

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.

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

$\sum_{i=1}^n \epsilon_i^2$: variance of the residuals: (unexplained variance)
 df : degrees of freedom; $n_{\text{observations}}$ minus $n_{\text{model_parameters}}$

And here is the formula: . 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 Satterwaites 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.