

Goodness-of-Fit, Regression, and ANOVA**Objectives:**

1. To practice goodness-of-fit tests
2. To practice using JMP for multiple regression analysis and Analysis of Variance

Getting Started: There are two files to download today. Log onto your machine and download the `icecream.txt` and `hotdogs.txt` datafiles from the class webpage:

<http://www.soe.ucsc.edu/classes/ams007/Fall107/> After downloading these files, start JMP.

Part I. Goodness-of-Fit Tests

Let's check that JMP's random number generator is fair. From the JMP Starter window, create a "New Data Table". Double-click on the header for `Column 1` to bring up the column info window. We'll generate 100 random integers from 1 to 4. To do this, go to the line for `Initial Data Values` and where the drop-down menu says `Missing/Empty`, click and change that to `Random`. By default, `Random Integer` should already be selected, and it should be set to 100 rows. However, by default is it going to make random numbers from 1 to 100. We only want numbers from 1 to 4, so where it says `between 1 and 100`, change that 100 to 4, then hit "OK". You should now have 100 rows of numbers with values from 1 to 4.

We will treat the values of this variable as categories, so tell JMP to think of this variable as nominal by clicking on the blue triangle on the left side of the window (next to `Column 1`) and choose `Nominal`. To begin the analysis, go to the JMP Starter window, and choose `Basic` and "Distribution".

Question #1 How many of each category did you get?

Question #2 How many did you expect to get in each category?

Question #3 What are your hypotheses for testing whether the random number generator is fair?

Question #4 Compute your test statistic by hand.

Question #5 What is the sampling distribution of the test statistic? (Which distribution, how many degrees of freedom?)

Question #6 The critical value is 7.815. What do you conclude?

Now we'll do the analysis using JMP. From the distribution analysis window, go to the Hot Spot by **Column 1** and choose **Test Probabilities**. A new section should appear at the bottom of the window. Enter your hypothesized probabilities into the boxes and click on "Done". In the table of **Test** information that appears, the line labeled **Pearson** is the goodness-of-fit test we have learned in this class.

Question #7 What is the test statistic computed by JMP? Does it match your answer to Question #4?

↪ **Question #8** What is the p-value given by JMP? (It's in the right-most column, labeled **Prob>Chisq**.) Is this consistent with your answer to Question #5?

Part II. Multiple Regression

Open the `icecream.txt` file, which contains weekly data on ice cream consumption (IC) in pints per capita. Possible independent (explanatory) variables are the price of ice cream per pint in dollars, the mean temperature in degrees F, and the year (0=1951, 1=1952, 2=1953).

Start by making a scatterplot matrix of the data. From the JMP Starter window, go to **Multivariate** and "Multivariate", and put all of the variables in as "Y, Columns", and hit "OK". You should get a bunch of plots, in fact all possible plots of one variable versus another. So in the top row are plots with ice cream consumption on the Y axis, and each of the possible independent variables on the X axes. In the leftmost column are the mirror images of the same plots, all with consumption now on the X axis.

Question #9 Looking at the top row of plots, do any of the independent variables appear to be highly correlated with ice cream consumption? (If so, state which one(s).)

Question #10 Do any of the "independent" variables appear to be correlated with each other? (Which one(s)?)

Question #11 What problems might we face if we include two highly correlated independent variables in a regression model?

If you hadn't already noticed, above the scatterplot matrix is the correlation matrix, which gives the correlations between each pair of variables.

Next we'll do a linear regression analysis. Start by fitting a model predicting ice cream consumption from all four independent variables. From the JMP Starter window, go to **Model** and "Fit Model", put **IC** as "Y", then put each of the independent variables in the bottom box (**Construct Model Effects**) by clicking on each one and then clicking on **Add** (or click on **price**, then shift-click on **Year** to highlight all four at the same time, then click on **Add** and it will add all four together). Once you have all four variables in the **Effects** box, click on the button for "Run Model" near the upper right. You should get several plots and tables of information about the regression model. The upper left plot shows the predicted values on the X axis and the observed values on the Y axis. If the fit was perfect, then all the points in this plot would fall along the $Y=X$ line, shown as the solid red line. If the model was useless, then the points would be randomly scattered around the mean of the observed values, shown by the dashed blue line. The dashed red lines show confidence intervals for predictions. We're not going to use this plot in this class, so you can hide it by clicking on the grey and blue diamond to its upper left (just left of **Actual by Predicted Plot**).

The other plots to the right are plots of leverage for each of the independent variables. Leverage is one possible measure of influence (influence is briefly discussed in Section 9.3 of the text). We're going to ignore these plots, so you can hide them by clicking on their grey and blue diamonds.

Question #12 The next section is **Summary of Fit**. What are the R^2 and Adjusted R^2 for this model?

Question #13 The **Analysis of Variance** section is for testing the fit of the overall model. What are the hypotheses being tested?

Question #14 The test statistic is an F, with degrees of freedom as shown. What is the p -value computed by JMP (shown under **Prob > F**)?

Question #15 What do you conclude about the overall model fit?

Question #16 The **Parameter Estimates** section has the fitted slope and intercept coefficients, as well as the p -values for testing the significance of the individual coefficients. What is the fitted regression equation?

Question #17 What are the hypotheses for testing if the slope coefficient for price is significant?

Question #18 What is the p -value for this test, and what do you conclude?

Question #19 Which variable is least significant?

↪ **Question #20** Below is the plot of residuals versus predicted values. Do you see any major problems in the residual plot?

We can only remove one variable at a time using the t tests, so we take out the least significant one, and then re-run the model. If there is any correlation between the independent variables, the coefficients and the p -values for the remaining variables may change. Re-run the model without the income variable by going back to the **Fit Model** window (which should be still open), clicking on **income** in the **Effects** box, then clicking on the “Remove” button and hitting “Run Model”.

Question #21 How much do the R^2 and Adjusted R^2 change?

Question #22 Is the overall model still significant?

Question #23 Interpret the meaning of the slope coefficient for `temp` in the context of this problem.

Question #24 Are any of the variables still not significant? Which one(s)?

Question #25 Continue to remove variables one-at-a-time and re-run the model each time, until all of the variables are significant. What is your final regression equation?

Question #26 What is the predicted ice cream consumption for an average temperature of 68 degrees in 1952?

Question #27 Of the independent variables that are in this model, are there any highly correlated with each other? (The pairs to look for are those from Question #10.) We would hope that there aren't any large correlations left by this point.

Part III. Analysis of Variance

Open the file `hotdogs.txt`, which contains a laboratory analysis of the calorie and sodium content of major hot dog brands, with hot dogs categorized by type: beef, poultry, and "meat" (mostly pork and beef, but up to 15% poultry meat). Here we'll look at comparing sodium content by type. We use an analysis of variance (ANOVA) to analyze a quantitative response (sodium) based on a nominal predictor (type). From the JMP Starter window, go to **Basic** and "Oneway". Put **Sodium** as "Y, Response" and **Type** as "X, Grouping" and hit "OK". You should get a plot of sodium for each type.

Question #28 What can you learn from the plot? (EDA items — can you say anything about differences in central location, differences in spread, or about outliers?)

Get the ANOVA output by clicking on the Hot Spot and choosing **Means/Anova**. More information will appear below the plot. Also, JMP adds information to the plot, showing the mean of each group by the widest part of the diamond, and the height of the diamonds is a measure of variability. ANOVA is a way of testing for differences in means among multiple groups. It is a generalization of a two-sample t test for more than two groups. It compares the variability between groups to see if it is large relative to the variability within groups. The formal test hypotheses compare whether all the group means are the same, or not all of them are.

Question #29 Based on the points and on the diamonds added to the plot, how much does the sodium content appear to differ by type of hot dog?

Question #30 Write out the hypotheses for this test. Be sure to define your notation.

Question #31 The **Analysis of Variance** table gives all the information for the formal test. What is the p -value for this test?

↪ **Question #32** What do you conclude about the relationship between hot dog type and sodium?

Quit JMP and please remember to **Log Off**.