



Newcastle
University

Lonza Workshop

September 23rd-24th, 2021

Dr. Lee Fawcett and Dr. Joe Matthews,
Newcastle University

Data summaries in JMP

Importing Data

Throughout these practical sessions we will be working with real datasets to demonstrate the methods studied in the lectures.

All the datasets needed for the practicals are available on the LSI sharepoint directory.

To import a dataset for analysis, first download the data file (JMP is very flexible when it comes to which file types you can import – it can handle pretty much anything!) and save it to a directory on your machine (it doesn't matter where – but remember where you save things!).

Next, open JMP and click **File**→**Open**, then locate and click on the data file you wish to open in JMP, and click **Open**.

Your data should now be imported into JMP and ready for you to analyse!

Movies Dataset

In this practical we will mainly be using the `movies` dataset. This contains data on over 50,000 movies from the International Movie Database (IMDb.com)

Import this dataset into JMP using the steps above, and have a look around. It contains the following variables:

- **title** – The name of the movie
- **year** – The year the movie was released
- **length** – The length of the movie (in minutes)
- **budget** – The budget of the movie in \$USD (if known)
- **rating** – The average rating the movie received from users on IMDb
- **votes** – The number of individuals that rated the movie
- **r1-10** – The (approximate) percentage of individuals that rated the movie 1-10
- **mpaa** – The movie's MPAA classification rating (if known)
- **Action, Animation, ...** – Indicator variables corresponding to the genre of the movie

Question 1: Box Plots

Box plots are an incredibly useful visual representation of the key summaries of a dataset. Let's see how they can be used to visualise movie ratings.

To construct a box plot:

- (i) Click Graph → Graph Builder
- (ii) In the icons under the Graph Builder heading, select the Box Plot image (roughly halfway along)
- (iii) From the column list, click and drag the variable we want to visualise into the centre of the plot area.
- (iv) To split the box plot by a categorical variable, click and drag the grouping variable to the area either just above or below the plot (to the areas marked X or Group X). The column headings denote the values of the categorical variable for each box plot (e.g. the "0" heading corresponds to the box plot with value "0" for the categorical variable)
- (v) You can change the main heading and the axis labels by clicking on them, and typing your preferred heading.

Questions:

- (a) Construct a boxplot of the ratings variable. What appears to be a typical rating for a movie on IMDb? How spread out are the ratings? What ratings do we consider to be "outliers"?
- (b) Split the ratings boxplot by Action. How do the ratings for action movies vs non-action movies compare? Which gets higher ratings on average? Which has the larger spread?
- (c) Repeat part (b) for the remaining genre categories (you can undo the last change you made to your plot by clicking Undo in the top left corner of the Graph Builder window). Which groupings showed the biggest/smallest differences?

Question 2: Histograms

Histograms present an alternative way of viewing continuous data. Let's take a look at the years our movies were released.

To construct a histogram:

- (i) Click Graph → Graph Builder
- (ii) In the icons under the Graph Builder heading, select the Histogram image (again halfway along – one to the right of the box plot image)
- (iii) From the column list, click and drag the variable we want to visualise to the bottom of the plot area (the region marked X).
- (iv) As before we can edit the headings/labels by clicking on them, and typing our preferred labels.

Questions:

- (a) Produce a histogram for the *year* variable. Which time period has the most films on IMDb? How symmetric is the distribution?
- (b) Experiment with adjusting the bar width of the histogram by clicking (and holding) on the x-axis, and moving the mouse up and down. What might be strengths/weaknesses in terms of visualisation, of having much thinner/wider groupings?
- (c) What advantages would using a histogram provide for visualising this data compared with using a box plot? When would a box plot be better?

Question 3: Bar and Pie Charts

We can graphically represent categorical data using pie charts or bar charts depending on the situation. Let's look at the `mpaa` classifications.

To construct a bar/pie chart:

- (i) Click Graph → Graph Builder
- (ii) In the icons under the Graph Builder heading, select the Bar image
- (iii) From the column list, click and drag the categorical variable we want to visualise to the bottom of the middle of the plot area.
- (iv) As before we can edit the headings/labels by clicking on them, and typing our preferred labels.
- (v) To switch to a pie chart, select the Pie chart image (on the right hand side of the graph images toolbar)

Questions:

- (a) Produce a bar chart and pie chart to summarise the `mpaa` variable. In which cases would we prefer a bar chart to a pie chart? When would we prefer the pie chart?

Question 4: Scatter Plots

Scatter plots allow us to visually assess how one variable changes with respect to another. Let's see if there's a relationship between how long a film is, and how well it scores.

To construct a scatter plot:

- (i) Click Graph → Graph Builder
- (ii) In the icons under the Graph Builder heading, select the Histogram image (again halfway along – one to the right of the box plot image)
- (iii) From the column list, click and drag the variable we want to visualise to the bottom of the plot area (the region marked X).
- (iv) As before we can edit the headings/labels by clicking on them, and typing our preferred labels.

Question 5: Numerical Summaries

To obtain a table of numerical summaries:

- (i) Click Analyze—>Distribution
- (ii) Select the variable(s) you are interested in seeing summaries for and drag it/them to the Y, Columns box.
- (iii) If you wish to split your summaries by group, click and drag your grouping variable to the By: box.
- (iv) Click OK

Questions:

- (a) Produce numerical summaries of the `length` variable. Compare the summaries of `length` when split by `mpaa`. Which classification had the longest films? Which had the largest spread of lengths?
- (b) What was the average known budget of films on IMDb? Why might this not be a good estimate of the average budget of films in general?

Exploratory multivariate data analysis in JMP

Question 1: Scatterplot matrices

Let's consider the construction of simple scatterplot matrices in JMP. Import the `heptathlon` dataset that we studied during the last session, and take a look around the dataset. You should have 29 rows in the dataset (one row for each athlete), and 8 columns (one for the athlete's name, and then one containing each athlete's performance across the seven events in the heptathlon).

To construct a scatterplot matrix:

- (i) Click `Graphs` → `Scatterplot Matrix`
- (ii) Select all of the heptathlon variables simultaneously by holding down the "Shift" button on the keyboard; selecting the first heptathlon variable in the `Select Columns` box (the 100 metres hurdles, or "H100M"); and then selecting the last heptathlon variable (the 800 metres run, or "R800M"). Then under the option to `Cast Selected Columns into Roles`, click on `Y, Columns`
- (iii) Hit `OK`

You should now have produced the scatterplot matrix shown in the lecture.

Questions:

- (a) Look down the first column of plots. What do these plots show? What is the nature of the association between performances in the 100 metres hurdles and the 800 metres? What about the relationship between performances in the 100 metres hurdles and the long jump?
- (b) How are performances in the shot put and the javelin throw associated?

Keep the heptathlon dataset open in JMP, you will need it for the next question!

Question 2: Correlation matrices

We will now consider the construction of correlation matrices in JMP. Once again, we will use the heptathlon dataset as an example.

To construct a correlation matrix:

- (i) Select Analyze—>Multivariate Methods—>Multivariate
- (ii) As you did earlier, select all the heptathlon variables simultaneously. Again, under the option to Cast Selected Columns into Roles, click on Y, Columns
- (iii) Click OK

You should see that JMP has now re-produced the scatterplot matrix from before, but also the correlation matrix from the lecture earlier.

Questions:

- (a) Can you explain how the correlation matrix supports your answers in questions 1(a) and 1(b)?
- (b) Using the correlation matrix, can you quickly identify the pairs of variables that are significantly associated? What about those that are not? Which pair of variables reveals the most significant association? We will return to the idea of statistical significance tomorrow.
- (c) Why do all the cells down the diagonal of the matrix contain '1'?

Keep the heptathlon dataset open in JMP, you will need it for the next question!

Question 3: 3-D scatterplots

JMP can be used to produce 3-D scatterplots. We will demonstrate their construction via the heptathlon dataset, so keep this dataset open in JMP!

To produce a 3-D scatterplot:

- (i) Select **Graphs** → **Scatterplot 3D**
- (ii) Drag-and-drop three variables of your choosing into the **Cast Selected Columns into Roles** list
- (iii) Click **OK**

You should now have produced a 3-D scatterplot in keeping with that shown in the lecture.

Question:

- (a) What can you say about the plot you have just produced? Are the relationships between the three variables you selected clear, as the plot stands?

You might be finding it rather challenging to answer the question in part (a). If so, why not try spinning your plot?

Spinning a 3-D scatterplot:

- (i) Click anywhere on your plot and hold the cursor
- (ii) With the cursor held, move the plot around

Question:

- (b) Rotate your 3-D scatterplot so you can easily see the association between a single pair of variables you have selected. You should see that you have a recreation of the corresponding plot in the scatterplot matrix in question 1(a).

Question 4: Principal Components Analysis

The primary focus of the lecture was on the use of Principal Components Analysis (PCA) as a way of exploring multivariate datasets. Recall from the lecture that, for easier interpretation of the principal components, we should make sure the direction of each variable has the same interpretation. In the lecture, we re-scaled data on the running events to make high values correspond to better performances.

Import the `heptathlon2` dataset into JMP. Pay particular attention to the data in the columns for the running events.

Performing a Principal Components Analysis:

- (i) Click on Analyze → Multivariate Methods → Principal Components
- (ii) Highlight all seven heptathlon events and move them all over into the Cast Selected Columns into Roles list
- (iii) Hit OK
- (iv) Make sure the results display window is entitled "Principal Components: on Correlations". If it is not, click the down-arrow just to the left of this title, and then select Principal Components → on Correlations

Let's now take a look at some of the output (and there's lots!).

Questions:

- (a) First of all, the principal components themselves. In your results display window, under the window title make sure the box for Loading Matrix is ticked. This matrix should now appear in your results window. The first column gives you the coefficients of your first principal component, so you should see that

$$Y_1 = 0.717X_1^* + 0.572X_2 - 0.002X_3 + 0.758X_4^* + 0.784X_5 - 0.071X_6 + 0.752X_7^*,$$

where the heptathlon variables have labels $X_1, X_2, \dots$ for the 100 metres hurdles, the high jump, $\dots$, and the stars indicate that performances in the running events have been re-scaled so that high values indicate better performances.

What is your second principal component? And the third? Compare your answers to the material from the lecture.

- (b) Recall that the aim of a PCA is to reduce the dimension of a multivariate dataset, perhaps so that it can be viewed in two dimensions using standard techniques. We would like most of the variability across our multivariate dataset to be explained by

just a few of the principal components (which are functions of all of the original variables in our dataset).

Under the results window heading, make sure `Eigenvalues` is ticked. Then look for the associated output in the updated results window. What percentage of the variability in the heptathlon dataset is explained by the first two principal components?

Constructing a PCA score plot:

Finally, let's produce a score plot. In fact, the score plot has probably already been posted to the results window as part of JMP's default reporting for a PCA; look under `Summary Plots` (if you can't see the summary plots, then make sure that `Summary Plots` is ticked under the results window heading. You should now have produced the score plot from the lecture).

Hover over specific points to reveal the row number in the spreadsheet associated with these points.

Questions:

- (a) Can you locate Jessica Ennis and Irina Karpova in the score plot?
- (b) How might the score plot be used?
- (c) Under the results window heading, make sure `Save Predicteds` is selected, and enter 2 in the `Specify number of components to save` box. Now look around the datasheet; can you see what JMP has done?

Probability models in JMP

Question 1: Binomial and Poisson probabilities

Let's consider the construction of simple probability distributions in JMP. We'll start by re-visiting the two examples from the lecture:

1. The number of tropical storms observed per year
2. The number of rats developing tumours after exposure to *Ochratoxin A*

Recall that in example 1, the number of tropical storms observed per year was assumed Poisson with rate $\lambda = 1.92$, where λ was estimated from the mean of the sample.

To compute Poisson probabilities:

- (i) In a fresh JMP session, click `File`→`New`→`Data Table`. You should see a new spreadsheet window appear
- (ii) In the first column, enter the numbers 0, 1, 2, 3, 4, 5, 6 and 7 in rows 1–8. These will represent observations on X : The number of tropical storms observed per year
- (iii) Double click on the header for the next available column; this should activate `Column 2`, with all entries initially appearing as `.`
- (iv) Having highlighted `Column 2`, click `Cols`→`Formula`. From the list down the left-hand-side, choose `Discrete Probability`→`Poisson Probability`
- (iv) Now enter the value 1.92 as the value for `lambda` in the formula box for `Poisson Probability`; select `k`, and double-click on `Column 1` from the left-hand-side
- (v) Hit `OK`

Questions:

- (a) Using the output generated above, what is the probability of observing exactly 6 tropical storms at this location in 2022?
- (b) What is the probability of observing no tropical storms at this location in 2022?
- (c) Repeat the steps above, but in (iii) and (iv) activate `Column 3`, and in (iv) choose `Poisson Distribution`. Look at the resulting output in `Column 3`. What has JMP computed now? What is the probability that this location will see no more than 4 tropical storms next year?

Recall that in example 2 the number of rats developing tumours was assumed Binomial, with $n = 5$ and $p = 0.35$; there were 5 rats in the sample, and from the literature we suspect that there is a 35% chance of a rate developing tumours having been exposed to the toxin.

To compute Binomial probabilities:

- (i) Click File→New→Data Table. You should see another new spreadsheet window appear
- (ii) In the first column, enter the numbers 0, 1, 2, 3, 4 and 5 in rows 1–6. These will represent observations on X : The number of rats developing tumours
- (iii) Just as you did in the Poisson example, activate a new column. This time, from Cols→Formula choose Binomial Probability from the Discrete Probability options; enter your values for p and n , and Column 1 for k , and hit OK

Questions:

- (d) What is the probability that exactly four rats will develop tumours?
- (e) Follow what you did in part (c) to obtain the cumulative distribution function. Use the output to find the probability that at most, 3 rats will develop tumours.

Question 2: More Binomial and Poisson probabilities

- (a) A recent study published in the *Annals of Internal Medicine* suggests that far more people than previously thought are carrying variants of rare genetic diseases, which could force us to redefine what is considered a healthy genome. In fact, they estimate that 20% of people are likely to carry such variants.

For a sample of 10 randomly chosen subjects, use JMP to find the full probability distribution for X : Number of subjects carrying variants of rare genetic diseases.

- (b) Stored fruit is subject to insect damage. For a sample of 150 fruits, the average number of insects per fruit is 1.55. Use JMP to find the probability that a randomly selected piece of fruit had exactly 0, 1, 2 and 3 insects.

Question 3: Studying Normal populations

Import the female haematocrit values used in the lecture into JMP.

- (a) Use techniques from session 1 this morning to summarise the data both graphically and numerically, to show that an assumption of Normality – with $\mu = 42.15$ and $\sigma = 3.44$ – is plausible.

To compute Normal probabilities:

- (i) Activate Column 2 of the Haematocrit spreadsheet, and enter 42.15, 45, 48, 53 into rows 1, 2, 3 and 4 of this column
- (ii) Now let's find the probability that a randomly chosen female student will have a haematocrit value of less than 42.15%, 45%, 48% and 53%. Activate Column 3 and highlight this column
- (iii) Select Cols → Formula
- (iv) From the options down the left-hand-side, select Probability → Normal Distribution; in the formula editor box, press the caret (^) symbol twice – you should see spaces appear to enter x, mean and std dev in the formula for the Normal Distribution
- (v) For x, double-click on Column 2; for mean enter 42.15 and for std dev enter 3.44
- (vi) Click OK

Questions:

- (b) Given the output in Column 3, what is the probability that a randomly chosen female student has a haematocrit measurement of less than 42.15%? Why should you have known the answer to this question without any computation in JMP?!
- (c) What is the probability that a randomly chosen female student has a haematocrit measurement of less than 48%?
- (d) What is the probability of a randomly chosen female student having a haematocrit measurement of *more than* 45%?

Generating a Normal random sample:

- (i) Suppose we are told that male haematocrit values are Normally distributed, with $\mu = 45$ and $\sigma = 4$. To produce synthetic data for 126 male students, activate and highlight a new column in the spreadsheet containing female haematocrit values. Give this column a new name – for example, “Ht Male (simulated)”
- (ii) Select **Cols** → **Formula**
- (iii) Select **Random** → **Random Normal**
- (iv) Double-click inside the brackets of the **Random Normal** formula, and enter the mean and standard deviation separated by a comma
- (v) Hit **OK**

Questions:

- (e) Explore the simulated dataset by producing a histogram. Also find the mean and standard deviation of your simulated sample.
- (f) What is the probability that a male student will have a haematocrit value less than 48%? How does this compare to that for a female student?

Question 4: Normal Q-Q Plots

Sticking with the female Haematocrit values, let's produce a Q-Q plot to assess Normality in the data.

Generating a Normal Q-Q plot:

- (i) Highlight the column containing the female Haematocrit values, and then select Analyze→Distribution
- (ii) In the Y, Columns box drag and drop Ht Female, and click OK
- (iii) You should now see a histogram, a boxplot and some data summaries. From the drop-down options under the title "Ht Female", select Normal Quantile Plot
- (iv) From the drop-down options under the title "Ht Female", you can also select Continuous Fit→Normal, to superimpose the Normal PDF on the histogram

Questions:

- (a) How would you interpret the Normal Q-Q plot produced above?
- (b) The *Exponential distribution* is a probability distribution for continuous data that are very left-skewed (rather than symmetric, as with Normally-distributed data). This distribution has a single rate parameter λ controlling the extent of asymmetry in the distribution. Can you follow the steps outlined in question 3 to generate a random sample from an exponential distribution with rate $\lambda = 2$, and produce a Normal Q-Q plot for this sample? What does it show?

Confidence Intervals and Hypothesis Testing in JMP

Question 1: Confidence Intervals

Here we analyse the Big Class dataset, containing the name, age, sex, height and weight of 40 martial arts students.

To produce a confidence interval:

- (i) Click Analyze—>Distribution
- (ii) Click and drag the response variable to the Y box.
- (iii) If you wish to calculate separate confidence intervals for different values of a variable, drag the grouping variable to the By: box.
- (iv) Click OK.
- (v) To change the size of the confidence intervals, right click the red arrow next to Summary Statistics and click Customize Summary Statistics. In the box next to Enter (1-alpha) for mean confidence interval enter the level of confidence you'd like (e.g. for a 90% confidence interval enter 0.9) and click OK.

Questions:

- (a) Produce a 95% confidence interval for height. How would we interpret this interval?
- (b) Produce 90% and 99% intervals for height. How do they compare to the 95% interval?
- (c) Produce separate 95% confidence intervals for height grouped by Sex. How do these intervals compare? What can we say about the difference in heights between males and females in this group?

Question 2: One Sample t -test

To carry out a one-sample t -test:

- (i) Click Analyze—→Distribution
- (ii) Click and drag the response variable to the Y box.
- (iii) Click OK.
- (iv) Right click the red arrow next to the response variable heading, and click Test Mean
- (v) Enter the proposed mean under the null hypothesis.
- (vi) Check the box if you want the non-parametric Wilcoxon Signed-Rank test carried out also.
- (vii) Click OK.
- (viii) Right click on the red arrow next to the response variable heading and click Normal Quantile Plot to produce a Q-Q plot to check the model assumptions.

Questions:

- (a) Investigate whether there is evidence to support the mean weight of students being above 110lbs, i.e. test the hypothesis

$$H_0 : \text{Mean} = 100,$$

$$H_1 : \text{Mean} > 100$$

What p -value we do get? What does that mean regarding our null hypothesis? What does mean for our investigation?

- (b) Produce a Q-Q plot for the weight data. Does the Normality assumption appear satisfied?
- (c) Repeat the hypothesis test in part (a), but this time only for the male students. Does the conclusion change?
- (d) Do the t -test assumptions hold in the male only case?
- (e) Repeat the test from part (c) using the non-parametric Wilcoxon signed-Rank test. Do the conclusions change?

Question 3: Two Sample t -test

To carry out a two-sample t -test:

- (i) Click Analyze—→Fit Y by X
- (ii) Click and drag the response variable to the Y, Response box.
- (iii) Click and drag the predictor variable to the X, Factor box.
- (iv) Right click the red arrow next to the Oneway Analysis of... heading in the resulting window, click Unequal Variances
- (v) If the resulting p -values are greater than 0.05, go to step (vi), otherwise go to step (vii).
- (vi) Right click the red arrow next to the Oneway Analysis of... heading and click Means/Anova/Pooled t .
- (vii) Right click the red arrow next to the Oneway Analysis of... heading and click t test.

Questions:

- (a) Produce a box plot of `weight` split by `sex`. Do the weights appear significantly different between the two groups? Do the spreads look similar?
- (b) Run a two-sample t -test to test if `weight` is significantly different between the two groups (remember to check if the variances can be assumed equal). What p -value do we get? What does that mean for the null hypothesis? What does it mean in terms of our investigation?
- (c) Rerun the test from part (b) as a non-parametric Mann-Whitney test. Do the conclusions change?
(To do this right click the red arrow next to the Oneway Analysis of... heading, click Nonparametric—→Wilcoxon Test)

Performing ANOVA in JMP

Question 1: Simple one-way ANOVA

The following data are from an experiment to compare the effects of four catalysts on the yield of a chemical process:

Catalyst	Observations				
A	4.85	4.89	5.60	5.08	3.42
B	9.43	6.77	5.39	7.39	
C	4.67	5.49	5.96	3.71	5.51
D	5.50	4.90	6.00	5.81	4.68

Open a fresh JMP datasheet by clicking File→New→Data Table.

- Now enter all the data shown in the table above in a single column (Column 1)
- Next, enter the associated Catalyst label for each observation in the next column (Column 2)
- Re-label Column 1 as 'Yield' by double-clicking on the column heading and changing the Column Name. Also change the heading of Column 2 to 'Catalyst'

You have now created a 'stacked data table', which is necessary for performing ANOVA in JMP. Your data table should look like this:

Yield	Catalyst
4.85	A
4.89	A
5.60	A
5.08	A
3.42	A
9.43	B
⋮	⋮
4.68	D

Questions:

- Using any techniques you have learnt in this course so far, check the assumption that the four samples are Normally distributed.
- Also check the underpinning assumption of equal variances in the four populations from which these samples have been drawn.
- Write down your null hypothesis.

Performing a one-way ANOVA:

- (i) Click on Analyze—→Fit Y by X
- (ii) In the Cast Selected Columns into Roles list, select 'Yield' as the Y, Response and 'Catalyst' as the X, Factor
- (iii) Hit OK
- (iv) From the drop-down menu of the results sheet (Oneway Analysis of Yield By Catalyst), select Means/ANOVA, and make sure you can see the resulting ANOVA table
- (v) From the same drop-down menu, select Compare Means—→All Pairs, Tukey HSD, and make sure you can see the "HSD Threshold Matrix" and the "Connecting Letters Report".

Questions:

- (d) Look at the ANOVA table produced in step (iv) above. What does this tell you about your null hypothesis in part (c)? Can you say what this means in non-technical terms?
- (e) Now look at the output produced in step (v), and draw your conclusions about the test you have performed.

Question 2: Simple one-way ANOVA

Perform a one-way analysis of variance to check for significant differences in the mean width of cuckoos' eggs laid in the nest of different birds. We discussed this dataset in the lecture; the data can be found in the file `cuckoo.jmp`.

Question 3: ANOVA with blocking

In this question, we will re-produce the analysis of the Carrots data used in the lecture. You should open the datafile `carrots.jmp`.

Questions:

- (a) Follow the steps in question 1 to perform a one-way ANOVA on the carrots data in JMP. However, in step (ii), you should also make sure you enter the Block.
- (b) Look at the output produced in part (a), and make sure you can see:
 - Why we conclude that there are significant differences in mean yield between the ten different strains of carrots
 - Why we conclude that the nuisance factor (plot) is *not* significant
 - Why we concluded that there were three main groupings in terms of the yields produced by the different strains of carrots, and why average yields produced by strain D are significantly different to all others
- (c) Produce output in JMP to check the assumptions underpinning this ANOVA

Question 4: Factorial experiments

Open the datafile `rats.jmp`. The aim of this question is to reproduce the analysis of the rats gain dataset used in the lecture.

Questions:

- (a) Check the assumption of Normality in this dataset. Technically, you should check for Normality in *both directions*; that is, for weight gains in each of the three protein 'source' groups, and in both of the protein 'level' groups.

Performing a two-way ANOVA:

- (i) Click on Analyze→Fit Model
- (ii) Enter the 'Gains' data as the Y variable
- (iii) Now, select both 'Source' and 'Level' by pressing the Ctrl key and selecting 'Level' and 'Source' from the Select Columns list. Both of these variables should now be highlighted in blue
- (iv) In the Construct Model Effects box, select Macros and the choose Full Factorial; you should now see 'Level', 'Source' and 'Level*Source' in the model box – which means we are checking for the significance of both variables *and* interaction between them
- (v) Click Run
- (vi) In the resulting output, make sure you can see the 'Analysis of Variance' table and the 'Effect Tests' table
- (vii) In the output, scroll across to the right until you see plots for the interaction term ('Level*Source'). From the drop-down menu select LS Means Plot to produce the interaction plot shown in the lecture.

Questions:

- (b) From the output in (vi) above, can you see why source of protein is significant but level of protein is not?
- (c) Can you see why there is a significant interaction, and can you use the plot produced in (vii) to understand this interaction?

Regression Modelling in JMP

Question 1: Linear regression

The dataset we will be analysing in this question is the Wage dataset, containing the wage and other data of 3000 individuals from the mid-atlantic region of the USA. The goal is to be able to accurately model wage using the other information available.

The dataset contains the following variables:

- **year**: The year of the observation
- **age**: The age of the worker
- **maritl**: The worker's marital status
- **race**: The worker's race
- **education**: The level of education the worker has
- **jobclass**: The sector of job the worker has
- **health**: The health of the worker
- **healthins**: If the worker has health insurance
- **wage**: The worker's raw wage annual wage (in \$ USD)

Let's investigate the effect of a worker's characteristics on their wage using simple linear regression.

To carry out a linear regression:

- (i) Click Analyze—→Fit Model
- (ii) Click and drag the response variable, to the box next to Y. (Alternatively, select your response variable and click Y for the same result.)
- (iii) Click and drag your predictor variable(s) to the Construct Model Effects box
- (iv) Click “Run”
- (v) In the resulting window, right click the red arrow next to Response. Click Row Diagnostics—→Plot Residual by Predicted & Plot Residual by Normal Quantiles to produce the diagnostic summary plots.
- (vi) To calculate the predicted value, confidence interval and prediction interval for a new individual, before running the regression, add a row to the data table with values of the predictor variable(s) for the new individual (but leave the response variable entry blank). After running your regression, right click the red arrow next to Response, click Save Columns—→Predicted Values & Mean Confidence Interval & Indiv Confidence Interval.

Questions:

- (a) Produce a scatter plot of wage against age. Does it look like there's a relationship? Does it look linear?
- (b) Run a simple linear regression on wage using age. What is the effect of age on wage? Is this effect significant?
- (c) Produce the regression diagnostic plots – do the assumptions appear satisfied? Thinking about the variables, why might the assumptions of linear regression not be valid here?
- (d) Rerun the analysis with `logwage` as the dependent variable – are the diagnostic plots any better?
- (e) Find the predicted `logwage` for an individual aged 18, along with a 95% confidence interval and 95% prediction interval.
- (f) Rerun the regression on `logwage`, this time including all predictor variables (i.e. all variables aside from wage and `logwage`). Which variables are significant? What effects do the characteristics have on expected wage?
- (g) Reproduce the diagnostic plots for the multiple linear regression. Do the assumptions appear satisfied?

Question 2: Quadratic regression

Recall the example from the lecture, investigating the relationship between signal peak heights (Y) for known concentrations of albumin ($X \mu\text{g mL}^{-1}$). These data can be found in the file `albumin.jmp`. Open this datafile in JMP.

Question:

- (a) Produce a scatterplot of X against Y .

Performing a quadratic regression:

- (i) Follow the steps in previous questions to perform your regression (i.e. Analyze→Fit Y by X...
- (ii) From the drop-down menu in your output window, select Fit Polynomial→2,quadratic
- (iii) From the drop-down menu, also select Fit Line

Questions:

- (b) Is the quadratic model worthwhile here?
- (c) Write down the quadratic model in full.
- (d) Use methods introduced earlier to check the assumptions underpinning this regression model fit.

Question 3: Logistic regression

Open the file `leukaemia.jmp`. We will now reproduce the binary logistic regression example from the lecture.

Follow the steps in previous questions to perform a multiple regression (i.e. Analyze→Fit Model...). As the dependent variable, "REMISS", is binary, JMP should automatically suggest using Nominal Logistic in the Personality box. Now follow the process of backwards elimination, introduced in the lectures, to show that the best-fitting binary logistic regression model is the *simple binary logistic regression model*

$$P(Y = 1) = \frac{e^{-3.778+2.897X_4}}{1 + e^{-3.778+2.897X_4}}.$$