAL
 /SCATTER.
```

Note: newer versions of SPSS seem to have dropped the IGRAPH and moved to a new command GGRAPH with slightly different syntax.

Appendix 2: Example with two midterm scores

Here are data from 31 students who took two midterms (from Ryan, Joiner, and Ryan, 1985). We'll run several analyses to illustrate some of the points in these lecture notes.

```
 data list free /  test1 test2.
 begin data
50 69
66 85
73 88
84 70
57 84
83 78
76 90
95 97
73 79
78 95
48 67
53 60
54 79
79 79
76 88
90 98
60 56
89 87
83 91
81 86
57 69
71 75
86 98
82 70
95 91
42 48
75 52
54 44
54 51
65 73
61 52
end data.
```

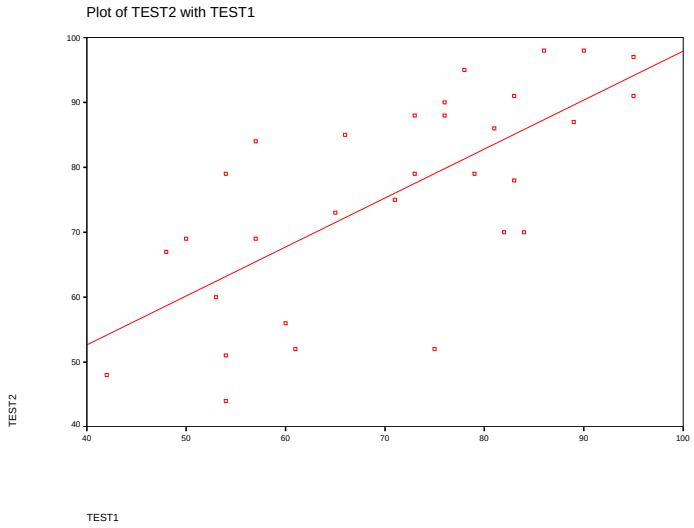
First let's look at the scatter plot of these two variables (could also use the scatterplot command on the windows interface of SPSS).

```
GRAPH
 /scatterplot(bivar)= test1 with test2
 /missing = listwise.  .
```

Second, we'll compute the correlation between the two midterms.

```
  correlation test2 test1
   /print= twotail
   /statistics=all.

Variable     Cases           Mean          Std Dev
```

Figure 6-8: SPSS scatter plot




```
TEST2           31        75.7742          15.9179
TEST1           31        70.6452          14.8314

Variables              Cases    Cross-Prod Dev    Variance-Covar

TEST2     TEST1           31         4979.5161          165.9839
                       - -  Correlation Coefficients  - -

             TEST2       TEST1

TEST2         1.0000       .7031
             (   31)     (   31)
             P= .        P= .000

TEST1          .7031      1.0000
             (   31)     (   31)
             P= .000     P= .

(Coefficient / (Cases) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

The Cross-Prod Dev term is what we call S_{XY}, the sum of squared cross products, and the Variance-Covar term is what we call the covariance. Note that the correlation is 0.7031 which equals the covariance divided by the product of the two standard deviations (i.e., $\frac{165.9839}{15.9179 \times 14.8314}$). The **CORRELATION** command also tests the null hypotheses.

Third, we perform a regression of test 2 on test 1. Note the *regression dependent.variable*

on independent.variables(s) terminology in the previous sentence.

SPSS tidbit: you may need to re-adjust your SPSS default value to print everything listed here. For example, in the windows version, click on Edit/Preferences/Output and change 80 to 132; in the Unix version, click on Options/Preferences/Output.

```
 regression variables= test1 test2
   /statistics = r anov  coeff ci
   /dependent=test2
   /method=enter test1
   /residuals=defaults sepred outliers(cook)
   /casewise=defaults sepred all
   /scatterplot (test2, test1).

Multiple R           .70307
R Square             .49431
Adjusted R Square    .47687
Standard Error     11.51311

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1          3757.42041       3757.42041
Residual            29          3843.99895        132.55169

F =      28.34683       Signif F =  .0000


---------------------- Variables in the Equation -----------------------

Variable              B        SE B    T   Sig T   95% Confdnce Intrvl B    Beta

TEST1           .754575     .141726  5.324 .0000    .464713     1.044438    .703068
(Constant)    22.467092   10.223577  2.198 .0361   1.557529    43.376655


Casewise Plot of Standardized Residual

*: Selected   M: Missing

  Case    TEST2       *PRED      *RESID     *SEPRED
       1     69.00     60.1959      8.8041      3.5829
       2     85.00     72.2691     12.7309      2.1701
       3     88.00     77.5511     10.4489      2.0946
       4     70.00     85.8514    -15.8514      2.8033
       5     84.00     65.4779     18.5221      2.8312
       6     78.00     85.0969     -7.0969      2.7096
       7     90.00     79.8148     10.1852      2.2027
       8     97.00     94.1518      2.8482      4.0237
       9     79.00     77.5511      1.4489      2.0946
      10     95.00     81.3240     13.6760      2.3157
      11     67.00     58.6867      8.3133      3.8179
      12     60.00     62.4596     -2.4596      3.2450
      13     79.00     63.2142     15.7858      3.1370
      14     79.00     82.0785     -3.0785      2.3828
      15     88.00     79.8148      8.1852      2.2027
      16     98.00     90.3789      7.6211      3.4352
      17     56.00     67.7416    -11.7416      2.5597
      18     87.00     89.6243     -2.6243      3.3231
      19     91.00     85.0969      5.9031      2.7096
      20     86.00     83.5877      2.4123      2.5357
      21     69.00     65.4779      3.5221      2.8312
      22     75.00     76.0419     -1.0419      2.0684
```

```
     23     98.00     87.3606     10.6394      3.0019
     24     70.00     84.3423    -14.3423      2.6202
     25     91.00     94.1518     -3.1518      4.0237
     26     48.00     54.1593     -6.1593      4.5561
     27     52.00     79.0602    -27.0602      2.1580
     28     44.00     63.2142    -19.2142      3.1370
     29     51.00     63.2142    -12.2142      3.1370
     30     73.00     71.5145      1.4855      2.2172
     31     52.00     68.4962    -16.4962      2.4788
Case # TEST2       *PRED      *RESID     *SEPRED
         -3.0      0.0      3.0


Residuals Statistics:

                   Min           Max     Mean  Std Dev   N

*PRED          54.1593       94.1518  75.7742  11.1914  31
*ZPRED         -1.9314        1.6421    .0000   1.0000  31
*SEPRED         2.0684        4.5561   2.8518    .6579  31
*ADJPRED       55.3029       94.5903  75.7765  11.1939  31
*RESID        -27.0602       18.5221    .0000  11.3196  31
*ZRESID        -2.3504        1.6088    .0000    .9832  31
*SRESID        -2.3928        1.6598   -.0001   1.0122  31
*DRESID       -28.0455       19.7143   -.0023  12.0003  31
*SDRESID       -2.6245        1.7143   -.0098   1.0418  31
*MAHAL           .0006        3.7303    .9677    .9325  31
*COOK D          .0001         .1206    .0300    .0324  31
*LEVER           .0000         .1243    .0323    .0311  31

Total Cases =       31

Durbin-Watson Test =   1.69958
```

You can use this output to build confidence intervals and prediction intervals around predicted values.

Finally, we will compute a prediction on Y for a new value of X. I am not aware of a user-friendly way of doing this on SPSS. The kludge[10] is to include the new value as a case in the data file. It will have a missing code for the dependent variable. I chose 9999 to be

[10]Yes, *kludge* is a real word meaning something that has been thrown together with poorly matching parts, forming a distressing whole (aptly describing SPSS). Here is the entry from the Oxford English Dictionary:

> kludge klud. slang (orig. U.S.). Also kluge. [J. W. Granholm's jocular invention: see first quot.; cf. also bodge v., fudge v.] 'An ill-assorted collection of poorly-matching parts, forming a distressing whole' (Granholm); esp. in Computing, a machine, system, or program that has been improvised or 'bodged' together; a hastily improvised and poorly thought-out solution to a fault or 'bug'.
>
> - 1962 J. W. Granholm in Datamation Feb. 30/1 The word 'kludge' is..derived from the same root as the German Kluge.., originally meaning 'smart' or 'witty'... 'Kludge' eventually came to mean 'not so smart' or 'pretty ridiculous'.
> - 1966 New Scientist 22 Dec. 699/1 Kludges are conceived of man's natural fallibility, nourished by his loyalty to erroneous opinion, and perfected by the human capacity to apply maximum effort only when proceeding in the wrong direction.
> - 1979 Personal Computer World Nov. 71/3 Kludge, a local modification or patch in a computer program to overcome some error or design fault.

Figure 6-9: SPSS histogram of residuals

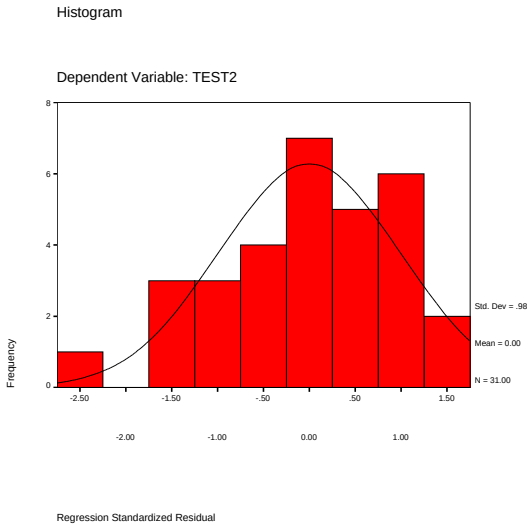
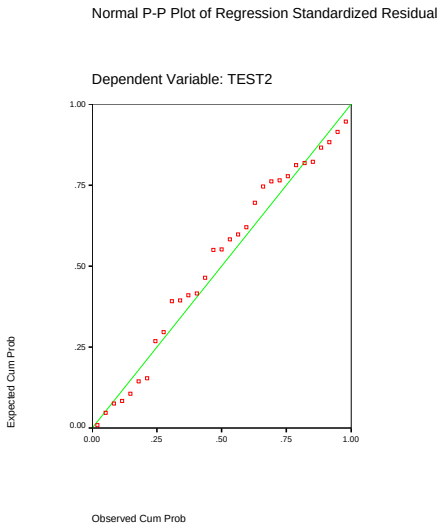


Figure 6-10: SPSS normal plot of residuals



the missing code. Note the only difference between the previous regression and the present regression is that the data file has an extra case at the end with a missing value code for the 2nd column of data.

```
 data list free /  test1 test2.
 begin data.
50 69
66 85
73 88
84 70
57 84
83 78
76 90
95 97
73 79
78 95
48 67
53 60
54 79
79 79
76 88
90 98
60 56
89 87
83 91
81 86
57 69
71 75
86 98
82 70
95 91
42 48
75 52
54 44
54 51
65 73
61 52
68 9999
end data.

  missing value test2(9999).

  regression variables= test1 test2
    /statistics = r anov coeff ci
    /dependent=test2
    /method=enter test1
    /residuals=defaults sepred outliers(cook)
    /casewise=defaults sepred all
    /scatterplot (test2, test1).
```

```
Multiple R           .70307
R Square             .49431
Adjusted R Square    .47687
Standard Error     11.51311

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           1          3757.42041       3757.42041
```

```
Residual            29          3843.99895        132.55169

F =      28.34683       Signif F =  .0000


---------------------- Variables in the Equation -----------------------

Variable               B         SE B     95% Confdnce Intrvl B       Beta

TEST1           .754575     .141726      .464713     1.044438    .703068
(Constant)    22.467092   10.223577     1.557529    43.376655


----------- in ------------

Variable           T  Sig T

TEST1          5.324  .0000
(Constant)     2.198  .0361


Casewise Plot of Standardized Residual

*: Selected   M: Missing
 Case    TEST2       *PRED      *RESID     *SEPRED
       1     69.00     60.1959      8.8041      3.5829
       2     85.00     72.2691     12.7309      2.1701
       3     88.00     77.5511     10.4489      2.0946
       4     70.00     85.8514    -15.8514      2.8033
       5     84.00     65.4779     18.5221      2.8312
       6     78.00     85.0969     -7.0969      2.7096
       7     90.00     79.8148     10.1852      2.2027
       8     97.00     94.1518      2.8482      4.0237
       9     79.00     77.5511      1.4489      2.0946
      10     95.00     81.3240     13.6760      2.3157
      11     67.00     58.6867      8.3133      3.8179
      12     60.00     62.4596     -2.4596      3.2450
      13     79.00     63.2142     15.7858      3.1370
      14     79.00     82.0785     -3.0785      2.3828
      15     88.00     79.8148      8.1852      2.2027
      16     98.00     90.3789      7.6211      3.4352
      17     56.00     67.7416    -11.7416      2.5597
      18     87.00     89.6243     -2.6243      3.3231
      19     91.00     85.0969      5.9031      2.7096
      20     86.00     83.5877      2.4123      2.5357
      21     69.00     65.4779      3.5221      2.8312
      22     75.00     76.0419     -1.0419      2.0684
      23     98.00     87.3606     10.6394      3.0019
      24     70.00     84.3423    -14.3423      2.6202
      25     91.00     94.1518     -3.1518      4.0237
      26     48.00     54.1593     -6.1593      4.5561
      27     52.00     79.0602    -27.0602      2.1580
      28     44.00     63.2142    -19.2142      3.1370
      29     51.00     63.2142    -12.2142      3.1370
      30     73.00     71.5145      1.4855      2.2172
      31     52.00     68.4962    -16.4962      2.4788
      32   9999.00     73.7782           .      2.1015
  Case #    0:........:........:0      TEST2       *PRED      *RESID     *SEPRED
           -3.0      0.0      3.0

[REMAINDER OF OUTPUT OMITTED]
```

From the output we see that the prediction for X = 68 (the new value we added) is 73.78. This is simply the intercept plus the product of 68 and the slope. The value 73.78 is both the prediction of a single ($\hat{Y}$) value and the prediction of the expected value (E($\hat{Y}$)).

The standard error of the expected value is SEPRED = 2.1015 (as seen in the output). The standard error for the case of predicting a single value is, as given in Equation 6-50,

$$\begin{aligned} \text{st. error}(\hat{Y}) &= \sqrt{\text{MSE} + \text{SEPRED}^2} \\ &= \sqrt{132.55 + 2.1015^2} \\ &= 11.70 \end{aligned}$$

The "confidence interval" around the expected value E($\hat{Y}$) is

$$\begin{aligned} \text{E}(\hat{Y}) &\pm t\text{SEPRED} \\ 73.78 &\pm t \times 2.1015 \end{aligned}$$

and the "prediction interval" around the single value $\hat{Y}$ is

$$\begin{aligned} \hat{Y} &\pm t\text{st.error}(\hat{Y}) \\ 73.78 &\pm t \times 11.70 \end{aligned}$$

The t value in both of these intervals is based on the same degrees of freedom associated with the residual term (i.e., in the case of simple linear regression with one slope and one intercept, the degrees of freedom is N - 2, where N is the number of subjects). Obviously, the "prediction interval" is wider (in this case it is much wider) than the "confidence interval." The reason the prediction interval is wider is because the standard error of $\hat{Y}$ is greater than the standard error of $\overline{Y}$, as it should be since in the former we are predicting a single score but in the latter we are predicting a mean.

Appendix 3: Relevant R syntax

R commands for correlation and regression are very simple.

Correlation

To correlate two variables x and y, the R command is (also show how to test a correlation against null hypothesis of 0)

```
> #make up some data; y is linearly related to x with noise epsilon
> x <- rnorm(20)
> y <- -1 + 2*x + rnorm(20,0,2)
> cor(x,y)

[1] 0.6362254

> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 3.4987, df = 18, p-value = 0.002564
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2696120 0.8417558
sample estimates:
      cor
0.6362254
```

See the cor help file (?cor) for more information on details such as how to handle missing data.

More generally, if you have many variables to correlate, assemble them in a matrix and submit that matrix to the cor command as in

```
> #make up some data
> z <- rnorm(20)
> cor(cbind(x,y,z))

          x          y          z
x 1.0000000  0.6362254  0.1389318
y 0.6362254  1.0000000 -0.3097723
z 0.1389318 -0.3097723  1.0000000
```

If you want a covariance matrix (variances in the diagonal, covariances in the off-diagonal) just use the var command instead of the cor command.

Regression

The regression command is lm() and it uses the usual "model" notation. So if y is the dependent variable and x is the predictor, just enter the structural model direction in the lm() command. It is good practice to save the output of the lm() command into an object so you can then operate on that object, such as produce a summary, extract residuals, do a residual plot, etc.

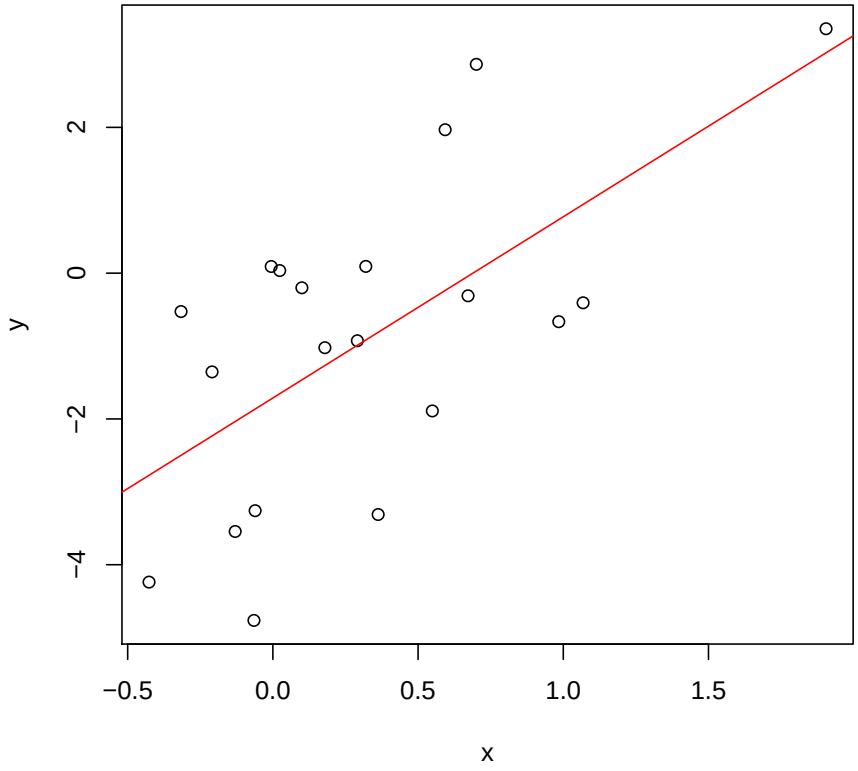
```
> data <- data.frame(x,y)
> lm.out <- lm(y ~ x, data=data)
> summary(lm.out)

Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8891 -1.4164  0.1539  1.3706  2.8362

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.7128     0.4507  -3.800  0.00131 **
x             2.4860     0.7106   3.499  0.00256 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.728 on 18 degrees of freedom
Multiple R-squared:  0.4048,        Adjusted R-squared:  0.3717
F-statistic: 12.24 on 1 and 18 DF,  p-value: 0.002564

> anova(lm.out)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1 36.532  36.532  12.241 0.002564 **
Residuals 18 53.719   2.984
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Scatterplots

R has nice plotting routines that work nicely with the other statistical commands. So it is easy to produce a scatterplot and superimpose the regression line. For example, if I

already ran the regression command above and saved the results into the object that I called "output", I can do one command that produces the plot and a followup command that extracts the coefficients from the saved object "output" and uses the coefficients in the abline() command to produce the regression line. One can also place confidence bands around the regression line, identify outliers, etc.

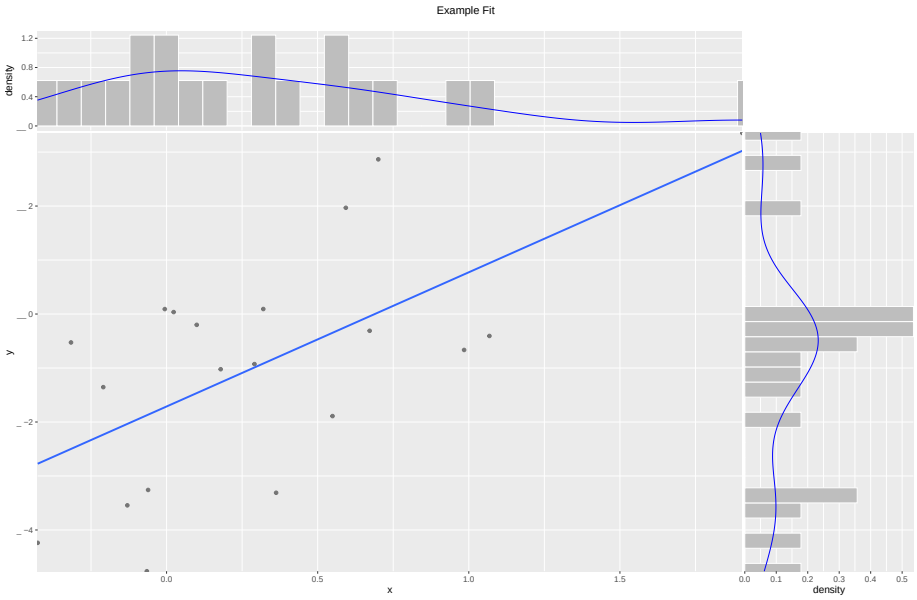
```
> plot(x,y)
> #add regression line
> abline(coef(lm.out), col="red")
```



There are fancier plots one can do with the library ggplot2, and other add-ons people have contributed. Here is one nice plot that combines distributions as well as scatterplots (see http://www.win-vector.com/blog/2015/06/wanted-a-perfect-scatterplot-with-marginals/ but there are many other examples on the web).

```
> #library(devtools)
> #install_github("WinVector/WVPlots")
> library(WVPlots)
```

```
> #adapted from help file
> ScatterHist(data, "x", "y", title="Example Fit",smoothmethod="lm")
```



Prediction

To make predictions save the lm object and use the predict() command. You can get both prediction or confidence interval as well as print standard error.

```
> predict(lm.out, newdata=data.frame(x = 3), interval="confidence", se.fit=T)

$fit
      fit      lwr      upr
1 5.74534 1.672938 9.817741

$se.fit
[1] 1.938388

$df
[1] 18

$residual.scale
[1] 1.727531

> predict(lm.out, newdata=data.frame(x = 3), interval="prediction", se.fit=T)

$fit
      fit       lwr      upr
```

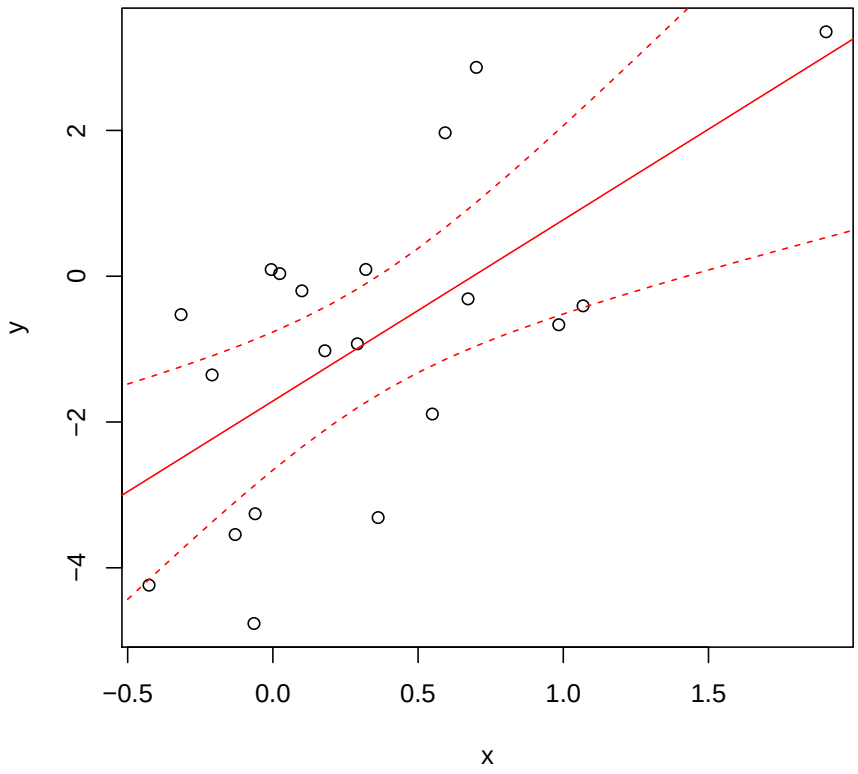
```
1 5.74534 0.2903363 11.20034

$se.fit
[1] 1.938388

$df
[1] 18

$residual.scale
[1] 1.727531
```

You can use this to piece together a graph with regression bands. The function ggplot in the ggplot2 package produces graphs that are much prettier and it provides many more features to overall graphing. Here I show the more direct way to plot a regression line with its error band.

```
> plot(x,y)
> #add regression line
> abline(coef(lm.out), col="red")
> range.predictor <- range(x)
> #generate 20 x scores spanning the range of x
> new.x <- pretty(range.predictor, 20)
> predictions <- predict(lm.out, new=data.frame(x = new.x), interval="confidence")
> predictions

           fit         lwr        upr
1  -2.95576811 -4.43284225 -1.4786940
2  -2.70716503 -4.06199410 -1.3523360
3  -2.45856195 -3.69708756 -1.2200363
4  -2.20995887 -3.33995874 -1.0799590
5  -1.96135579 -2.99306499 -0.9296466
6  -1.71275271 -2.65959913 -0.7659063
7  -1.46414963 -2.34345762 -0.5848416
8  -1.21554655 -2.04886345 -0.3822296
9  -0.96694347 -1.77948375 -0.1544032
10 -0.71834039 -1.53723994  0.1005592
11 -0.46973731 -1.32152448  0.3820499
12 -0.22113423 -1.12946038  0.6871919
13  0.02746885 -0.95698102  1.0119187
14  0.27607193 -0.79993775  1.3520816
15  0.52467501 -0.65474105  1.7040911
16  0.77327809 -0.51854920  2.0651054
17  1.02188117 -0.38921181  2.4329742
18  1.27048425 -0.26513263  2.8061011
```

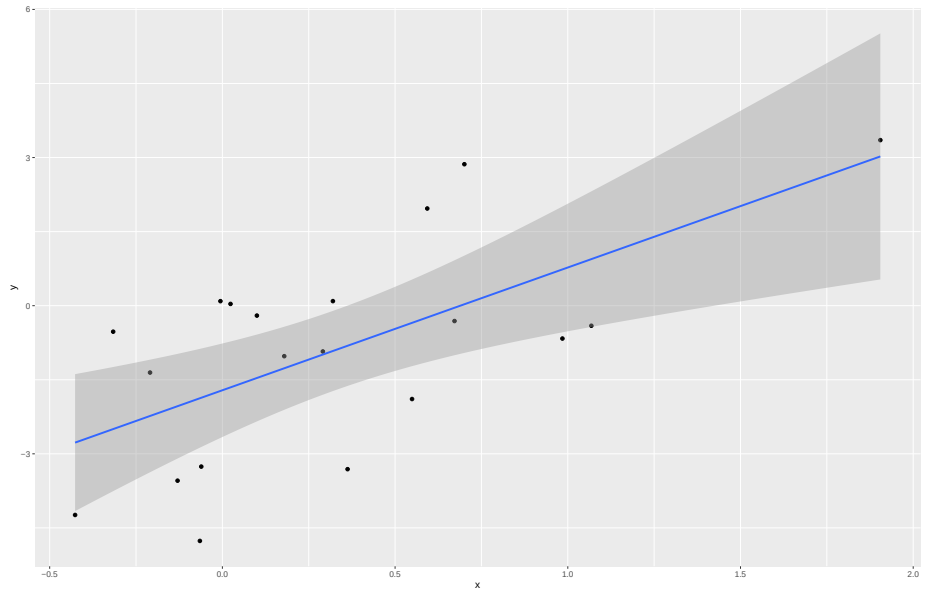
```
19  1.51908733 -0.14513173  3.1833064
20  1.76769041 -0.02833329  3.5637141
21  2.01629349  0.08591858  3.9466684
22  2.26489657  0.19812044  4.3316727
23  2.51349965  0.30865272  4.7183466
24  2.76210273  0.41781041  5.1063950
25  3.01070581  0.52582490  5.4955867
26  3.25930889  0.63287977  5.8857380
```

```
> lines(new.x, predictions[,2], col="red",lty=2)
> lines(new.x, predictions[,3], col="red",lty=2)
>
> #
```



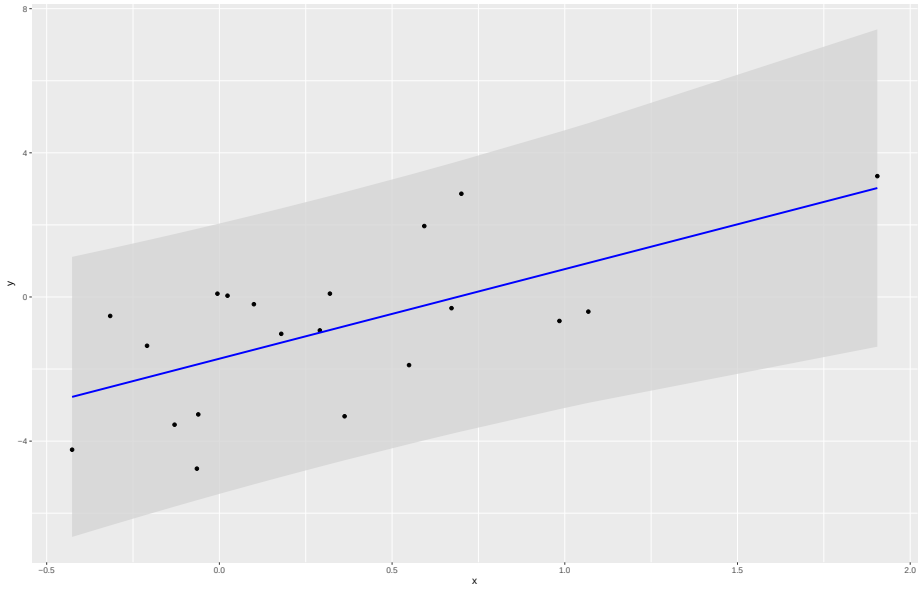
The package ggplot2 offers some more advanced capabilities for graphics. Here is a confidence band around a regression line.

```
> library(ggplot2)
> ggplot(data, aes(x=x,y=y)) + geom_point() + stat_smooth(method="lm")
```



But some things are not built into ggplot2 yet, such as prediction bands around regression lines. One needs to do some computation first and then plot the results of that computation. If you replace the word "prediction" with "confidence" in the code chunk below you have a confidence band plot analogous to the one that is built into ggplot.

```
> out <- lm(y ~ x, data = data)
> # cbind the prediction intervals to the dataframe data
> data.pred <- cbind(data, predict(out, interval = "prediction"))
> ggplot(data.pred, aes(x=x,y=y)) +
+    geom_ribbon(aes(ymin = lwr, ymax = upr),fill="lightgray",alpha=.75) +
+    geom_point() +
+    geom_line(aes(y = fit), color = "blue", size = 1)
```



Bayesian Approach: Regression Example

For a review of the basic Bayesian approach review LN1 and LN2. Here I show an example with a simple regression just using the default settings and priors.

```
> library(rstan)
> library(brms)
> #rstan_options(auto_write = TRUE)
> #to speed up bayesian computation use available cores on your computer
> options(mc.cores = parallel::detectCores())
> #for illustration, using default priors
> out.bayes <- brm(y ~ x, data=data, iter=20000, thin=5)
> summary(out.bayes)

 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: y ~ x
   Data: data (Number of observations: 20)
Samples: 4 chains, each with iter = 20000; warmup = 10000; thin = 5;
         total post-warmup samples = 8000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    -1.71      0.49    -2.69    -0.75       8091 1.00
x             2.49      0.77     0.95     3.99       7881 1.00

Family Specific Parameters:
```

```
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     1.85      0.34     1.34     2.65       7740 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
> plot(out.bayes)
> #prediction, compare to lm prediction interval above
> predict(out.bayes,newdata=data.frame(x = 3))
```

```
     Estimate Est.Error      Q2.5   Q97.5
[1,] 5.783673  2.814804 0.2123825 11.3121
```

```
>
> #plot with 95% shaded; see also the tidybayes package
> #library(bayesplot)
> #mcmc_areas(as.matrix(out.bayes), regex_pars = "x", prob=.95)
```

Figure 6-11: Bayesian Plots from Regression Analyses