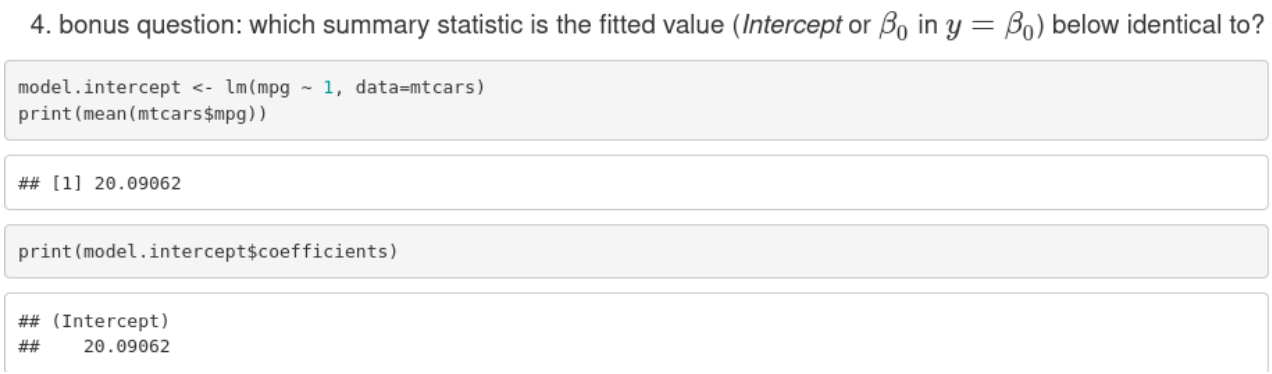
General feedback

Practical exercise 1



4. bonus question: which summary statistic is the fitted value (*Intercept* or β_0 in $y = \beta_0$) below identical to?

```
model.intercept <- lm(mpg ~ 1, data=mtcars)
print(mean(mtcars$mpg))
```

```
## [1] 20.09062
```

```
print(model.intercept$coefficients)
```

```
## (Intercept)
##    20.09062
```

It is the mean

Applying the inverse *logit* function to the linear predictors returns the fitted values

You should make a quadratic model for the logistic model – do not plot and compare the old quadratic model

Be careful that you get it right when – some have stated that there is only 3 % chance of the Pontiac having automatic transmission – it's the other way around. It's 3 % chance of having manual transmission

Please check that the documents knits before handing it in

Practical exercise 2

Remember that the function *sigma* can be used to retrieve the residual standard deviation.

$$\text{residual standard deviation} = \sqrt{\left(\frac{\sum_{i=1}^{n} \epsilon_i^2}{df}\right)}$$

And here is the formula: $\sum_{i=1}^{n} \epsilon_i^2$: variance of the residuals: (unexplained variance) df: degrees of freedom; n_observations minus n_model_parameters . Remember that *df* needs to be calculated for each model. There is no fixed one

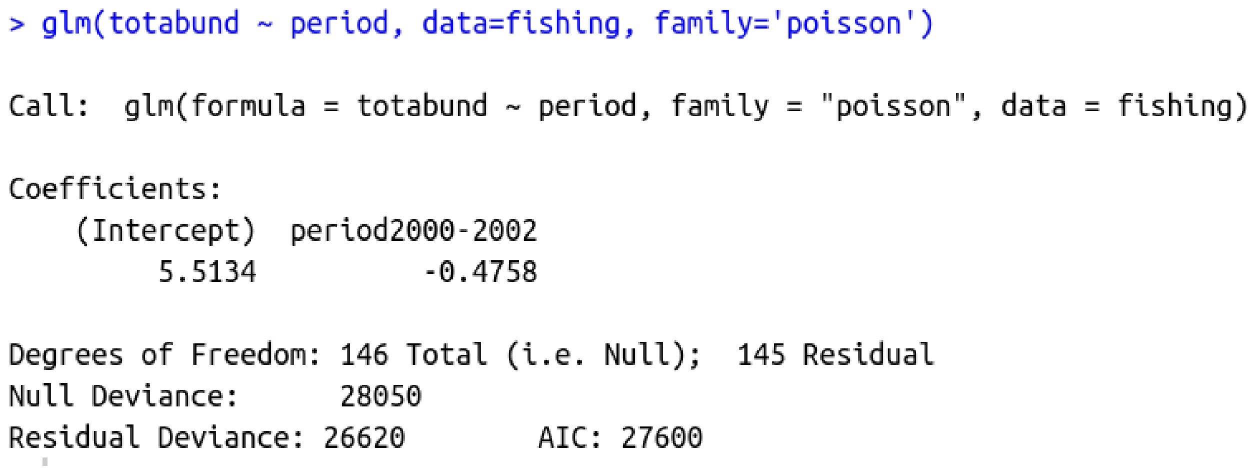
Also remember that the assumption of the general linear model is **not** that the data, *y*, is normally distributed, but that the **residuals**, ε, are.

When you knit into a pdf, make sure that text isn't cut out. html is the safe option.

If you use the Sattherwaites approximation for the calculation of the degrees of freedom, be prepared to argue why this is appropriate.

Practical exercise 3

Remember when you apply the inverse function, e.g. *exp*, to an estimated coefficient it doesn't necessarily express a meaningful number. With the following coefficients, for example:



```
> glm(totabund ~ period, data=fishing, family='poisson')

Call:  glm(formula = totabund ~ period, family = "poisson", data = fishing)

Coefficients:
    (Intercept)  period2000-2002
         5.5134          -0.4758

Degrees of Freedom: 146 Total (i.e. Null);  145 Residual
Null Deviance:      28050
Residual Deviance: 26620        AIC: 27600
```

The intercept (1977-1989) can be found by doing *exp(5.5134),* 248 fish. For the following period it would be *exp(5.5134 - 0.4758),* i.e. 154 fish. *exp(-0.4758)* isn't very meaningful in itself.

In general, when doing model comparison, do not use single t-tests or z-tests between two levels, e.g. *singles* and *pairs*, to justify whether or not a given factor should be included among the independent variables, in this example *task*. Instead compare the model overall on measures such as AIC and residual standard variance – or the log-likelihood ratio test which we will come to in week 5.

It is tempting to show the full summary of a model, but in general only show the parts that are relevant to the question. That allows me to understand what you want me to focus on. For example, don't use *summary*, if you want to just show or comment on the fixed effects, then just use *fixef*.

Make sure that you model PAS as a factor, not as a numeric variable.

Practical exercise 5

Remember to also model the no correlation between target.frames and subject can be done by using "||" instead of "|"

Remember that if you want the probability of being correct on say PAS2, you cannot just take inv.logit on its estimated value – inv.logit has to be applied on the sum of the Intercept and PAS2. The beta value for PAS indicated the log odds, not the probability

In Exercise 7, remember to fit a line for each PAS, don't collapse them

Remember to try different optimizers, if the default doesn't converge

Practical exercise 8

In the covariance matrix in 1.1.iii, you may find it easier to interpret if you standardise the dataset first (you weren't asked to do it – so it is not necessary for the exam). The unstandardised plot is dominated by a high covariance in the middle of the plot – which may make it harder to see other colours (compare to the Z covariance matrix in Exercise 10)

Addendum to the above: in this case, it didn't change the visibility much. In plt.imshow(), you can set vmin and vmax – if you have standardised, you can try vmax=200

Be careful when using logical testing between two arrays, e.g. a == b. There might be small differences in them because of numerical imprecision – use something like np.isclose(a, b) instead.

Regarding noise in 1.1.viii – think about what happens if you average a lot of samples that has a structured part (signal) and an unstructured part (noise), which is μ=0 with some standard deviation. What is retained in the limit where you have unlimited samples?

Practical exercise 10

Regarding the five last dimensions – if they are noise dimensions why is the inclusion of them in 2.1.ii not increasing classification accuracy (it's flat) - we would expect an increase if they were noise, since we more have more parameters (predictor variables) now to that noise.

And conversely, why doesn't it decrease in 2.2.ii – if you are introducing noise, why doesn't prediction become worse, as it normally does when you are fitting noise.