NCRM National Centre for Research Methods

StatJRs eBook interface and Statistical Analysis Assistants

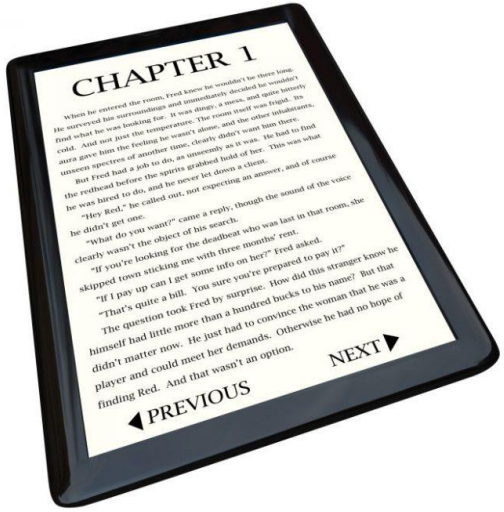
Professor William Browne

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
UNIVERSITY OF Southampton
MANCHESTER 1824 The University of Manchester
THE UNIVERSITY OF EDINBURGH

Ebooks

+

=



An electronic book is a book-publication in digital form.
In the US more books are published online than distributed in hard copy in book shops.

Statistical (and Mathematical) eBooks

- The idea is can we incorporate statistical content into an eBook? Of course a statistical textbook is no different on paper to any other document when it comes to creating a pdf file (aside from maybe more equations!)
- The difference is in what ‘enhancements’ we can add and so the idea here is combining the text book with the statistics package i.e. interactive examples, allowing the user to include their own dataset etc.

Multilevel modelling with the 'tutorial' dataset

Finished

← Previous 1 2 3 4 5 Next → Go to page

Navigate through pages of eBook

Hierarchical table of contents (can be expanded / collapsed at each node)

Overview

This eBook provides a bri orial dataset.

We are developing eBook t statistics. They're an interactive environment, and dynamic content will appear tailore

You progress through the ocks at the top and bottom of the page, or via the hierarchical table of contents on the le mes available as a result of your choices).

EBook functionality is still being developed, so you may notice the odd thing here or there yet to be finessed (such as the large number of decimal places sometimes returned!), but we nevertheless wanted to introduce you to what we hope you find to be an interesting means of exploring statistics, and we would very much appreciate any comments you have.

Note that there may be a short delay until all available contents on a particular page are uploaded - you can keep an eye on progress either via the gauge in the top-left corner of the browser window, or by looking at the command window running in the background.

NB: if your eBook crashes, then you can reload the eBook by choosing Debug > Reload eBook from the black bar towards the top of this window. That will wipe you're previous choices, I'm afraid, but it will (hopefully) breathe life back into the software!

The tutorial dataset

The **tutorial** dataset is one of the example datasets provided with the Stat-JR package (as well as with the software package MLwiN) and is summarised below. This dataset was selected from a much larger dataset of examination results from six inner London Education Authorities (school boards). A key aim of the original analysis was to establish whether some secondary schools were more 'effective' than others in promoting students' learning and development, taking account of variations in the characteristics of students when they started secondary school. The analysis then looked for factors associated with any school differences found. Thus the focus was on an analysis of examination performance after adjusting for student intake achievements.

Exploring the tutorial dataset

We'll be modelling **normexam** as the response (ble: as the summary below indicates, this represents the students' exam score at age 16, normalised to have an approximately standard Normal distribution.

In fact, you can view the full dataset via the **Resources** button, which you can find in the black bar at the top of this window. In the resulting

Multilevel modelling with the 'tutorial' dataset

Finished

← Previous 1 2 3 4 5 Next → Go to page

Summary table of tutorial dataset

Column name	n	Missing	Min	Max	Description
school	4059	0	1	65	Numeric school identifier
student	4059	0	1	198	Numeric student identifier
normexam	4059	0	-3.67	3.67	Students' exam score at age 16, normalised to have approximately a standard Normal distribution.
cons	4059	0	1	1	A column of ones. If included as an explanatory variable in a regression model, its coefficient is the intercept.
standlrt	4059	0	-2.93	3.02	Students' score at age 11 on the London Reading Test (LRT), standardised using Z-scores.
girl	4059	0	0	1	Students' gender: 0=boy; 1=girl
schgend	4059	0	1	3	School gender: 1=mixed; 2=boys' school; 3=girls' school
avslrt	4059	0	-0.76	0.64	Average LRT score in school
schav	4059	0	1	3	Average LRT score in school, coded into 3 categories: 1=bottom 25%; 2=middle 50%; 3=top 25%
vrband	4059	0	1	3	Students' score in test of verbal reasoning at age 11, coded into 3 categories: 1=top 25%; 2=middle 50%; 3=bottom 25%

Plotting variables

Here you can graphically-explore the **tutorial** dataset.

In the first two sections, below, you can produce a densityplot and XY plot, respectively; here you can re-specify your choice of variables

Multilevel modelling with the 'tutorial' dataset

Finished

← Previous | 1 | 2 | 3 | 4 | 5 | Next → | Go to page

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Type of plot: densityplot

Submit

about

...then, once you have made your choices, **your plot will appear here:**

Cross-tabulation

Here you can create a table of means and standard deviations for one variable, conditioned on another variable. The first question asks which variable to condition on: a column will be produced for each value of this variable, and so for it to be a useful guide to your data it is best if the variable you choose here consists of relatively few, discrete categories (e.g. **girl**, **schgend**, etc). If you don't want to condition on any variables, you can simply choose **cons**.

What variable do you want to condition your columns on?: school

What variable do you want to produce means etc for?: school

Submit

about

14:50
13/06/2012

Finished

Multilevel modelling with the 'tutorial' dataset

← Previous 1 2 3 4 5 Next → Go to page

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Which variable would you like to use to construct x-axis panel: schgend

Which variable would you like to use to construct y-axis panel: vrband

Do you want the variable name included in panel bar, or just the level: Yes

Submit

about

...then, once you have made your choices, **your plot will appear here:**

Cross-tabulation

Here you can create a table of means and standard deviations for one variable, conditioned on another variable. The first question asks which variable to condition on: a column will be produced for each value of this variable, and so for it to be a useful guide to your data it is best if the variable you choose here consists of relatively few, discrete categories (e.g. **girl**, **schgend**, etc). If you don't want to condition on any variables, you can simply choose **cons**.

What variable do you want to condition your columns on?: school

What variable do you want to produce means etc for?:

Running RScript

Multilevel modelling with the 'tutorial' dataset

← Previous | 1 | 2 | 3 | 4 | 5 | Next → | Go to page

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Which variable would you like to use to construct x-axis panel: schgend

Which variable would you like to use to construct y-axis panel: vrband

Do you want the variable name included in panel bar, or just the level: Yes

Submit

about

...then, once you have made your choices, **your plot will appear here:**

Cross-tabulation

Here you can create a table of means and standard deviations for one variable, conditioned on another variable. The first question asks which variable to condition on: a column will be produced for each value of this variable, and so for it to be a useful guide to your data it is best if the variable you choose here consists of relatively few, discrete categories (e.g. **girl**, **schgend**, etc). If you don't want to condition on any variables, you can simply choose **cons**.

What variable do you want to condition your columns on?: school

What variable do you want to produce means etc for?:

Finished

Multilevel modelling with the 'tutorial' dataset

← Previous | 1 | 2 | 3 | 4 | 5 | Next → | Go to page

Your choice of plot

Finally, here you have more flexibility in specifying a plot of your choice. For more information on what the various options mean, please refer to the **PlotsViaR template eBook**...

Which variable would you like to use to construct x-axis panel: schgend

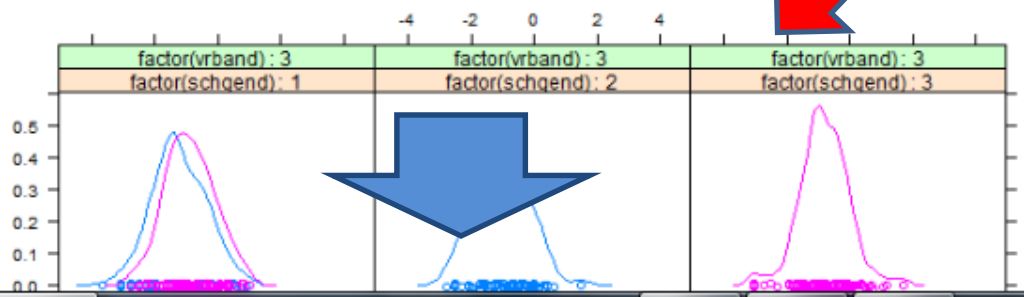
Which variable would you like to use to construct y-axis panel: vrband

Do you want the variable name included in panel bar, or just the level: Yes

Submit

about

Here is the plot you requested:

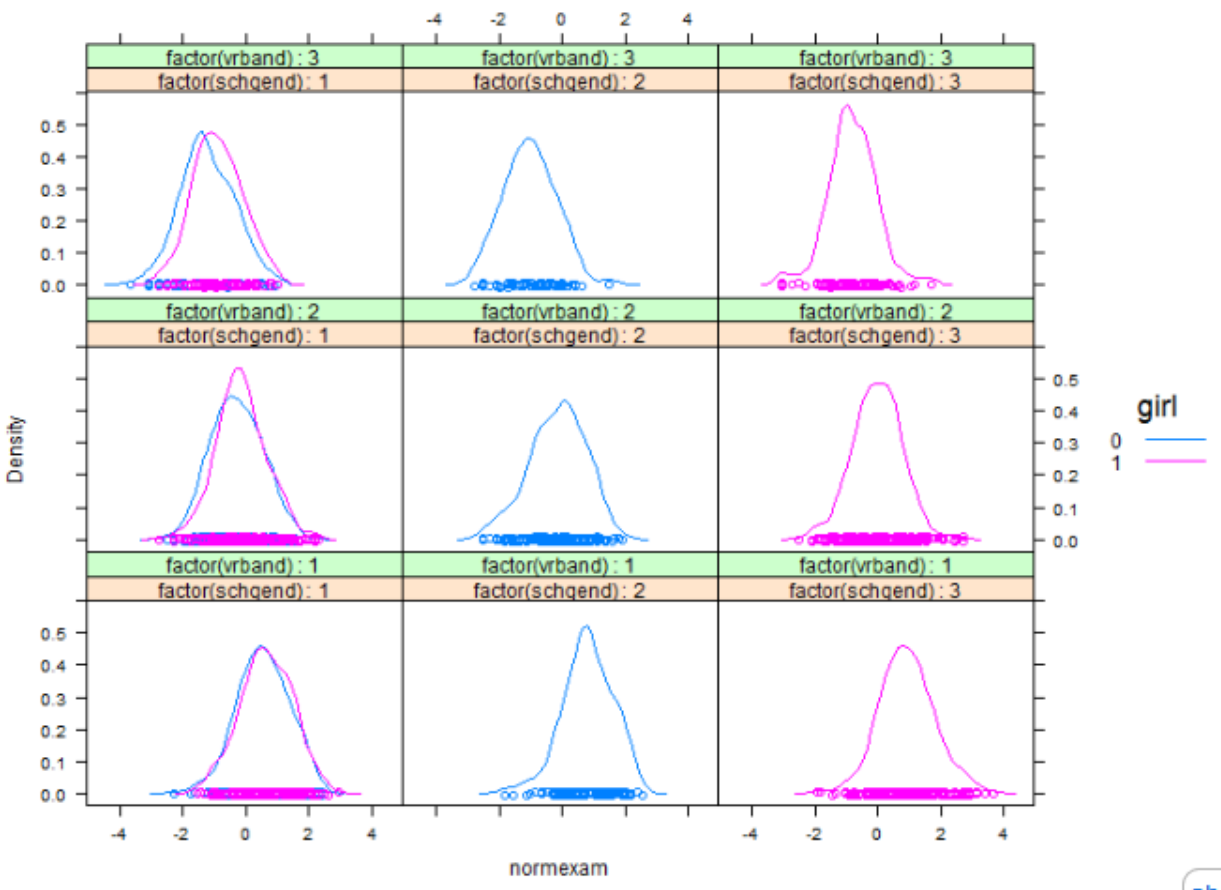


14:56
13/06/2012

Multilevel modelling with the 'tutorial' dataset

Finished

← Previous 1 2 3 4 5 Next → Go to page

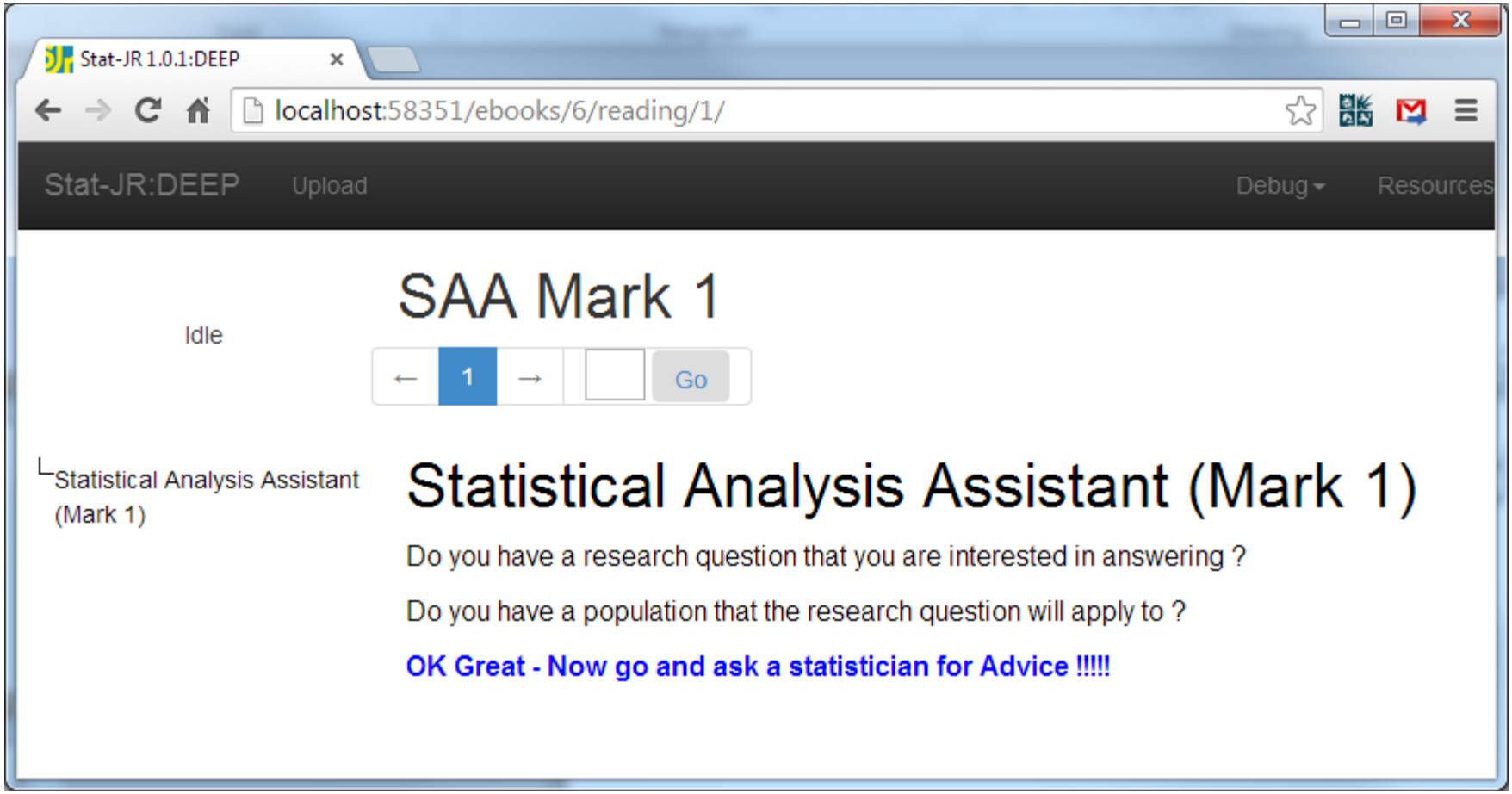


about

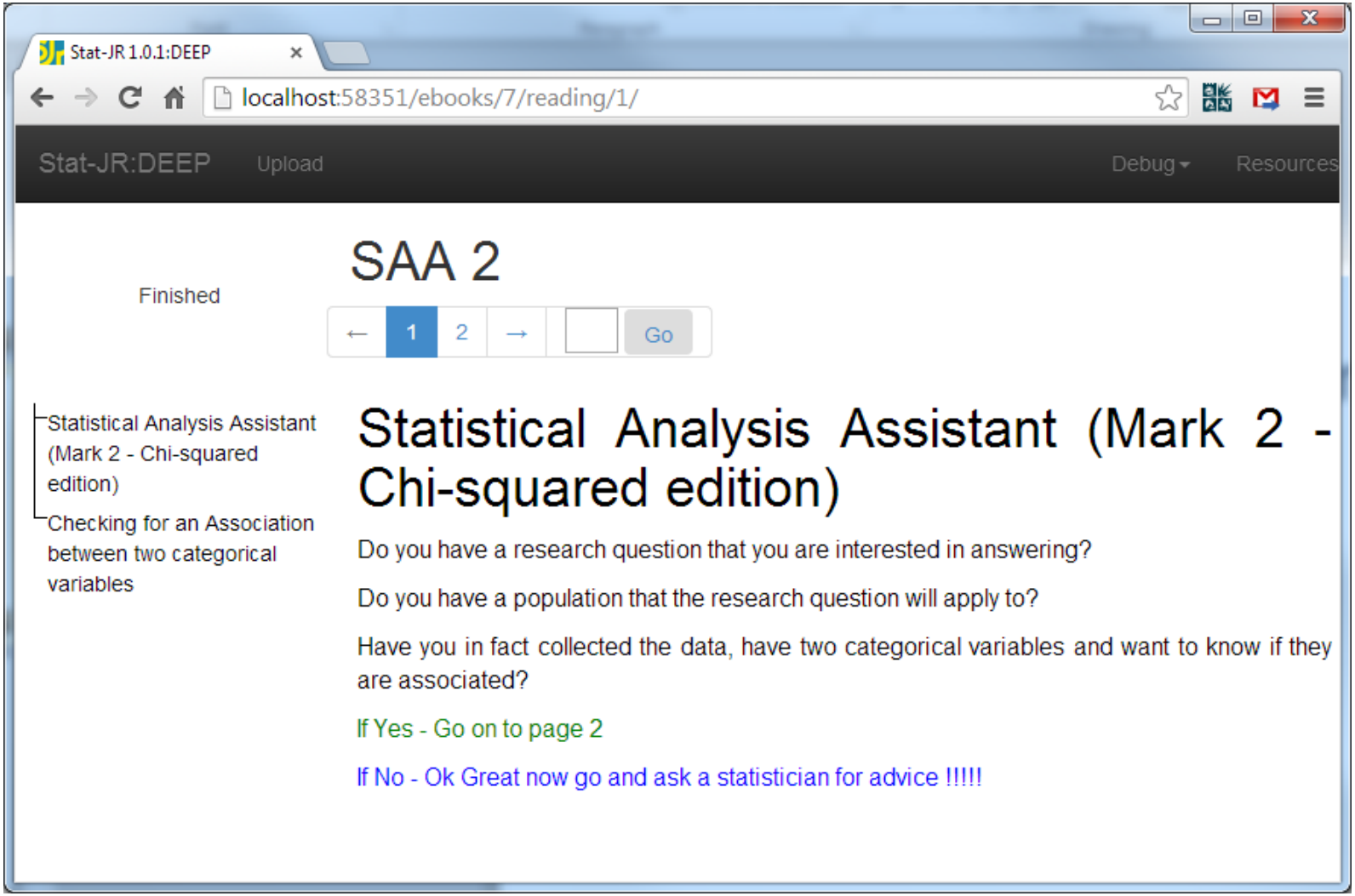
Statistical Analysis Assistants

- We adapt our eBook system to allow workflows that will be constructed to describe how the steps in a statistical analysis fit together.
- There may be many SAAs adapted to different researcher's approaches – e.g. one might want to answer a research question/analyse a dataset as a specific expert might do it.
- Opinion is divided on how far one can take the idea – from nowhere to complete automation i.e. pour in the dataset at the top and let the computer sort it out.
- Probable end point will be somewhere in between or in fact a series of SAAs that lie on this continuum.
- Easiest to start with automating single operations.

A statistical analysis assistant we are all happy with!



One Step further



Finished

SAA 2

Checking for an Association between two categorical variables

You will be presented below with the choice of categorical variables to choose. Having chosen them you will then get the output to your analysis

First categorical variable: cscat

Submit

about

Second categorical variable: nsucc

Submit

about

To do a chi-squared test we start by tabulated observed counts and totals:

Observed	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	188	1559	303	2050
nsucc=1.0	139	1536	440	2115
Total	327	3095	743	4165

Finished

SAA 2

← Previous 1 2 Next → Go to page

To do a chi-squared test we start by tabulated observed counts and totals:

Observed	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	188	1559	303	2050
nsucc=1.0	139	1536	440	2115
Total	327	3095	743	4165

We can therefore work out the expected counts from the margins of the observed data

And so we expect

E(cscat=0.0,nsucc=0.0)= Total cscat=0.0* Total nsucc=0.0/grand total = 327*2050/4165=160.95
E(cscat=1.0,nsucc=0.0)= Total cscat=1.0* Total nsucc=0.0/grand total = 3095*2050/4165=1523.35
E(cscat=2.0,nsucc=0.0)= Total cscat=2.0* Total nsucc=0.0/grand total = 743*2050/4165=365.7
E(cscat=0.0,nsucc=1.0)= Total cscat=0.0* Total nsucc=1.0/grand total = 327*2115/4165=166.05
E(cscat=1.0,nsucc=1.0)= Total cscat=1.0* Total nsucc=1.0/grand total = 3095*2115/4165=1571.65
E(cscat=2.0,nsucc=1.0)= Total cscat=2.0* Total nsucc=1.0/grand total = 743*2115/4165=377.3

So the table of expected counts is

Expected	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	160.95	1523.35	365.7	2050.0
nsucc=1.0	166.05	1571.65	377.3	2115.0
Total	327.0	3095.0	743.0	4165.0

We next look at differences between what we observe and expect in each cell. We square these values so that every difference is positive and scale by the expected counts so that more frequently expected cells arent overly influential. So for example for cscat=0.0, nsucc=0.0 (O-E)^2/E = (188-160.95)^2/160.95=4.55. This statistic is shown in tabular form below

Finished

SAA 2

← Previous 1 2 Next → Go to page

Expected	cscat=0.0	cscat=1.0	cscat=2.0	Total
nsucc=0.0	160.95	1523.35	365.7	2050.0
nsucc=1.0	166.05	1571.65	377.3	2115.0
Total	327.0	3095.0	743.0	4165.0

We next look at differences between what we observe and expect in each cell. We square these values so that every difference is positive and scale by the expected counts so that more frequently expected cells arent overly influential. So for example for cscat=0.0, nsucc=0.0 (O-E)^2/E = (188-160.95)^2/160.95=4.55. This statistic is shown in tabular form below

(O-E)^2/E	cscat=0.0	cscat=1.0	cscat=2.0
nsucc=0.0	4.55	0.83	10.75
nsucc=1.0	4.41	0.81	10.42

The test statistic for a chi-squared test is found by summing the values of this table so

Chisq=4.55+0.83+10.75+4.41+0.81+10.42=31.77

This is compared with a chi-squared table with degrees of freedom = (number of columns -1)x(number of rows - 1) =

(2-1)x(3-1)=2

Looking up the chi-squared table the value for P=0.05 is 5.99 and for P=0.01 = 9.21

as 31.77 > 9.21 our P value is less than 0.01 and we have strong evidence to reject the null hypothesis (at the P=0.01) level

The p-value is in fact less than 0.0001

about

Adding contextual text to a single operation

As we have seen with the Chi-squared example it is easy to enhance a single statistical operation like a statistical test.

We can easily expose the steps required for the test in this case –

1. The tabulation of the observed counts
2. The calculation of the corresponding expected counts
3. The calculation of the test statistic and degree of freedom
4. The interpretation of the test, the P value and what it means in words.

What is harder is to then put what the result means into context.

Statistical tests and tables are fairly easy to enhance with intelligent textual information whilst graphs and figures are harder to enhance. Generally one has to calculate a statistic related to the figure and work with that e.g. skewness and histograms as shown later.

‘The Warlock of Firetop Mountain’ approach

- The first of a genre of interactive books published in 1982 and lapped up by 10 year olds like myself!
- A combination of book and flowchart
- Worked something like:

‘The goblin advances towards you, shouting words that you can’t understand, do you try to make conversation (turn to page 231), run past the goblin (turn to page 176) or draw your sword and fight (turn to page 134)’

- Basically underpinning the book was effectively a flowchart disguised by random page movements with a variety of endings (99% of them involved you dying), possible loops etc.

‘The Warlock of Firetop Mountain’ approach

- The first of a genre of interactive books published in 1982 and lapped up by 10 year olds like myself!
- A combination of book and flowchart
- Worked something like:

‘The goblin advances towards you, shouting words that you can’t understand, do you try to make conversation (turn to page 231), run past the goblin (turn to page 176) or draw your sword and fight (turn to page 134)’

- Basically underpinning the book was effectively a flowchart disguised by random page movements with a variety of endings (99% of them involved you dying), possible loops etc.

The use of Flowcharts in Statistics

- The equivalent exists in (at least) basic statistical analysis and a variety of books have flowcharts to guide the uninitiated to the appropriate test.
- The branching rules are usually things like – how many variables do you have?, what type are they?, is a normality assumption appropriate?
- The example flowcharts usually then say you need a t test / Mann Whitney test / ANOVA etc.
- One could expand this idea to include branches where we haven't written material – i.e. the equivalent of ending up dead would be the default 'go and ask a statistician' end point – possibly taking your answers to the flow chart with you.

The use of Flowcharts in Statistics

- The equivalent exists in (at least) basic statistical analysis and a variety of books have flowcharts to guide the uninitiated to the appropriate test.
- The branching rules are usually things like – how many variables do you have?, what type are they?, is a normality assumption appropriate?
- The example flowcharts usually then say you need a t test / Mann Whitney test / ANOVA etc.
- One could expand this idea to include branches where we haven't written material – i.e. the equivalent of ending up dead would be the default 'go and ask a statistician' end point – possibly taking your answers to the flow chart with you.

Where might this go?

- The flow chart idea is appealing as it may to some degree mimic a statistical consultation.
- If the system is flexible enough then each statistician can tune the SAA to their own approach to analysis and to how much they feel can be comfortably automated.
- Where there is uncertainty / options in what one should do this could be incorporated
- E-books can contain hyperlinks so that further background on proposed statistical methods or examples can be easily found

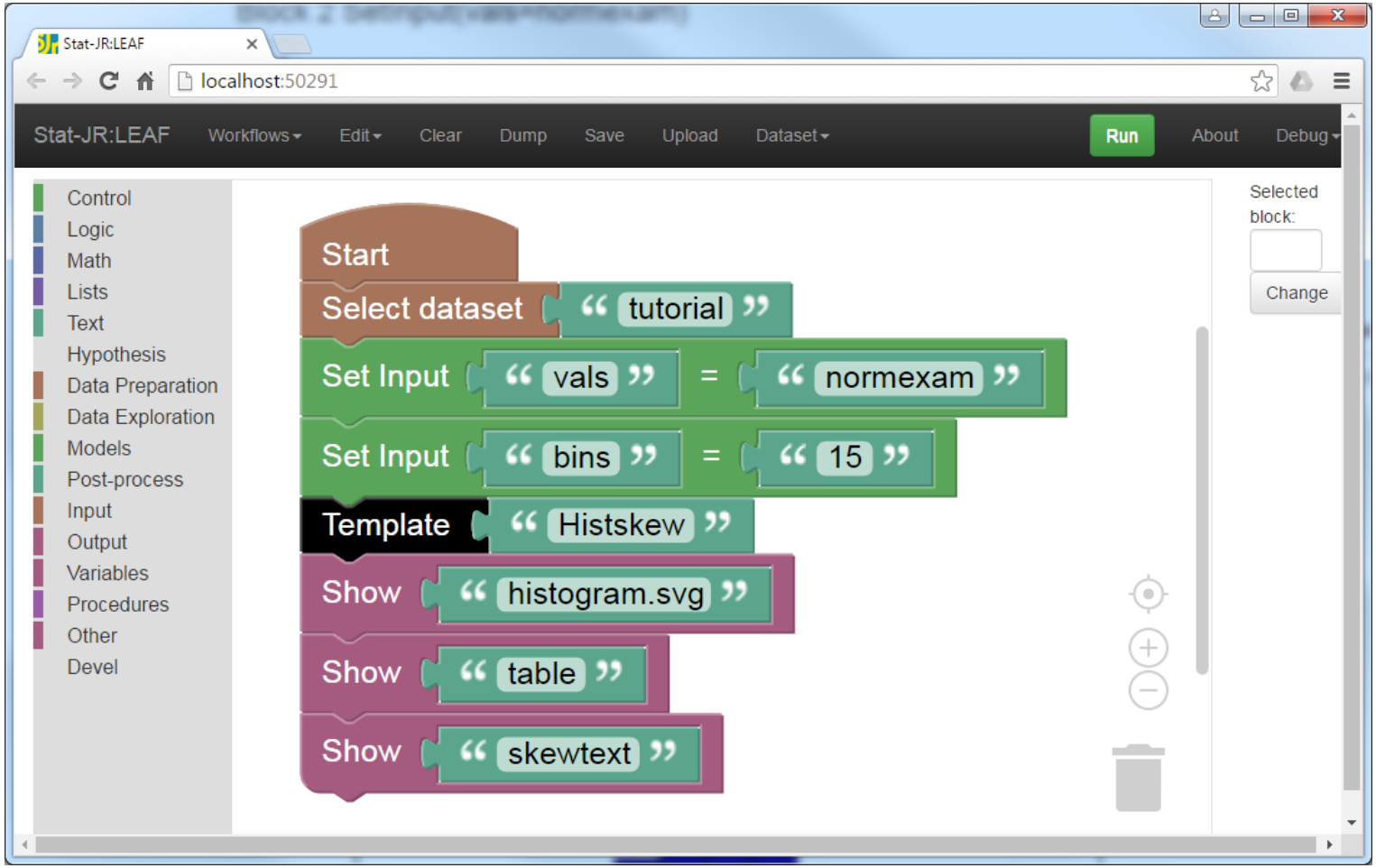
Where might this go?

- The flow chart idea is appealing as it may to some degree mimic a statistical consultation.
- If the system is flexible enough then each statistician can tune the SAA to their own approach to analysis and to how much they feel can be comfortably automated.
- Where there is uncertainty / options in what one should do this could be incorporated
- E-books can contain hyperlinks so that further background on proposed statistical methods or examples can be easily found

Workflows and StatJR LEAF

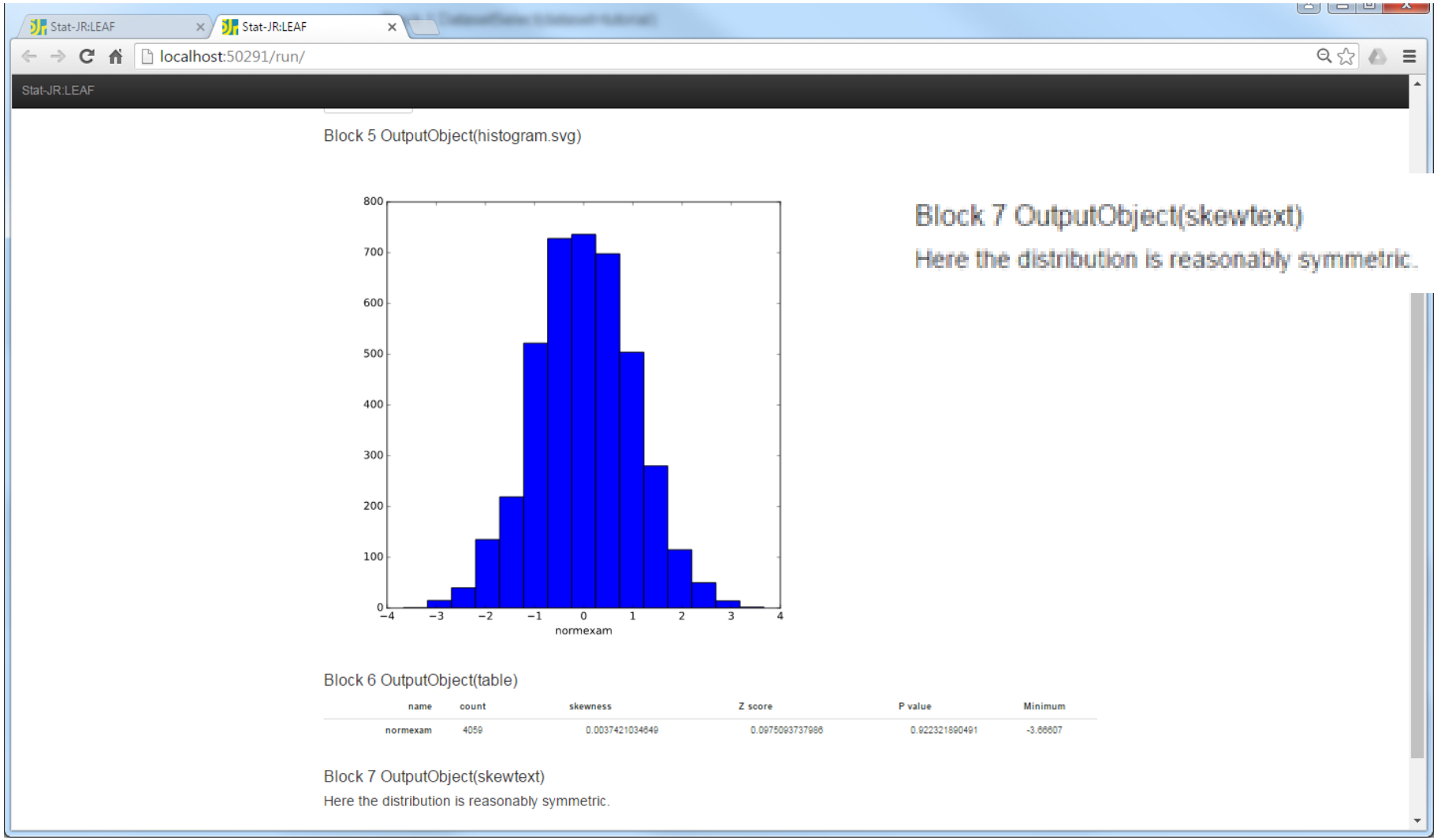
- Workflows allow the sequencing of a series of operations to perform an analysis.
- StatJR LEAF is based around a new front end written using the Blockly system.
- It allows the user to link up templates themselves in a user-friendly visual way.
- Work flows can then be included in eBooks.
- We will use this system in the SAAs.

Skewness / Histogram workflow



- Here is a logfile style workflow.
- Basically we select a dataset then fit a histogram to a variable and display several objects.

Skewness / Histogram workflow



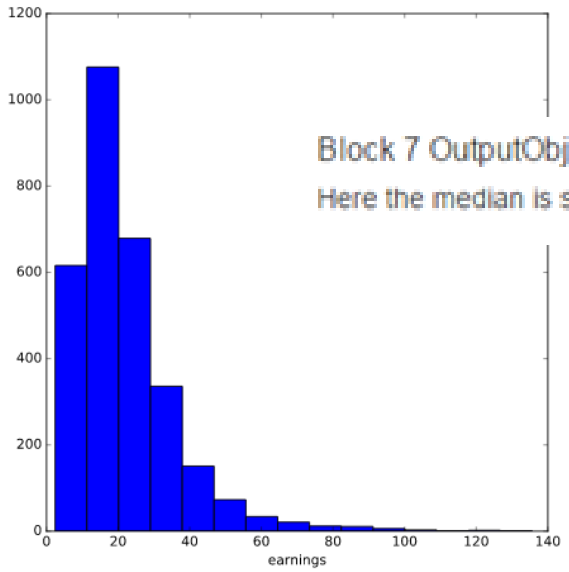
Skewness / Histogram workflow

Stat-JR:LEAF | Stat-JR:LEAF | Stat-JR:LEAF

localhost:50291/run/

Stat-JR:LEAF

Block 5 OutputObject(histogram.svg)



Block 7 OutputObject(skewtext)

Here the median is smaller than the mean and there is significant skew to the right.

Block 6 OutputObject(table)

name	count	skewness	Z score	P value	Minimum
earnings	3022	2.192631959915	31.537733852492	0.64122554237e-218	2.4

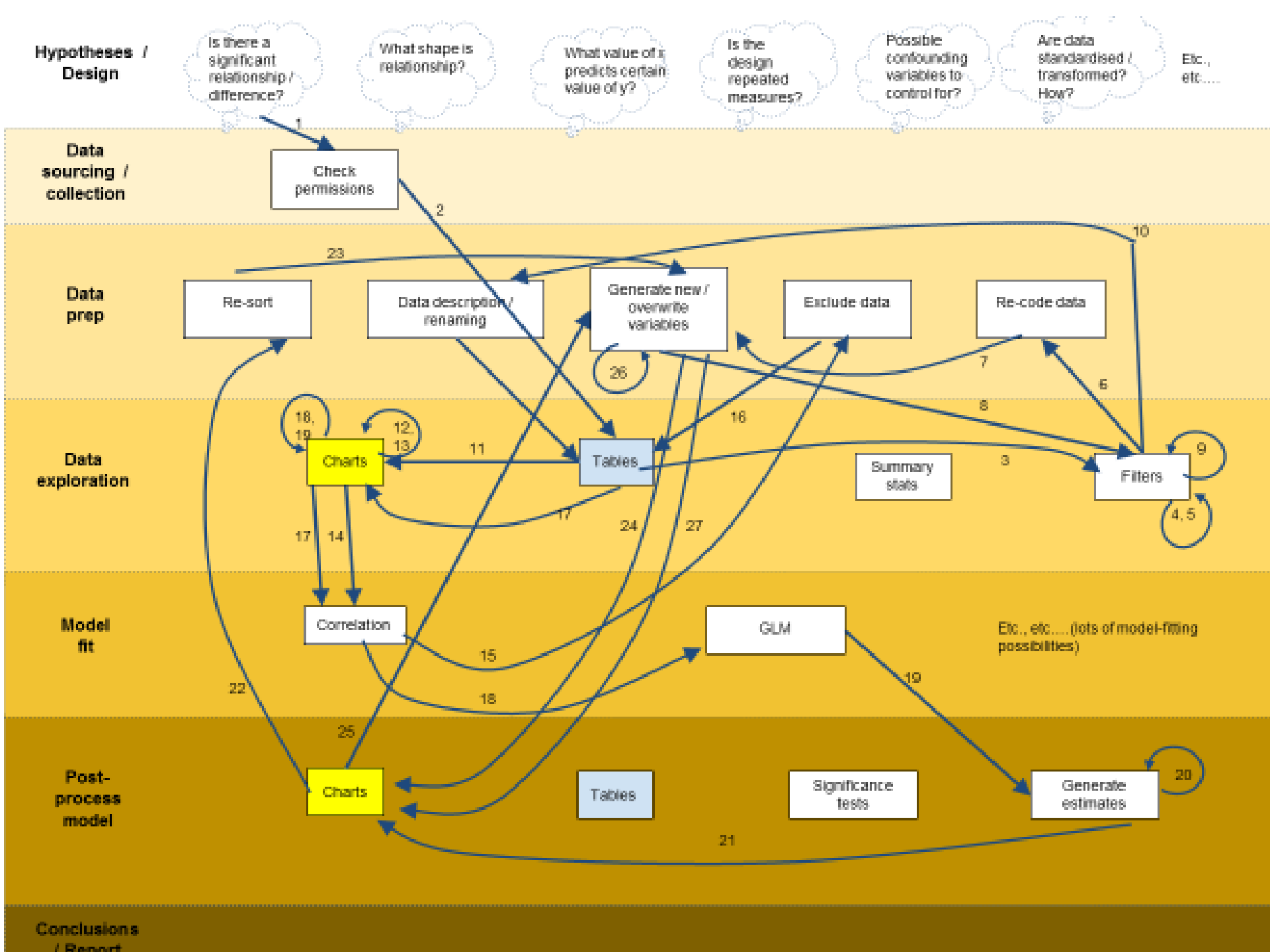
Block 7 OutputObject(skewtext)

Here the median is smaller than the mean and there is significant skew to the right.

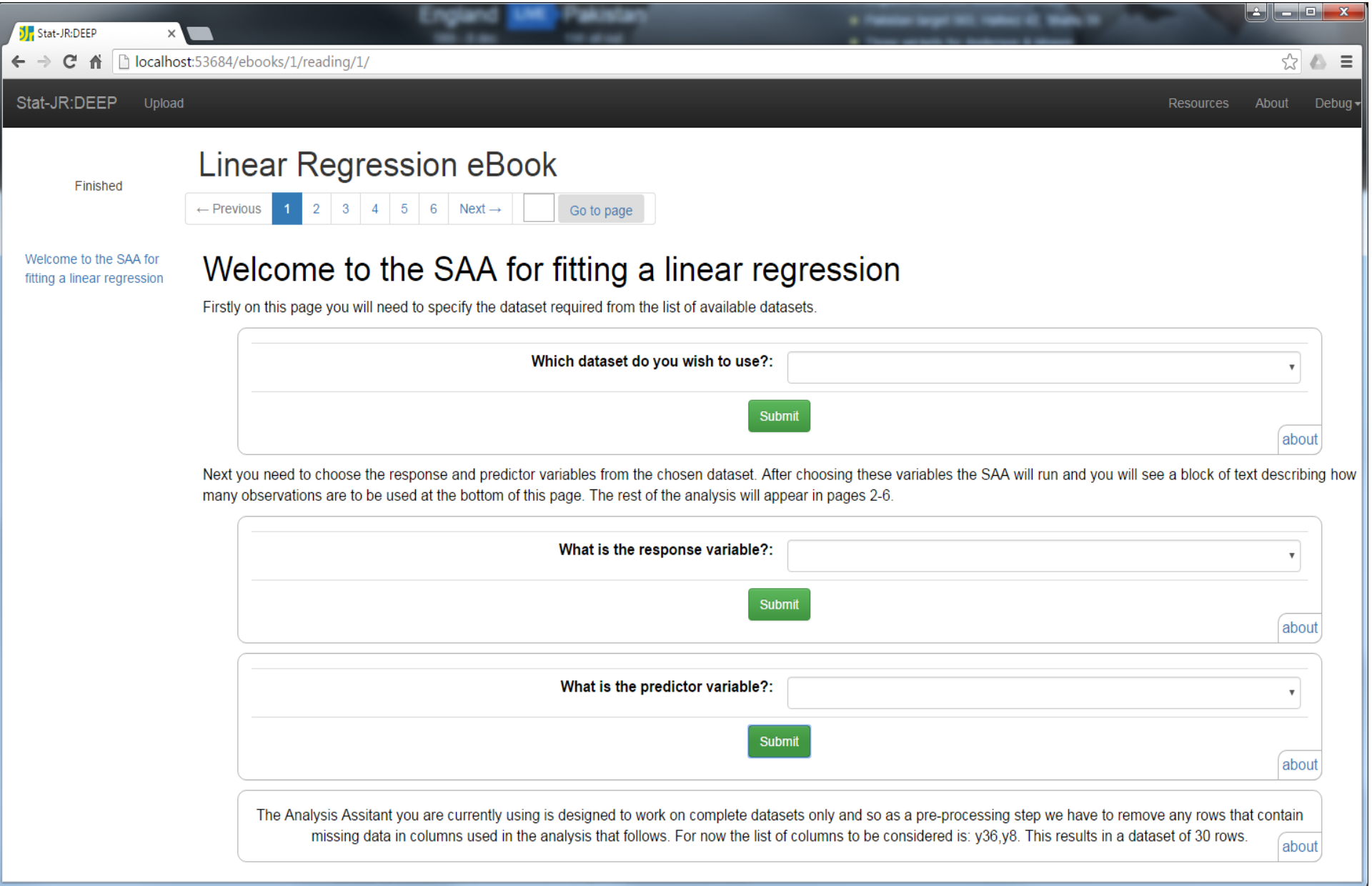
Provenance

More complex operations – linear regression

- When we looked at the chi-squared test earlier we already broke the test down into a series of steps which formed the test.
- For a regression analysis we might have additional steps to translate from simply a test to an analysis.
- We might do some initial exploratory data analysis and possible transform variables.
- We will clearly do the model fit itself but we will probably then also do some post-processing steps – for example analysis of the residuals and plotting the model predictions
- We will demonstrate an SAA for a linear regression but first show an example of a flow-chart for a real analysis.


Hypotheses / Design
Is there a significant relationship / difference?
What shape is relationship?
What value of x predicts certain value of y?
Is the design repeated measures?
Possible confounding variables to control for?
Are data standardised / transformed? How?
Etc., etc....
Data sourcing / collection
Check permissions
Data prep
Re-sort
Data description / renaming
Generate new / overwrite variables
Exclude data
Re-code data
Data exploration
Charts
Tables
Summary stats
Filters
Model fit
Correlation
GLM
Etc., etc.....(lots of model-fitting possibilities)
Post-process model
Charts
Tables
Significance tests
Generate estimates
Conclusions / Report
1
2
3
4, 5
6
7
8
9
10
11
12, 13
14
15
16
17
18
18, 19
19
20
21
22
23
24
25
26
27

Linear regression eBook



Linear regression eBook

Stat-JR:DEEP

localhost:53684/ebooks/1/reading/1/#page2

Stat-JR:DEEP Upload Resources About Debug

Finished

Linear Regression eBook

← Previous 1 2 3 4 5 6 Next → Go to page

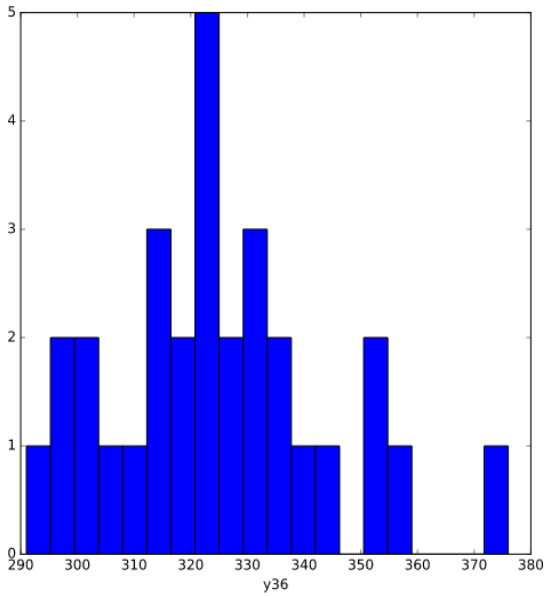
Welcome to the SAA for fitting a linear regression

We will begin our analysis of the dataset by doing some basic data exploration.

You have chosen y36 as your response variable and so a first step is to take a look at this variable and assess its suitability for a normal model. The summary statistics for the variable are in the table below

Observations	30
Mean	324.8
Standard Deviation	19.132
Median	323.5

We also look at a histogram of y36 to see if it is approximately normally distributed. Although in modelling the response in terms of a set of predictors it is what is unexplained (the model residuals) that need to be normally distributed, it is still useful to look at the response variable as a very skewed variable will often lead to very skewed residuals.



Here the distribution is reasonably symmetric with skewness value 0.518.

There are no obvious outliers in y36.

about

Linear regression eBook

Stat-JR:DEEP

localhost:53684/ebooks/1/reading/1/#page4

Stat-JR:DEEP Upload Resources About Debug

Finished

Linear Regression eBook

← Previous 1 2 3 4 5 6 Next → Go to page

Welcome to the SAA for fitting a linear regression

Exploratory analysis continued - Looking at variables in pairs

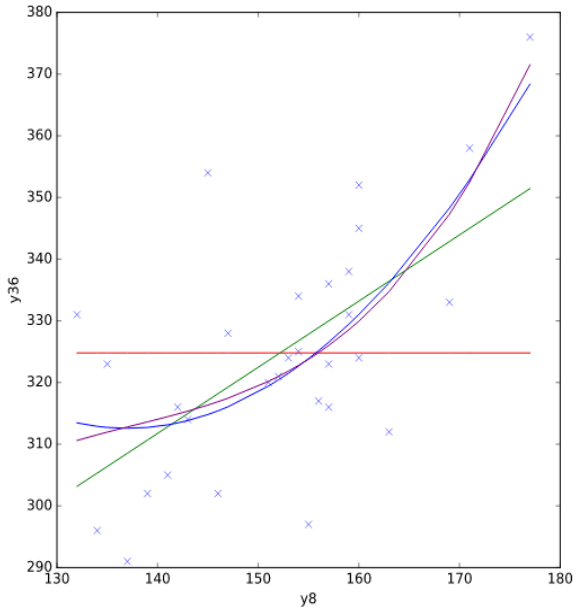
Once we are happy with our response variable and our predictor variable we now want to have a preliminary look at them together before progressing to the linear regression.

For the predictor we can look at correlations with the response and scatterplots with best fitting curves to see if there is a linear relationship.

Predictor : y8

The Pearson correlation between y36 and y8 is 0.615 (s.e. 0.000295).

The Spearman correlation between y36 and y8 is 0.559 (s.e. 0.00131).



The graph includes best fitting curves for a constant, linear, quadratic and cubic relationship between y36 and y8.In this case a linear relationship is most appropriate.

about

Linear regression eBook

Stat-JR:DEEP

localhost:53684/ebooks/1/reading/1/#page5

Stat-JR:DEEP Upload Resources About Debug

Finished

Linear Regression eBook

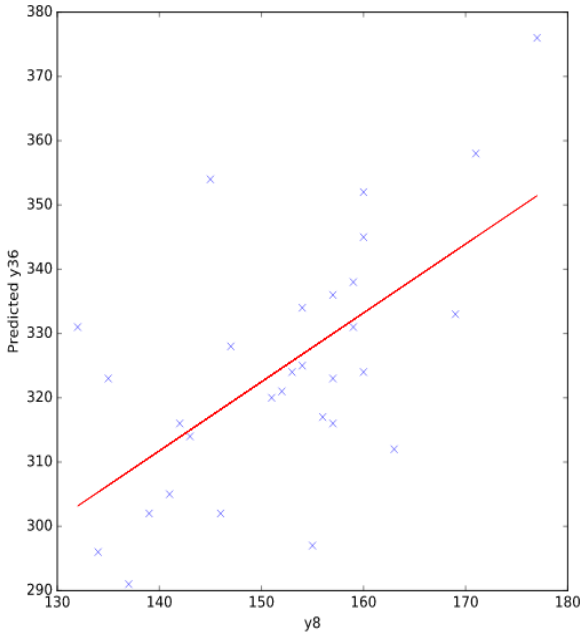
Welcome to the SAA for fitting a linear regression

Here we simply fit the linear regression model for our chosen predictor

Variable	Coefficient	SE	P value	Significance
y8	1.073	0.26	< 0.001	***
Intercept	161.6	39.61		
sigmasq	243.6			

We can plot a predicted regression line to describe the model. This is shown below:

Prediction plot for response vs y8

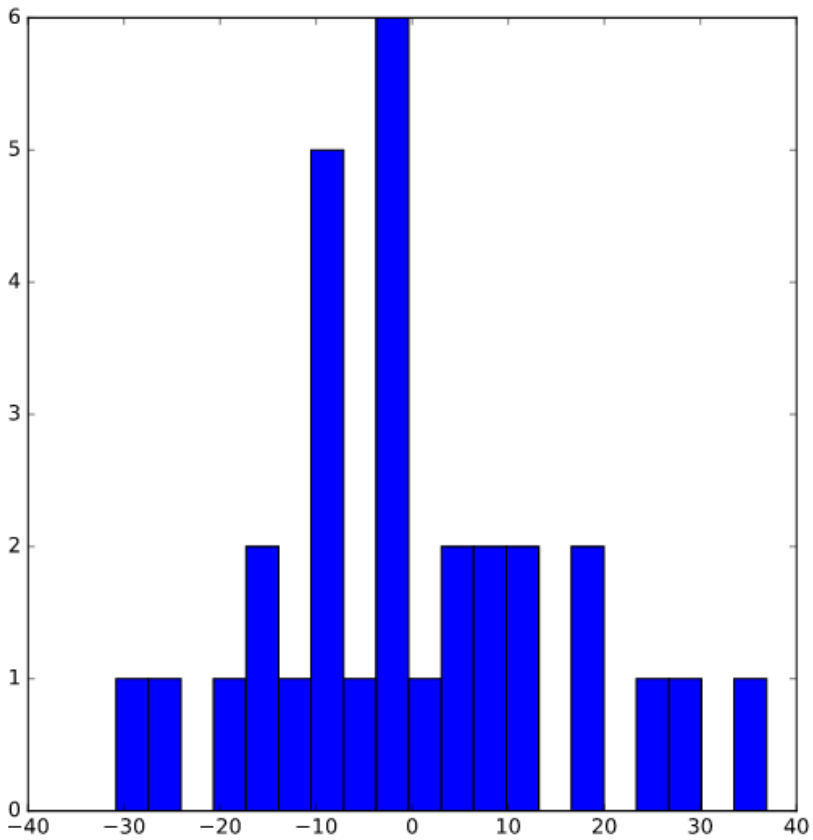


about

Linear regression eBook

Here we look at the residuals from the model and plot them in various ways.

Histogram of residuals

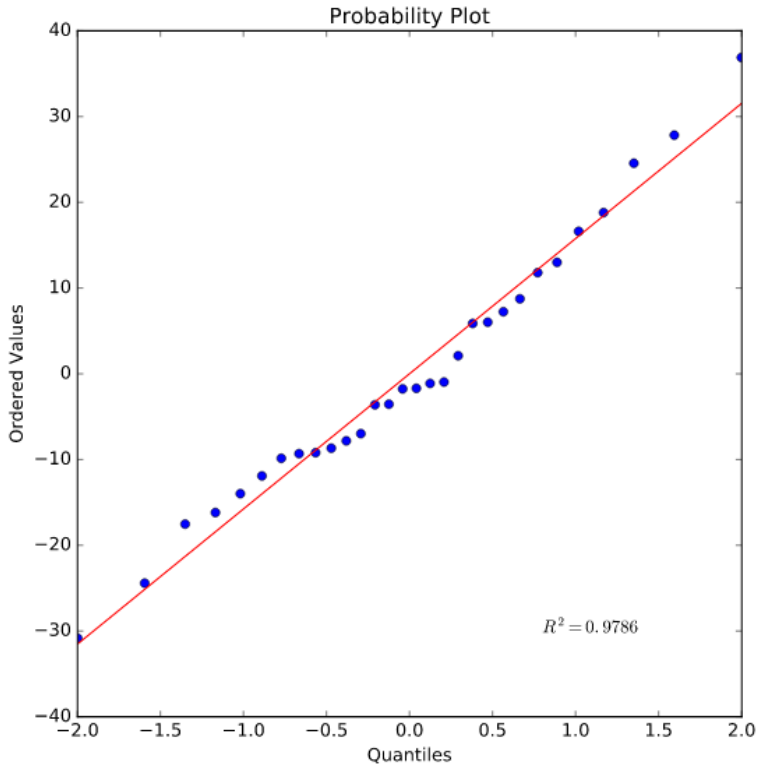


Here the distribution is reasonably symmetric with skewness value 0.399.

There are no obvious outliers in the residuals.

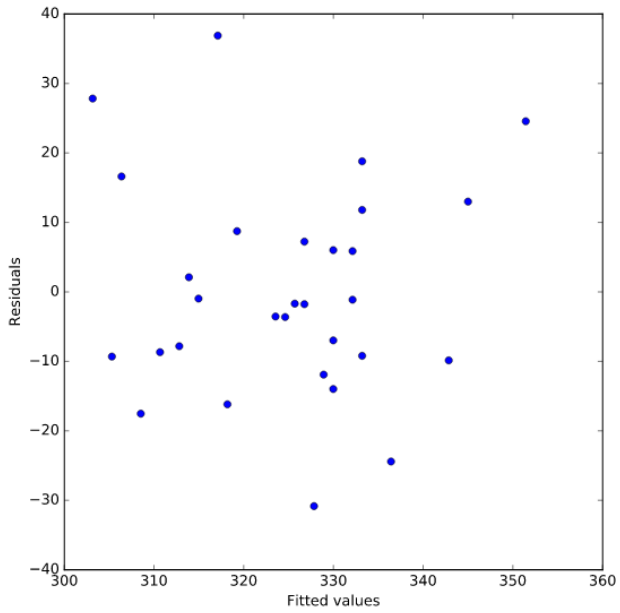
Linear regression eBook

Quantile-Quantile Plot of residuals



If the residuals are fairly normally distributed then the points in this graph should be close to the red line.

Residuals vs fitted values

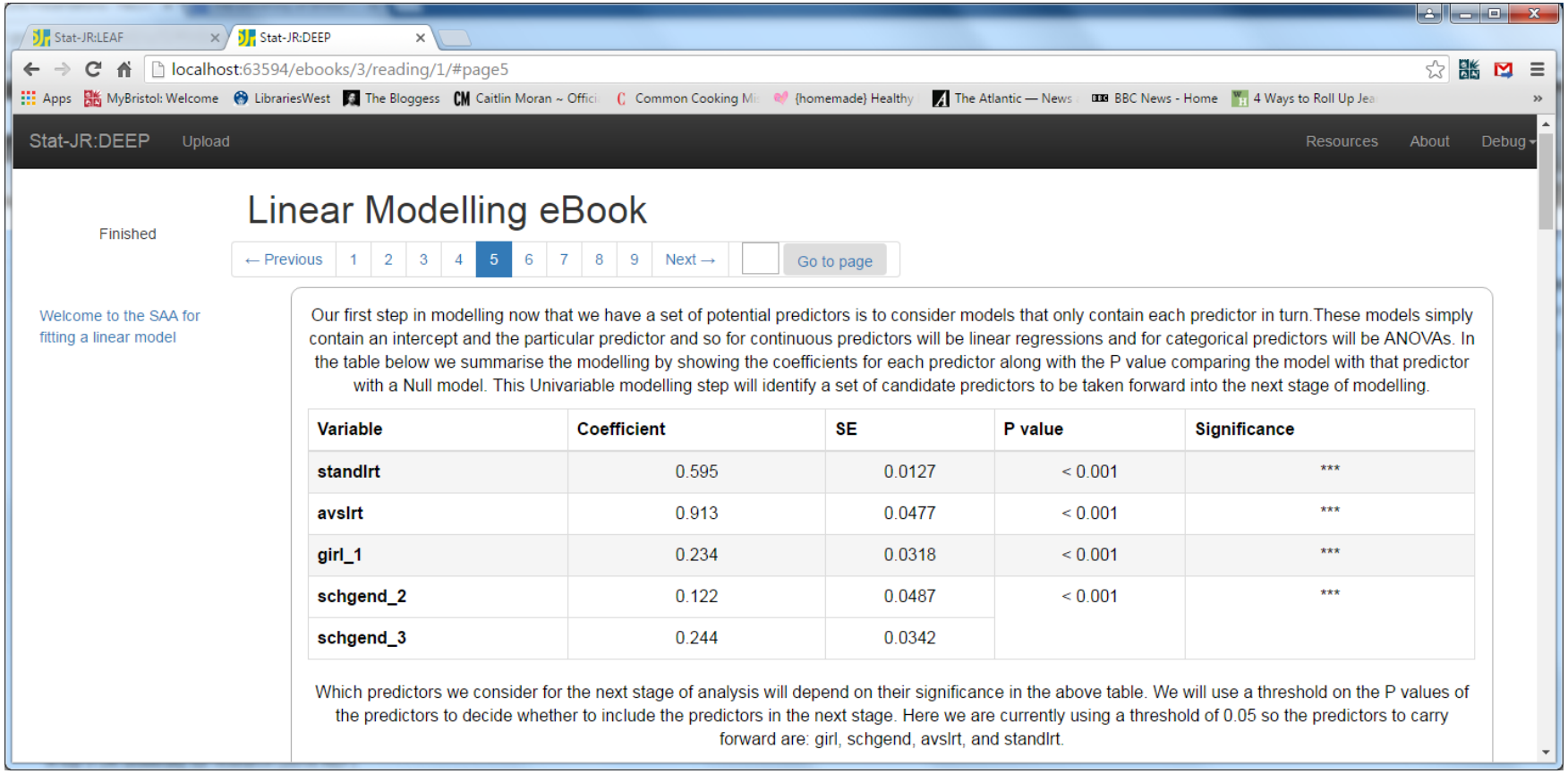


Here you should consider whether there are any patterns in this plot. Ideally we would like to see similar variability of the residuals across the range of fitted value

Moving to general linear models

- Here we have to deal differently with categorical predictors both in how they are included in the model and in also in how we perform exploratory data analysis on them.
- We might perform ‘univariable analysis’ where each predictor is considered in isolation and a separate model is fitted.
- We can then consider ‘multivariable analysis’, possibly via some stepwise style approach to find a ‘best’ model.
- Residual analysis is straightforward to extend to general linear models but what is more of a challenge is automation of prediction plots when say one has 3 continuous and 4 categorical predictors!
- One possible solution is to plot against each predictor in turn holding the others at their mean or offering a bespoke prediction tool.

Linear Modelling eBook

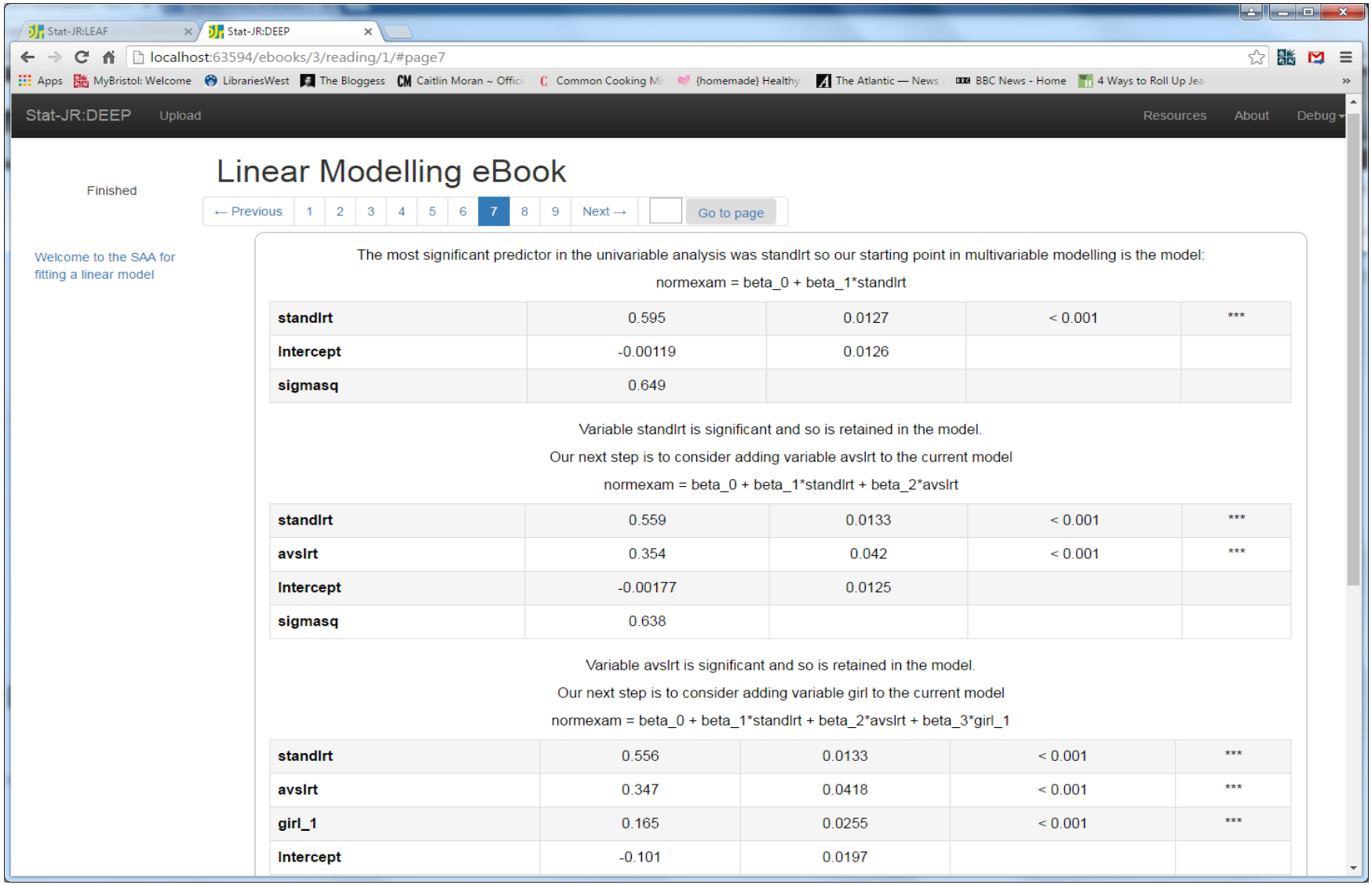


Our first step in modelling now that we have a set of potential predictors is to consider models that only contain each predictor in turn.These models simply contain an intercept and the particular predictor and so for continuous predictors will be linear regressions and for categorical predictors will be ANOVAs. In the table below we summarise the modelling by showing the coefficients for each predictor along with the P value comparing the model with that predictor with a Null model. This Univariable modelling step will identify a set of candidate predictors to be taken forward into the next stage of modelling.

Variable	Coefficient	SE	P value	Significance
standlrt	0.595	0.0127	< 0.001	***
avslrt	0.913	0.0477	< 0.001	***
girl_1	0.234	0.0318	< 0.001	***
schgend_2	0.122	0.0487	< 0.001	***
schgend_3	0.244	0.0342		

Which predictors we consider for the next stage of analysis will depend on their significance in the above table. We will use a threshold on the P values of the predictors to decide whether to include the predictors in the next stage. Here we are currently using a threshold of 0.05 so the predictors to carry forward are: girl, schgend, avslrt, and standlrt.

Linear Modelling eBook



More on Statistical Analysis Assistants

- We have produced a far wider selection of SAAs than we have covered in these slides.
- We have SAAs that deal with other response types – for example binary responses and counts.
- We also have SAAs for multilevel models.
- We also have SAAs that use Bayesian MCMC methods.
- For more details see http://www.bristol.ac.uk/cmm/media/software/statjr/downloads/manuals/1-06/manual-saa.pdf

NCRM National Centre for Research Methods

For more information visit
www.ncrm.ac.uk