Statistical reports

Regression, 2010

Niels Richard Hansen
June 10, 2010

This document gives some guidelines on how to write a report on a statistical analysis. The document is organized into sections that are reasonable to include in a report, each with some remarks on what to put into those sections, and simple examples on how to write and document your analysis.

Below are some (personal) rules that I believe are useful. Note that this is about writing a report on a statistical analysis that is more exploratory than confirmatory.

- *Clarity over formality.* Try to write in clear text and in simple terms the actions you take and decisions you make. Don't try to cover it up with formalities or pretend that you can come up with a priory justifications of all things done and tried.

- *Try, don't speculate.* Investigate and analyze the data and see what happens. Don't write long texts with speculations about what could or could not be interesting/problematic/fun etc.

- *Report, comment and conclude.* Report your findings, make relevant comments. Try to make it easy for the reader to follow your actions.

- *Thorough over extensive analysis.* For this report I prefer if you are careful and thorough with the different aspects of the analysis rather than try out a huge number of different things.

Introduction

General content: All reports should start with an introduction stating the purpose of the analysis/study, the background and anticipations, a verbal description of the data used (but no statistics here), perhaps the experimental setup and/or study design if relevant. Some find that it is a good idea to include major conclusions drawn from the analysis at the end of the introduction. Some may believe that this is confusing.

Your report: You should have an introduction, but you have little background knowledge. It does not need to be more than a few paragraphs and max half a page. Try to narrow down what your work is going to be about.

Descriptive analysis or Preliminary analysis

General content: Before any formal statistical modeling is introduced the data should be described statistically. This includes summary computations and plots of means, medians, variances, histograms, correlations, scatter plots etc. for continuous variables (response and explanatory variables) and tabulations and perhaps cross-tabulations of discrete variables. Associations between discrete and continuous variables may sometimes be better illustrated with histograms decomposed according to the value of the discrete variable rather than using a scatter plot. Continuous variables may also – experimentally – be cut into groups, which can be used for cross-tabulations.

We may also include various plots of the response variable against explanatory variables decomposed according to other discrete variables or other continuous variables that are cut into groups.

In addition, we should investigate if there are missing observations and where they are.

Your report: Though this is certainly an important part of a real statistical analysis, this does not play a huge role in the course. You should "take a look" at the data and report some descriptions, but the initial analysis can be a small research project in itself. I don't want your report/work to grow out of control because of an extensive preliminary analysis. No more than 2-3 pages.

The objective: This section documents and describes data. It provides ideas about relations and the quality and nature of the data. On the basis of the preliminary analysis you may find that you need to exclude certain variables/observations from the further analysis, e.g. if there are missing observations or strange/unreasonable observations. For the purpose of the regression analysis you should in particular be aware of correlations between explanatory variables and the range of the variables. If, for instance, age is an explanatory variable, but we have observed only individuals in the range 35-42 years, then if the objective is to investigate risk factors for a heart disease, the narrow age range may be problematic.

Main analysis or Statistical analysis

General comments: This is the section where you set up one or more statistical models, fit the models to data, investigate the model fit and compare models.

Many of the practical tools for this section have been developed during the course. The example analysis of the South Africa Heart Disease data provides an illustration of some of the ways one can proceed. You don't need to include all output for all models considered. Select the most relevant/interesting parts, and comment in details on the selected material. You are welcome to briefly comment on things you tried but don't want to report.

You should also here be explicit about model assumptions and criticize and discuss these in the light of the different model diagnostic tools you apply.

Your report: This should be the major part of your report. Around 8-10 pages is appropriate (depending a little on how many figures/tables you include, and how large they are).

Example: (This could be the first part of an analysis section)

As stated in the introduction our main objective is to find risk factors for myocardial infarction. Our response variable is thus the dichotomous variable `chd` giving whether or not the individual has suffered from myocardial infarction. We will in this section set up and analyze different regression models for the probability of `chd=1` given the explanatory variables.

As a fundamental model assumption throughout the analysis we assume

- independence of the response variables given the explanatory variables in our data set.

Initially we will, furthermore, assume that

- the logistic regression model is correct,
- that the explanatory variables enter linearly in the linear predictor (no transformations needed),
- and that there are no interactions among the explanatory variables.

Thus the initial model we consider is the logistic regression model with the linear predictor given as

`age+famhist+ldl+tyepa+tobacco+sbp+obesity+alchohol`

The resulting parameter estimates, standard errors and p-values for the parallel tests of the parameters being equal to 0 are given in Table 1. From this model an immediate observation is that the coefficients for the three variables `sbp`, `obesity`

Variables	Estimate (std. err)	z value	Pr(>|z|)
Intercept	−6.4170 (1.2402)	−5.17	0.0000
`age`	0.0490 (0.0106)	4.62	0.0000
`famhistPresent`	0.9234 (0.2277)	4.06	0.0000
`ldl`	0.1824 (0.0583)	3.13	0.0017
`typea`	0.0389 (0.0123)	3.17	0.0015
`tobacco`	0.0796 (0.0266)	2.99	0.0028
`sbp`	0.0067 (0.0057)	1.18	0.2400
`obesity`	−0.0422 (0.0294)	−1.43	0.1517
`alcohol`	0.0003 (0.0045)	0.06	0.9501

Table 1: Parameter estimates for the initial logistic regression model with all potential explanatory variables included in the model.

and `alchohol` do not seem to be significantly different from 0. To investigate this further we did an analysis of deviance where we excluded all the three variables. This resulted in a difference in deviance of 3.14, and using the χ^2-distribution with 3 degrees of freedom the corresponding p-value is 0.37. Thus we can not reject the hypothesis that all these three variables can be excluded from the model.

From hereon I would investigate model diagnostics, include plots and discuss those. If there are any obvious problems (perhaps outliers), I would reiterate the initial analysis. Then I would turn to the three model assumptions stated above. I would investigate if there are any transformations that could be suggested – most importantly for those variables excluded, and/or if there are interactions that should be included. Then I might turn to the choice of link function, but for the particular example with the retrospective design, the link function is more or less required to be the logit link. In any case, it rarely makes a big difference which link function you choose for a dichotomous regression model. Finally, I might consider automatic model selection procedures. Sometimes it is necessary to iterate some of the processes. There is no fixed formula here. You might, for instance, like to be more mathematical in your formulations of models, or more systematic/mathematical in the formulations of hypotheses investigated than I was above.

Results

General comments: In the light of the analysis one or a few models come out of the analysis as appropriate models. Reporting the models in a complete way including confidence intervals for parameter estimates of interest is important. You should also interpret the resulting model and parameter estimates.

Your report: This should be a short section, 1-2 pages. One may include the results as a subsection of the previous section.

Variables	Estimate (std. err)	95% conf. int.
`age`	0.0505 (0.0103)	[0.03;0.07]
`ldl`	0.0704 (0.0592)	[−0.05;0.19]
`typea`	0.0369 (0.0123)	[0.01;0.06]
`tobacco`	0.0839 (0.0262)	[0.03;0.14]
`ldl:famhistPresent`	0.2009 (0.0421)	[0.12;0.28]

Table 2: Estimated parameters, standard errors and 95% confidence intervals for the resulting logistic regression model.

Example:

Our resulting model is a logistic regression model with a linear predictor of the form

```
age+ldl+tyepa+tobacco+ldl:famhist
```

The estimated parameters are given in Table 2 together with 95% confidence intervals. The confidence intervals are computed using the profile log-likelihood. We observe from the table that all explanatory variables contribute with an increase of the risk for increased value of the variable. Considering age, for instance, the odds increases by a factor $e^{0.505} \simeq 1.66$ per 10 years. We also observe that the significance of the cholesterol level as measured by `ldl` depends upon whether there is a family history or not. This is both in terms of statistical significance and in terms of the estimated parameters. If there is no family history the odds is estimated to increase by only a factor $e^{0.0704} \simeq 1.07$ per unit of `ldl`. This is not even statistical significant on a 5% level. If there *is* a family history the odds is estimated to increase by a factor $e^{0.2009+0.0704} \simeq 1.31$ per unit of `ldl`, and this is a highly statistically significant effect.

Notes: We left out the intercept parameter above on purpose. Since this is a retrospective study it does not really hold any relevant information. Try to report only those parameters in the model that are of relevance. In particular, don't report subject specific parameters should there be such parameters in the model.

Discussion and/or Conclusion

General comments: Some like to have a separate section where they discuss the results before the conclusion, e.g. typically treating various problems with the data and/or the analysis that could be brought up to criticize the conclusions. Some like to combine conclusion and discussion. In general, I would say that any technical discussion about the appropriateness of the statistical models should be in the analysis section above and not here. More general methodological issues, such as lack of known important explanatory variables, data quality issues, suspicion of a bias in the sample etc. are appropriate to discuss here.

Don't just repeat the reporting of the resulting model here, but put the conclusion in the context of the whole data set, e.g. are there variables not included in the resulting model, and what does that mean?

Your report: Don't spend too much time speculating. Highlight your most important/relevant findings and make sure that the conclusion matches the purpose from the introduction. If you have any reason to cast doubt about the conclusion you may find it appropriate to have a small discussion. Max half a page.

Appendices

You can include R-code and/or R-output in an appendix if you like. I don't think it should be included in the report directly, and I don't think you should constantly refer to e.g. parameter estimates or the like in the appendix.