BIOL811 Applied Biostatistics II

Complete the activities using the MS Excel and the ANCOVA.xlsx workbook. Answer the questions in the spaces provided on the worksheet of ANCOVA.xlsx. Upload the completed workbook and your Systat and JMP output files using the ANCOVA Assignment Submission link on the course Blackboard site.

Part 1: Analysis of Covariance (ANCOVA)

Recap: We have explored how to analyze the results of an experiment with more than one categorical treatment (i.e. a factorial design) using a two-way ANOVA. We have also looked at analyzing results from studies with a single quantitative treatment variable using Simple Linear Regression.

Objective: It is possible to design experiments with one (or more) categorical treatment(s) plus one (or more) quantitative (numeric) factor(s). In some of these experiments, the quantitative factor is included for the same reason that a blocking factor is included in an RCB design. That is, simply to remove some of the variability from the error term so that the effect of the categorical treatment is easier to detect. Other times, the quantitative factor is of interest and it is included in the analysis so that a hypothesis can be tested related to it. In either case, an analysis that includes categorical and quantitative experimental factors is often referred to as an Analysis of Covariance, and the quantitative variable is called a “covariate”.

Activity 1: The ANCOVA.xlsx workbook contains fictional systolic blood pressure data for a group of male and female human volunteers. On the Blood Pressure ANOVA tab of the workbook is a template for doing a "one-way" (meaning it has only one treatment group) ANOVA to determine if there is a difference in the average blood pressure for males versus females. By now, you should be able to complete the ANOVA on your own. When you have, type the answers to the following questions at the bottom of the Excel worksheet.

Questions 1: What is the null hypothesis that you are testing with this ANOVA?

Questions 2: Based on your ANOVA, what would you conclude?

Activity 2: The Blood Pressure Regression tab of ANCOVA.xlsx shows the same BP data plotted against the Age of the subject. It also includes a template for a Simple Linear Regression Analysis of BP versus Age. Complete the Regression. Type the answers to the following question in the spaces provided at the bottom of the Regression template.

Questions 3: What is the null hypothesis that you are testing with the Linear Regression Analysis?

Questions 4: Based on the results of your Regression, does Age affect BP?

Activity 3: The Blood Pressure ANCOVA tab of the workbook shows a plot of the BP data against Age, but the points for the male subjects are in a different color than the points for the females. It also includes a template for an Analysis of Covariance (ANCOVA). As you will see, it simply combines the deviations for the Male and Female groups with the deviations due to age. The Means and Deviations Table from your ANOVA has been copied to this page, as have the parameters for the best-fit line that you calculated for the Regression. You should also notice that the Analysis of Covariance worksheet is very
similar to the one you used for the RCB ANOVA. The difference is that the calculations for the Age deviations are the distance from each point to the regression line.

To complete the analysis:

a) Calculate the yhat values the same way you did for the Regression analysis
b) Fill in the Deviations column for Gender using the appropriate values from the Means and Deviation table.
c) Calculate the Age deviations by subtracting the overall mean (ybar) from yhat values.
d) The equation for this analysis is BP = Mean BP + DevGender + DevAge + DevError. To calculate the Error Deviation in each row, subtract the Mean BP, the DevGender and the DevAge from the BP measurement.
e) Square all of the Deviations into the Deviations² columns and add up each column to get the Sums of Squares for Gender, Age and Error.
f) Fill in the ANCOVA Table. Calculate p-values for Gender and Age.

Question 5: What null hypotheses are you testing with this analysis?

Question 6: Based on your ANCOVA, can you reject either or both of the null hypotheses?

Question 7: Why did the ANCOVA find a significant effect of Gender, when the ANOVA didn't? Did the SSGender from the ANOVA change in the ANCOVA? Did the SSAge from the Regression change in the ANCOVA? How about the SSError?

Activity 4: Repeat the ANCOVA in Systat. Start by importing the Gender, Age and BP data from Excel into a new Systat data sheet. The cut and paste method is probably easiest, but you will need to edit the column headings. Remember, Gender is a categorical variable, so the column heading will need to end with a dollar sign. Age is a numeric (quantitative) variable, so it should not end with a dollar sign.

Pull down Analyze -> General Linear Model -> Estimate Model. Drag BP to the Dependent(s) box. Drag Gender$ and Age to the Independent(s) box. Note that we have one categorical and one quantitative Independent variable. Press OK.

Compare the Systat ANOVA table to the ANCOVA Table you completed in Excel.

It is possible to test for an Interaction between a quantitative and categorical variable. The hypothesis you would be testing in this study would be that the effect of Age is independent of Gender. Technically you will be testing to see if the slope of the regression line for Males is the same as the slope of the regression line for Females.

Try this in Systat. Pull down Analyze -> General Linear Model -> Estimate Model. BP should already be in the Dependent(s) box and Gender$ and Age should already be in the Independent(s) box. Add Age to the Independent(s) box a second time, then select Gender$ in the Available Variables box and press Cross. Click OK. Look at the Gender$*Age interaction term in the Analysis of Variance Table.

Question 8: Based on your Systat analysis, does age have a different effect on BP in Males than it does in Females?

Save your Systat output to hand in.
Activity 5: Repeat the ANCOVA in JMP. Copy your data from Excel and use File->New->Data Table in JMP. Paste the data into the new table and edit the column headings (double-click on the heading). You should have three columns of data. In the middle block to the left of the data, you will see the names of the 3 columns. Gender will have a bar chart icon, which means it contains "categorical" data. The other two column names (Age and BP) are labeled with a blue triangle, which indicates it contains numeric data.

Pull down the Analyze menu and select Fit Model. In the Fit Model window, select BP in the Select Columns box. Click on Y in the Pick Role Variables box. Click on Gender and add it to the Construct Model Effects box. Add Age to the Construct Model Effects box. Click on Gender in the Select Columns box and holding down the Ctrl key, click on Age. Release the Ctrl key and click on Cross in the Construct Model Effects box. This should create the interaction term Gender*Age in the Construct Model Effects box.

Leave all of the other options as they are and press Run.

To simplify the output display, leave the Regression Plot, but close the Leverage Plots, the Actual by Predicted Plot and the Residual by Predicted Plot by clicking on the little gray triangles in their upper left corners.

In the Summary of Fit display, the RSquare value (like Multiple Squared R in Systat) indicates the percentage of the variability in the BP measurements that is accounted for (explained) by the combined effects of Gender, Age and the Gender by Age interaction.

The third row of the Analysis of Variance display (labeled C. Total) shows the Total Degrees of Freedom and Total Sum of Squares (SS) for the ANOVA. The second row is the DF, Sum of Squares (SS) and Mean Square (MS) for the Error term. The first row is the combined DF, SS and MS for Gender, Age and Gender by Age, which are detailed in the Effects Tests display (you may need to open this).

On the right side of the window are the Means and Standard Errors for each level of Gender.

Compare the JMP p-values to you Systat output.

Save your JMP output to hand in.

Part 2: Non-linear Regression in Systat and JMP

Recap: For studies with a quantitative experimental factor, you learned how to fit a straight line that minimizes the sum of the squared deviations (SS\text{Error}) between the line and the data points. You calculated the SS\text{Treatment} by adding up the squared distances between the overall mean and points on the line corresponding to each data point. From the SS\text{Treatment} and SS\text{Error}, you were able to calculate the MS\text{Treatment} and MS\text{Error} and an F-ratio (MS\text{Treatment}/MS\text{Error}). The p-value associated with that F-ratio is the probability the treatment has no effect on the response. You also saw that it is possible to fit a straight line through data and get a p-value less than 0.05 even when the relationship
between the treatment and response is clearly not linear. A Simple Linear Regression is not a test for linearity.

Objective: In studies where the relationship between the experimental factor and the response can be represented by a curve, and where a mathematical formula for the curve can be defined, it is possible to fit that curve to the data. This is done by adjusting the parameters of the formula to minimize the sum of the squared deviations between the curve and the data points ($SS_{\text{Error}}$). This is analogous to adjusting the slope and intercept of a straight line to find the best fit.

The first objective of this lesson is to use Systat and JMP to fit curves to data. The second is to determine whether the curves fit through two sets of data are similar or significantly different.

Activity 1: On the Bayberry PS Data sheet of the ANCOVA.xlsx workbook are the results of a field/lab study on the effects of light level on the photosynthetic rate of bayberry plants (*Myrica pensylvanica*) that are acclimated to full sunlight or to shade. Up to a point, the photosynthetic rates of plants increase with greater light. A mathematical model that often can be made to fit photosynthesis vs light data fairly well is as follows:

$$PS = \frac{P_{\text{max}} \cdot \text{Alpha} \cdot \text{Light}}{P_{\text{max}} + \text{Alpha} \cdot \text{Light}}$$

Where $PS$ is the photosynthetic rate (mgO$_2$/m$^2$/h), Light is light level ($\mu$moles/m$^2$/s), $P_{\text{max}}$ is the maximum photosynthetic rate (when the plant has all the light it can use) and Alpha is the initial slope of the curve. Alpha is an indication of how the plant responds in low light levels.

Copy the 3 columns of data from the Bayberry PS Data worksheet and paste it into a new data table in Systat. The light level is in the first column (change the column name to "Light"). The photosynthetic rate measurements for the sun acclimated plants are in the second column (name it "Ps_Sun") and the shade adapted plant photosynthesis measurements are in the third column ("Ps_Shade").

Before doing the analysis, make sure the Systat automatic graph output is turned on. Pull down the Edit menu and select Options. In the navigation box at the left, select Output. On the right side of the Output dialog box, check Display Statistical Quick Graphs. (It's possible it is already checked). While you are in this dialog box, also make certain that the Output Results Length is set to Medium and the number of Decimal places is 3. Click OK.

Pull down the Analyze menu and select Regression -> Nonlinear -> Estimate Model (or in the Systat shortcut menu, you can click the icon that looks like a scatter plot with a blue curved line through the data). Drag PS_SUN from the Available variable(s) box into the Dependent box. In the Model expression: box, enter the formula from above. It should look like this $P_{\text{max}} \cdot \text{Alpha} \cdot \text{Light}/(P_{\text{max}} + \text{Alpha} \cdot \text{Light})$. To get Systat headed in the right direction, you will need to supply a rough guess for each of the parameters ($P_{\text{max}}$ and Alpha) in the formula. To do that, click on Options (left side of the dialog box) and in the Starting values: box enter 300,1 or whatever guesses you want to try for $P_{\text{max}}$, Alpha. Click Model (on the left side of the dialog box). Click OK. If all you see is a Scatter Plot, click on the *Untitled.syo tab at the top (or double-click on the yellow icon in the Workspace window labeled “MODEL PS_SUN=Pmax...”.) At the bottom, you
should see a plot of the data with the line that Systat fit by adjusting the parameters in the formula you gave it. Scroll up to the Parameter Estimates table. You will see the values that Systat estimated for Pmax and Alpha along with their standard errors (labeled ASE). Record the Pmax and Alpha estimates in the spaces provided on the "Curve Comparison" worksheet of the ANCOVA.xlsx workbook. In Systat, above the Parameter Estimates table you will find the $r^2$ values. In this case, Light (actually the curve) explains nearly all of the variation in the photosynthesis. In the Sum of Squares box, notice that the $SS_{Error}$ (the row labeled Residual) is very small compared with the $SS_{Total}$. Copy the $SS_{Error}$ value and paste it in the space provided on the "Curve Comparison" worksheet of the ANCOVA.xlsx workbook.

Repeat the analysis for the photosynthesis measurements from the shade acclimated plants. All you will need to do is open the Nonlinear Regression dialog box and replace PS_SUN in the Dependent: box with PS_SHADE and click OK. Scroll up to the Sum of Squares and Parameter Estimates tables. Copy the parameter estimates for Pmax and Alpha and the Residual SS ($SS_{Error}$) and paste them in the spaces provided on the "Curve Comparison" worksheet of the ANCOVA.xlsx workbook.

Activity 2: From the parameter estimates for the Sun and Shade acclimated plant data, it should be apparent that the curves are different. In other word, the light under which the plants are grown affects it photosynthesis vs light relationship. In other instances, this difference may not be so obvious. In those cases, we can do an F-test to test the null hypothesis that the two curves are not really different. To see how this works, we will do the analysis on the two curves we just fit. The logic of the test goes like this: if the photosynthesis versus light response of the shade and sun adapted plants is the same, then the amount of variability explained by fitting two separate curves to the data will be no greater than the amount explained by fitting one curve through all of the data. So what we need to do now is fit a single curve through the sun and shade data combined.

Click on the *Untitled1.syz* tab in Systat (or pull down View and select Data Editor). Copy all of the Light data and paste it at the bottom of the Light column (starting at row 23). You should now have two copies of the Light values. Copy all of the data in the PS_SHADE column and paste it at the bottom of the PS_SUN column (starting in row 23). Double-click on the PS_SUN heading and edit it to say PS_ALL. Right-click on the 3 at the left of Row 3 (the PS_SHADE row) and select Delete Variable.

Open the Nonlinear Regression dialog box (use the icon or Analyze -> Regression -> Nonlinear -> Estimate Model). The Model Expression should still be there. Remove PS_SHADE from the Dependent: box and replace it with PS_ALL. Click OK. The Scatter Plot will show the curve that Systat has fit through the combined data sets. Scroll up to the Sum of Squares and Parameter Estimates tables. Copy the parameter estimates for Pmax and Alpha and the Residual SS ($SS_{Error}$) and paste them in the spaces provided on the "Curve Comparison" worksheet of the ANCOVA.xlsx workbook.

Save your Systat output to hand in.
The “Curve Comparison” worksheet contains a template for an F-Test to determine what is referred to as “the coincidence of two regression lines”. Essentially we are going to test the null hypothesis that fitting two separate lines through the photosynthesis versus light data will not explain any more variability than a single line through both sets of measurements. It also includes a scatter plot to help visualize the analysis.

You have already filled in the Parameters and SS\text{Error} table using the data you recorded from the Systat Nonlinear Regressions. The row labeled “n” is the number of data points in the respective data sets. Fill in the values for n. Below the table, fill in the value for m, which is the number of parameters that were fit in each regression. In this case it was Pmax and Alpha, so m=2. Fill in the box for k, which is the number of curves you are comparing, which would be the Sun curve and the Shade curve.

Below this is what at first may appear to be a daunting formula, but if you look at each term separately, it’s really not all that complicated. The top line is $\text{SSE}_{\text{All}} - (\text{SSE}_{\text{Sun}} + \text{SSE}_{\text{Shade}})$. This is the amount of variability accounted for by splitting the data into two groups (Sun and Shade) compared to lumping it all together. Do this calculation in cell F28. All you need to do is add the Sun and Shade SSE values from your Parameters and SS\text{Error} table and subtract them from the SSE value in the column labeled All, so the Excel equation in F28 will be =F22-(C22+D22). In F29, you will calculate the df for the numerator of the F-ratio, which is $(m+1)*(k-1)$. If you enter the formula correctly, you should get 3. F30 is the sum of the Sun and Shade SSE values. Finally, F31 is the df for the denominator of the F-ratio. Finish the F-ratio calculation in H29 using $=\left(\frac{\text{F28}/\text{F29}}{\text{F30}/\text{F31}}\right)$. Use the FDIST function to calculate $p$ in cell F33.

**Question 1:** Are the two curves significantly different?

**Question 2:** Do Sun and Shade adapted Bayberry plants respond differently to increasing light?

Save your ANCOVA.xlsx workbook to hand in.

**Activity 3:** Nonlinear regression in JMP is a little more complicated. We’ll do one for just the Sun adapted plant data.

Open JMP and close the Tip-of-the-Day window. Pull down File -> New -> Data Table. Copy and Paste the 3 columns of data (not the headings) from the Bayberry PS Data worksheet of ANCOVA.xlsx. You should have three columns of data. Double click on where the headings and name the columns Light, PS_Sun and PS_Shade. Double click on where the heading would go for a fourth column. This will create a column with the heading Column 4. Right-click on the Column 4 heading and select Formula... This will open a dialog box. Pull down the Table Columns menu in the upper left of the dialog box and select Parameters. Click on New Parameter... Enter “Pmax” as the name and 300 as its value. Click OK. Click on New Parameter again and enter “alpha” as the name and 1 as its value. Click OK. Click on the red box that says No Formula. Press the left parenthesis key then click on Pmax in the Parameters box. Press the * key. Click on alpha in the Parameters box. Press the * key. Pull down the heading of the Parameters box and select Table Columns. Click on Light. In the formula that is being created, click on the
right parenthesis. Press the / key. This should put everything you created so far on the top half of a fraction. Press the ( key. Pull down the Table Column header again and select Parameters. Click on Pmax. Press the + key. Click on alpha in the Parameters box. Press the * key. Pull down the heading of the Parameters box and select Table Columns. Click on Light. Press the right parenthesis key. Click OK. Column 4 should fill with values that are more or less like the values in the PS_Sun and PS_Shade columns.

Pull down the Analyze menu and select Modeling -> Nonlinear. Put PS_Sun in the Y, Response box and put Column 4 in the X, Predictor Formula box. Click OK. It will draw a curve through the data using the parameter estimates that you provided (Pmax=300 and alpha=1) when you made Column 4. To have JMP find the parameter values for the best fit curve, press the Go button.

Compare JMP's parameter estimates and SSE value to the ones from Systat for the PS_SUN data. Save your JMP output to hand in.

Assignment: Hand in your completed MS Excel workbook and your Part 1 and Part 2 Systat and JMP output files.