Review of Correlation and Regression

A) correlation and regression deal with (usually) numeric variables that covary - i.e., change together

B) measures of how much two variables, X and Y, covary

1) covariance: \( \text{cov}(X,Y) = \frac{(\text{sum}(XY) - \text{sum}(X)*\text{sum}(Y))}{(n - 1)} \), where XY are cross products

2) correlation: \( r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)*\text{var}(Y)}} \)

   a) \( r \) is always in the range -1 to +1
   b) \( r = 0 \) means no linear relationship between X and Y
   c) only works well for linear relationships
   d) scatterplot

C) why things are correlated

1) spurious (www.tylervigen.com/spurious-correlations)
2) direct cause
3) indirect (mediated) cause
4) common cause

D) testing a correlation for statistical significance

1) null hypothesis: the correlation in the population is zero
2) \( t = r \times \sqrt{\frac{N - 2}{1 - r^2}}, \text{df = N - 2} \), where N is the number of pairs of scores

E) variance-covariance matrix

1) variances are on the main diagonal (the diagonal of numbers running from the upper left to the lower right)

2) covariances are values off the main diagonal

> `cov(AGG)`  # Leigh Ann Waslien's data (variances are highlighted)

<table>
<thead>
<tr>
<th>Variable</th>
<th>physag</th>
<th>verbag</th>
<th>anger</th>
<th>hostil</th>
</tr>
</thead>
<tbody>
<tr>
<td>physag</td>
<td>64.75</td>
<td>16.84</td>
<td>36.74</td>
<td>16.02</td>
</tr>
<tr>
<td>verbag</td>
<td>16.84</td>
<td>28.24</td>
<td>36.97</td>
<td>17.13</td>
</tr>
<tr>
<td>anger</td>
<td>36.74</td>
<td>16.17</td>
<td>38.42</td>
<td>17.13</td>
</tr>
<tr>
<td>hostil</td>
<td>16.02</td>
<td>12.85</td>
<td>17.13</td>
<td>38.42</td>
</tr>
</tbody>
</table>
F) Simple regression - one predictor

1) Regression equation: \( \hat{Y} = bX + a \) (predicted \( Y = \) "effect of \( X \)" times \( X \) plus a constant)
   a) \( a \) is the y-intercept (also called the constant)
   b) \( b \) is the slope of the regression line (also called the regression coefficient)
   c) \( \hat{Y} \) is a predicted value of \( Y \) (notice that the equation does not have \( Y \) in it!)

2) To calculate the regression equation
   a) Get the means for both variables
   b) Get the covariance
   c) Get the variance of \( X \) (the predictor or explanatory variable)
   d) \( b = \frac{\text{cov}(X,Y)}{\text{var}(X)} \)
   e) \( a = \text{mean}(Y) - b \times \text{mean}(X) \)

3) We say that "\( Y \) is regressed on \( X \)"

4) Predictions
   a) Fill in a value of \( X \) in the regression equation to get \( \hat{Y} \) for that value
   b) You cannot go the other way (from \( Y \) to \( X \))

5) Simple regression shows the total effect of \( X \) on \( Y \) (provided we can assume cause and effect)

6) Residual - prediction error: \( e = Y - \hat{Y} \)
   a) Positive residual means the prediction is too low
   b) Negative residual means the prediction is too high
   c) Residuals can be thought of as "\( Y \) with the effect of \( X \) removed"

7) We lose one degree of freedom for each term that must be estimated, thus: \( \text{df} = N - k - 1 \), where
   \( N \) is the number of pairs of scores and \( k \) is the number of predictors

8) The residual standard error is the standard deviation of the residuals calculated using \( \text{df} \) degrees of freedom - also called the standard error of estimate

9) \( R^2 \) (the square of the correlation) is the proportion of the variability in \( Y \) explained by \( X \)

10) Assumptions of linear regression
    a) Linear relationship - check with a residuals plot
    b) Normally distributed residuals - check with a normal probability (or QQ-norm) plot
    c) Homoscedasticity - check with a scale location plot (or look for a wedge in residuals plot)
    d) No influential cases - check with a Cook's distance plot

G) Dummy (0/1) coding - when a categorical variable has two levels

1) Types of "correlations"
   a) Two numeric variables - Pearson \( r \)
   b) A dummy-coded variable with a numeric variable - point-biserial correlation; related to the \( t \)-test: \( r^2 = t^2 / (\text{df} + t^2) \)
   c) Two dummy-coded variables - phi coefficient; related to the chi-square test of independence: \( \phi = \sqrt{\chi^2 / N} \)
2) simple regression using a dummy-coded IV
   a) the intercept (a) is the mean of the group coded 0
   b) the slope (b) is the difference between the means of the two groups
   c) therefore, the mean of the group coded 1 is \( a + b \)
   d) the test on the slope gives the same result as an independent measures t-test (two-tailed)

H) multiple regression - more than one predictor variable
   1) multiple regression produces a Y-intercept (constant) and a slope (regression coefficient) for each of the predictors
   2) each variable (predictor) is treated as if entered last - every predictor effect is controlled for all the other predictors in the model
   3) since all the indirect effects of other variables are removed, standard multiple regression shows direct effects (provided we can assume causal relationships or a "flow of influence")
   4) the regression equation: \( Y\text{-hat} = a + b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + \ldots \)
      a) \( b_1 \times X_1 + b_2 \times X_2 + b_3 \times X_3 + \ldots \) produces a composite variable and \( Y \) is regressed on this composite (at least that's one way of thinking about it)

5) results in R for Scottish Hill Racing regression (uncorrected, no interaction term)

Call:
\[ \text{lm(formula = time} \sim \text{dist + climb, data = hills)} \]

Residuals:
\[ \begin{array}{cccccc}
\text{Min} & \text{1Q} & \text{Median} & \text{3Q} & \text{Max} \\
-16.215 & -7.129 & -1.186 & 2.371 & 65.121 \\
\end{array} \]

Coefficients:
\[ \begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
(Intercept) & -8.992039 & 4.302734 & -2.090 & 0.0447 * \\
dist & 6.217956 & 0.601148 & 10.343 & 9.86e-12 *** \\
climb & 0.011048 & 0.002051 & 5.387 & 6.45e-06 *** \\
\end{array} \]

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.68 on 32 degrees of freedom
SD of residuals, df = 32
Multiple R-squared: 0.9191, Adjusted R-squared: 0.914
prop. expl. variability
F-statistic: 181.7 on 2 and 32 DF, p-value: < 2.2e-16 test on R-squared
a) interpretation of intercept - the predicted value of Y when all predictors are set to zero
b) interpretation of coefficient - the predicted change in Y when X is increased by 1
c) regression coefficients cannot be directly compared unless the predictors are measured on the same scale (i.e., have the same variance)
d) beta coefficients are obtained when all variables are converted to z-scores and the regression is run on those values
i) beta = b * sqrt(var(X) / var(Y))
ii) they can be directly compared, approximately
iii) never convert a dummy variable to z-scores!

6) dropping terms from a regression analysis
   a) terms with nonsignificant coefficients can be dropped and the regression redone
   b) never drop more than one term at a time

7) interactions - when changing the value of one variable changes the effect (coefficient) of the other variable, we say an interaction is occurring between the two variables
   a) always check for them - if it doesn't exist, redo the regression without the interaction term before you try to interpret the coefficients
   b) interaction is different from mediation effects and confounding
   c) interactions are often referred to as moderator effects
   d) if the interaction is retained in the model, lower order terms that are part of the interaction cannot be dropped even if they are not significant
   e) an interaction is modeled in the regression equation by multiplying the interacting variables

8) linear transform on a predictor will change the coefficient of that predictor, but it will not change anything else about the regression outcome

I) hierarchical regression (also called sequential regression)
   1) according to Keith (Multiple Regression and Beyond, 1st ed., 2006), hierarchical regression is appropriate when we can diagram the causal flow among the variables in a way similar to the following (Keith's book is highly recommended if you want more about regression)

2) that is, variable A has a (possible) direct effect on the response plus (possible) indirect effects through some or all of the other variables, variable B has a (possible) direct effect on the response plus (possible) indirect effects through some or all of the remaining variables, etc.

3) the interest in hierarchical regression is in seeing if adding each predictor to the regression in a sequential fashion significantly increases the R-squared value

4) in R, hierarchical regression can be done by doing several sequential standard regressions, or it can be approximated using the aov() function

5) if the order in which certain variables should be entered is uncertain, they can be entered in a block (you'll have to do some additional arithmetic in R to get this to work)
6) hierarchical regression shows total effects provided our causal model is correct
7) here is a hierarchical regression from Keith (2006) done in SPSS (the response was achievement test scores of high school students)

<table>
<thead>
<tr>
<th>Added to the Model</th>
<th>R</th>
<th>R Square</th>
<th>R Square Change</th>
<th>F Change</th>
<th>df1</th>
<th>df2</th>
<th>Sig.F Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>SES</td>
<td>0.430</td>
<td>0.185</td>
<td>0.185</td>
<td>200.709</td>
<td>1</td>
<td>885</td>
<td>0.000</td>
</tr>
<tr>
<td>Previous Grades</td>
<td>0.573</td>
<td>0.328</td>
<td>0.143</td>
<td>188.361</td>
<td>1</td>
<td>884</td>
<td>0.000</td>
</tr>
<tr>
<td>Self Esteem</td>
<td>0.577</td>
<td>0.333</td>
<td>0.005</td>
<td>6.009</td>
<td>1</td>
<td>883</td>
<td>0.014</td>
</tr>
<tr>
<td>Locus of Control</td>
<td>0.582</td>
<td>0.339</td>
<td>0.006</td>
<td>7.918</td>
<td>1</td>
<td>882</td>
<td>0.005</td>
</tr>
</tbody>
</table>

8) here is the same analysis done in R using the aov() method

```r
> print(anova(lm.out), digits=7)  # equivalent to using aov() to start with
Analysis of Variance Table
  Response: std.achieve.score
             Df Sum Sq Mean Sq  F value    Pr(>F)
SES          1 15938.6  15938.6 246.4953 < 2.22e-16 ***
prev.grades  1 12344.6  12344.6 190.9127 < 2.22e-16 ***
self.esteem  1   391.6   391.6   6.0566  0.0140453 *
locus.of.control 1  512.0  512.01  7.9182  0.0050028 **
Residuals    882 57031.0    64.66  
```

9) notice that they are not quite the same; I will demonstrate in class how to get from one to the other (hint: SS tot = 86217.92); I’ll leave some space for your notes (write small!)

J) simple mediation analysis

1) simple mediation analysis is appropriate when we can diagram the causal flow among three variables in the following way (the variables must all be significantly positively correlated)

```
X ---- M ---- Y (two parts of indirect effect of X on Y)
X ---- Y (direct effect of X on Y)
```

```
X ---- Y (total effect of X on Y (sum of above))
```
a) total effect is obtained by regressing Y on X (the coefficient of X is the total effect)
b) direct effect is obtained by regressing Y on M and X (the coefficient of X is the direct effect)
c) the two parts of the indirect effect are obtained by:
   i) regressing M on X (the coefficient of X is the direct/total effect of X on M)
   ii) regressing Y on M and X (already done; the coefficient of M is the effect of M on Y)
d) the indirect effect is then obtained by one of the following (equivalent) methods
   1) total effect minus direct effect
   2) multiplying the two parts of the indirect effect

2) significance tests
   a) the test of the total effect is given by the test on the coefficient of X in part (a) above
   b) the test of the direct effect is given by the test on the coefficient of X in part (b) above
   c) the test of the indirect effect is given by calculating Sobel's z: if A and B represent the
      two parts of the indirect effect, then \( z = \frac{AB}{\sqrt{A^2 \cdot SE(B)^2 + B^2 \cdot SE(A)^2}} \), where
      the standard errors (SE) are obtained from the regressions
   d) critical \( z = 1.96 \) (requires a large sample of \( N = 100 \) or more to be reasonably accurate)

3) possible conclusions from a mediation analysis

<table>
<thead>
<tr>
<th>direct effect</th>
<th>indirect effect</th>
<th>conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>not significant</td>
<td>significant</td>
<td>total mediation</td>
</tr>
<tr>
<td>significant</td>
<td>significant</td>
<td>partial mediation</td>
</tr>
<tr>
<td>significant</td>
<td>not significant</td>
<td>no mediation</td>
</tr>
</tbody>
</table>

K) regression and cause and effect
   1) correlation alone does not necessarily imply causality
   2) to suggest a causal relationship, we need additional information
      a) theoretical
      b) other studies - i.e., our thorough review of the literature
      c) control of possible confounds
      d) common sense

L) analysis of covariance
   1) ANCOVA is a method that allows removal of the effect of a numeric confound (the "numeric
covariate") before an ANOVA is done with a categorical IV
   2) ANCOVA can be done by using ANOVA (aov() in R) or regression (lm() in R) - it makes more
      sense to me to use regression (if by ANOVA, make sure you enter the numeric covariate first!)
   3) adjusted means - once the ANCOVA is calculated, means can be calculated for the groups
defined by the categorical IV that have been adjusted for the confound created by the
numeric covariate (one way to do this is to substitute the overall mean of the covariate
into the regression equation, then calculate a Y-hat for each of the groups)

K) regression and cause and effect
   1) correlation alone does not necessarily imply causality
   2) to suggest a causal relationship, we need additional information
      a) theoretical
      b) other studies - i.e., our thorough review of the literature
      c) control of possible confounds
      d) common sense

L) analysis of covariance
   1) ANCOVA is a method that allows removal of the effect of a numeric confound (the "numeric
covariate") before an ANOVA is done with a categorical IV
   2) ANCOVA can be done by using ANOVA (aov() in R) or regression (lm() in R) - it makes more
      sense to me to use regression (if by ANOVA, make sure you enter the numeric covariate first!)
   3) adjusted means - once the ANCOVA is calculated, means can be calculated for the groups
defined by the categorical IV that have been adjusted for the confound created by the
numeric covariate (one way to do this is to substitute the overall mean of the covariate
into the regression equation, then calculate a Y-hat for each of the groups)