Department of Statistics

COURSE STATS 330/762

Model answer for Assignment 3, 2012

Task 1. *You task is to write a report of no more than 6 pages (including appendices) describing the effect of gender on gambling behaviour. The client may wish to have your report peer reviewed by another statistician so you should support your conclusion by a suitable analysis, described in sufficient detail so that a reviewer can follow your reasoning. [25 marks]*

Rather than give a sample report, here is a summary of a likely structure, and a list of the things you should mention in your report:

Structure

You should have an executive summary, an introduction, a section describing the data and methods, a section summarizing the analysis in a non-technical way, and a conclusion. This should be followed by a technical appendix, describing the analysis in more detail.

Executive summary
This should give a one or two sentence description of the problem (Does gender have an effect on gambling?) and a short description of your answer. In fact there is strong evidence that there is a sex effect. For girls, gambling seems unaffected by income, status or verbal score. For boys, gambling is strongly related to income, but not the other variables.

Introduction

This should address the following points:

The contact made by the client
The question being posed (effect of gender on gambling)
The likely effect of the other factors (confounders)

Data and Methods

This should describe the data collection, the variables and how they are measured, and the decision to use regression analysis as a way of answering the question. You should explain how the model is to be used to answer the question (the gender effect) and in particular, what parameter(s) in the model are the crucial ones. Since the aim here is to interpret a

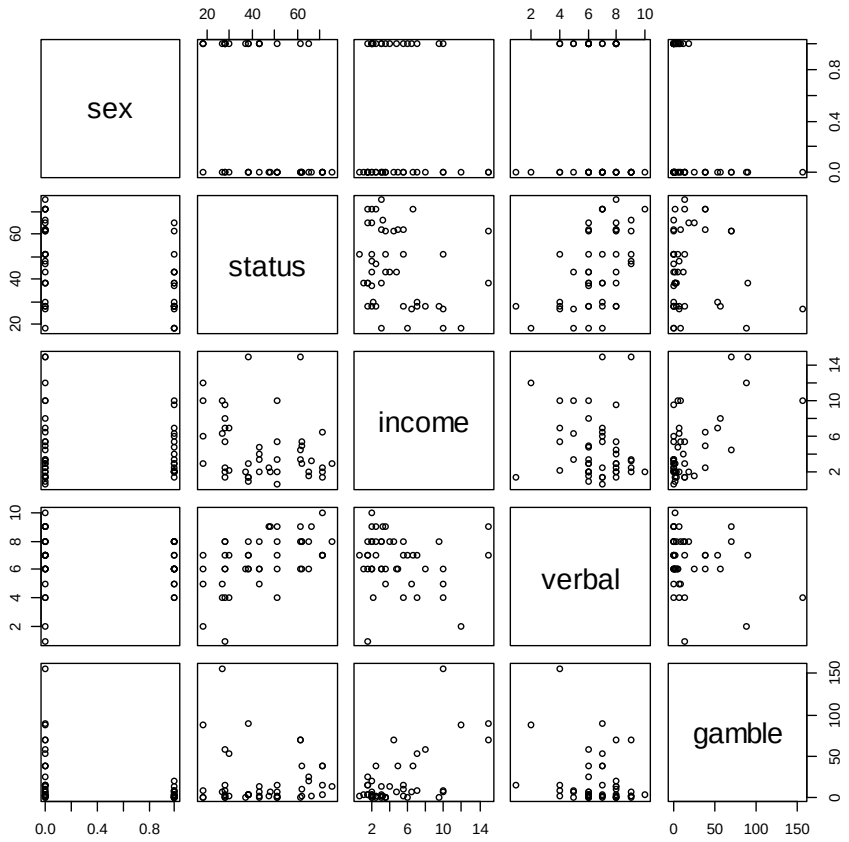
coefficient, and that we have other possible causes of gambling, we should initially include all available confounders in the model – i.e. fit the full model rather than a sub-model.

Analysis: You should clearly state

- The model being fitted
- The interpretation of the relevant coefficient
- Any caveats or reservations you might have about the model
- Other factors you think might affect gambling behaviour
- Translate the interpretation of the coefficient into everyday language

Conclusions/discussion: Repeat your conclusion from the analysis section, and any qualifications you might have, expressed in everyday language.

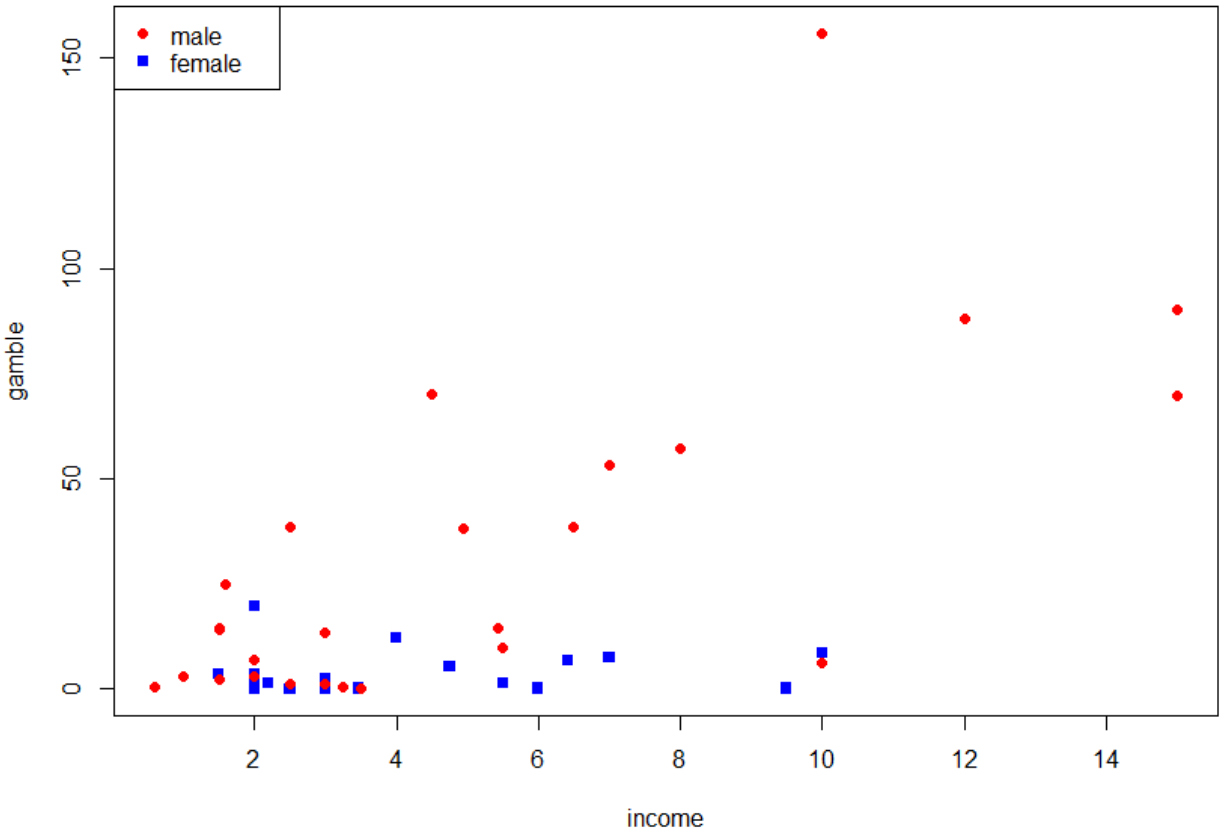
Technical appendix: Here I give more detail about the analysis: The first step, as always should be to explore the data graphically. A pairs plot is shown below:



Points to note

Clearly boys gamble more than girls. But is this due to their gender, or other factors? The pairs plot shows (weak) decreasing relationships between gamble and status and verbal, and a slightly stronger increasing relationship between gamble and income. There is not much relationship between the explanatory variables.

If we look more carefully at the plot of gamble versus income, and code the points by gender, we see the following:

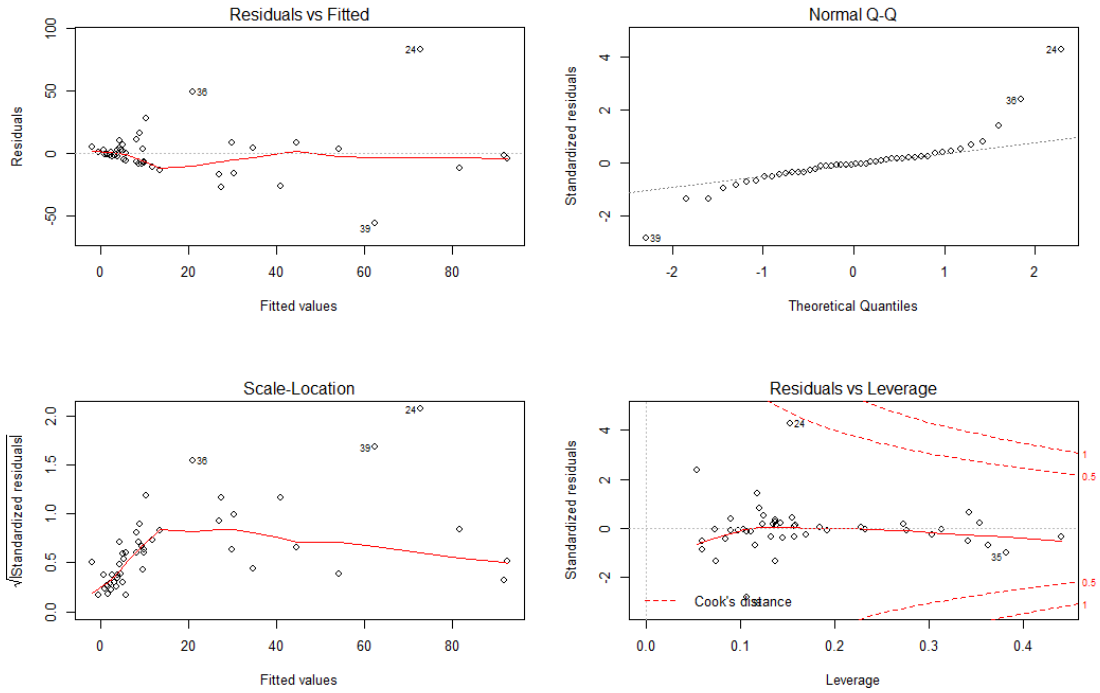


It seems that there are two sets of relationships: one for the males (an increasing relationship between income and gambling, with increasing variability about the regression line for higher incomes) , and a flat relationship for the females.

This strongly suggests that we need an interaction between sex and income, and possibly for the other explanatory variables as well. Accordingly we fit the model

gamble~ sex*income + sex*verbal + sex*status

```
> gamb.lm = lm(gamble~sex*status+sex*income+sex*verbal, data=gamble.df)
> summary(gamb.lm)

Call:
lm(formula = gamble ~ sex * status + sex * income + sex * verbal,
    data = gamble.df)

Residuals:
    Min      1Q  Median      3Q     Max
-56.422  -7.112  -1.163   3.779  83.266

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.1301    17.8832   1.629   0.1116
sex         -34.5080    35.3690  -0.976   0.3354
status       -0.1712     0.3361  -0.509   0.6134
income        6.0635     1.0625   5.707 1.44e-06 ***
verbal       -3.1020     2.4509  -1.266   0.2133
sex:status    0.3785     0.5544   0.683   0.4989
sex:income   -5.3822     2.4420  -2.204   0.0337 *
sex:verbal    2.9628     4.6335   0.639   0.5264
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.13 on 38 degrees of freedom
Multiple R-squared: 0.6288,     Adjusted R-squared: 0.5604
F-statistic: 9.196 on 7 and 38 DF,  p-value: 1.322e-06

> par(mfrow=c(2,2))
> plot(gamb.lm)
```



There are several large residuals, corresponding to the fanning out of the male gambling.

There is a suggestion in this summary that status and verbal are not important variables. (in fact, the model dropping status and verbal seems quite adequate.) It is pretty clear that income has a strong effect for males, and that the effect is different for males and females., since the interaction between sex and income is strongly significant. We could test if the slope for females is zero by using the test.lc function – the desired slope is the sum of the income and sex:income coefficients. In fact this is not significantly different from zero:

```
gamb.int.lm = lm(gamble~ status+sex*income+verbal, data=gamble.df)
> test.lc(gamb3.lm, c(0,0,0,1,0,0,1,0),0)
$est
[1] 0.6813044

$std.err
[1] 2.198697

$df
[1] 38

$t.stat
[1] 0.3098673

$p.val
[1] 0.7583549
```

The coefficient of sex is not significant. However, the term sex should be retained in the model since a zero true regression coefficient implies that the male and female lines have the same intercept, which does not seem sensible.

Another issue is that for the males (but not for the females) the variance of the response seems to increase with the mean. This impression is largely due to points 24, 39, 36. These seem to be genuine data points, so there is no real reason to exclude them. Still, we can see what difference they make to the regression by excluding them :

```
Call:
lm(formula = gamble ~ sex * status + sex * income + sex * verbal,
    data = gamble.df, subset = -c(24, 39, 46))

Residuals:
    Min      1Q  Median      3Q     Max
-23.173  -5.313  -0.584   3.017  47.271

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.21140   11.48099   0.889  0.37985
sex         -15.58924   22.26740  -0.700  0.48850
status        0.08883    0.22064   0.403  0.68969
income        6.07559    0.70161   8.660 3.18e-10 ***
verbal       -2.53011    1.56954  -1.612  0.11594
sex:status    0.11848    0.35311   0.336  0.73922
sex:income   -5.39429    1.54342  -3.495  0.00131 **
```

```
sex:verbal    2.39087    2.91691   0.820  0.41796
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.21 on 35 degrees of freedom
Multiple R-squared: 0.7693,     Adjusted R-squared: 0.7232
F-statistic: 16.68 on 7 and 35 DF,  p-value: 1.878e-09
```

Although the fit is now better (as the variance estimate is smaller), the conclusions are essentially unchanged.

It is possible to fit a variety of smaller models (e.g. without interactions between verbal and sex and status and sex, or even dropping all variables involving status and verbal). However, these do not involve large changes in income and income:sex coefficients, and the conclusions are essentially unchanged.

In summary, we conclude:

1. Gambling behaviour is different for males and females.
2. For females, none of the variables income, verbal or status seem to have much effect on gambling.
3. For males, gambling is strongly (positively) related to income, but not to the variables status and verbal.

Instructions to markers

Marks breakdown: give 5 for clear report structure (introduction , main section, conclusions, appendix), 5 for general clarity, proper spelling and understandable sentences, 10 for a good analysis and 5 for a clearly expressed conclusion.

The analysis must identify the importance of the income and sex:income interaction. It doesn't matter too much which model is used but it must include income, sex and the sex:income interaction. (if the sex:income interaction is omitted deduct 7 marks from the 10).

Task 2. *As well as examining the effect of gender, a secondary purpose of the research is to construct a scale which will allow the researcher to predict how much an individual is likely to gamble. Construct a prediction equation from the data supplied and estimate the prediction error. Test your prediction by predicting the gambling expenditure for a teenage boy whose parents have a socioeconomic status score of 71, who has an income of £2.50 and scores 9 out of 12 words correct on the verbal score.*
[15 marks]

We will calculate cross-validated prediction error estimates for the model fitted above and also examine submodels.

```
> allpossregs(gamb.int.lm)
```

```
      rssp  sigma2 adjRsq     Cp    AIC    BIC       CV status sex income verbal sex:income
1 27867.02 633.341  0.376 21.394 67.394 71.051 2725.058      0   0      1      0          0
2 19028.94 442.534  0.564  3.288 49.288 54.774 1911.681      0   0      1      0          1
3 17706.49 421.583  0.585  2.280 48.280 55.594 1873.113      0   0      1      1          1
4 17609.15 429.491  0.577  4.058 50.058 59.202 1888.442      0   1      1      1          1
5 17583.45 439.586  0.567  6.000 52.000 62.972 1946.511      1   1      1      1          1
```

The best model has a CV of 1873, but we don't usually consider models that have interactions without the main effects (in this case the model without sex is constraining the male line to go through zero.) So we pick the model omitting status.
(Note CV here is not the estimated prediction error, but is proportional to it.)

The prediction error can be estimated with the function cross.val:

```
> gamb3.lm = lm(gamble~sex*income+verbal, data=gamble.df)
> cross.val(gamb3.lm, nrep=1000)
Cross-validated estimate of root
mean square prediction error =  21.99572
>
> gamb4.lm = lm(gamble~sex*income+verbal+status, data=gamble.df)
> cross.val(gamb4.lm, nrep=1000)
Cross-validated estimate of root
mean square prediction error =  22.38842
>
> gamb5.lm = lm(gamble~sex*income, data=gamble.df)
> cross.val(gamb4.lm, nrep=1000)
Cross-validated estimate of root
mean square prediction error =  22.33889
```

It appears that the prediction based on **`gamble~sex*income+verbal`** has the smallest prediction error of these three models. We will base our prediction on his model.

```
> newdata.df = data.frame(sex = 0,income=2.5, verbal = 9, status=71)
> gamb3.lm = lm(gamble~sex*income+verbal, data=gamble.df)
> predict(gamb3.lm, newdata=newdata.df, interval="p")
       fit       lwr      upr
1 7.308571 -36.19019 50.80733
```

The amount spent is predicted to be *£7:31*

(the prediction based on the model dropping status and verbal is £13.35)

Markers: Give 5 marks for model selection, 5 marks for calculating the CV error, and 5 marks for the prediction. Record the prediction on a separate sheet to be delivered to me.

Total for assignment: 40 marks