Announcements

- Make sure to sign in on the google form (linked here)
- Pset 6 due October 28 at 5 pm
- Project proposal due October 28 at 5 pm

From RSS to BIC

When describing Bayes Information Criterion, the lecture notes leave the equation at

$$\text{BIC} = 2\ln(g(\text{SSE})) + (p+1)\log(n)$$

where g is some mysterious likelihood function. Wikipedia asserts (with citation but without proof) that for a Gaussian model, $\text{BIC} = n\ln(\text{RSS}/n) + p\ln(n)$ where their p includes the intercept. In this problem, we'll derive the result for ourselves in our usual notation.

1. First, recall that for a multiple regression model, $Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i} + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Also recall that for this distributional assumption, $\hat{\vec{\beta}}$ is the set of parameters that maximize the likelihood function of the whole model. Lastly, recall that in a multiple regression model, the maximum likelihood estimate for the residual variance is $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y_i})^2}{n}$ (and note that this is different from our unbiased estimator). Write the maximized likelihood function for the observed data as a function of $\hat{y}_i$, y_i, and $\hat{\sigma}^2$.
2. Write the maximized log likelihood function of the observed data as a function of the residual sum of squares (RSS). (You will find there are two terms that are constant regardless of the predictors; these can be dropped because we are only interested in comparing AIC between models.)
3. Find the Bayes Information Criterion (where the Bayes Information Criterion is $(p+1)\ln(n) - 2\ln(\hat{L})$ and $\hat{L}$ is the maximized likelihood function).

The Red Queen's R^2

1. Recall the formula for adjusted R^2_{adj}:

$$1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Consider a model with p predictors where the unadjusted R^2 is R^2_p. What unadjusted R^2_i would a model with i predictors need to have so that the adjusted R^2_{adj} remains unchanged?

2. For what p does adding an additional predictor require the smallest increase in unadjusted R^2 for the adjusted R^2 to remain the same? For what p does adding an additional predictor require the greatest increase? What are the increases in unadjusted R^2 in both cases?

Step procedures and cross validation

1. Given the following table, find the model produced by forward selection using an ESS F-test and starting from a model with only an intercept. (You should be able to do this with only a single test.)

Model Variables	**Residual sum of squares**	**Degrees of freedom**
None	7,200	38
X_1	6,600	37
X_2	6,980	37
X_3	6,760	37

The rest of this section will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here.

A few useful columns:

- `mad_gdppc`: GDP per capita
- `bi_fishes`: Number of endangered fish species
- `bi_fungi`: Number of endangered fungi species
- `bi_mammals`: Number of endangered mammal species
- `bi_reptiles`: Number of endangered reptile species
- `bi_molluscs`: Number of endangered mollusc species
- `bi_othinverts`: Number of other endangered invertebrate species

2. The next three questions will ask you to run forward, backward, and both-direction variable selection procedures. Briefly glance ahead and predict which model will have the highest R^2.

3. Run a forward variable selection procedure to predict log GDP per capita from endangered species statistics starting with an intercept only model and using an upper scope of all the two-way interaction terms for the variables listed above. Report this model's coefficient estimates, R^2, and AIC.

```
# TODO: Forward step model
```

4. Run a backwards variable selection procedure to predict log GDP per capita from endangered species statistics starting with all interaction terms of the variables listed above and using a lower bound of an intercept-only model. Report this model's coefficient estimates, R^2, and AIC.

```
# TODO: Backward step model
```

5. Run a both-direction variable selection procedure to predict log GDP per capita from endangered species statistics starting with a model including all variables listed above (but no interactions) and using a lower bound of an intercept-only model and an upper bound of a model with all the interaction terms. Report this model's coefficient estimates, R^2, and AIC.

```
# TODO: Both direction step model
```

6. Based on AIC, which model is the best? Why didn't the other procedures find the same model?

7. Recall from last week that we looked at various models incorporating the following variables:

- `wdi_araland`: Arable land (% of land area)
- `wdi_precip`: Average annual precipitation (mm per year)

Run k-fold cross validation with $k = 10, 20, 50$ to estimate out-of-sample RMSE for a LOESS model and a degree 2 polynomial model to predict the proportion of arable land from the country's average annual precipitation. Which model performs better for each k?

```r
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
set.seed(139)

for (ncross in c(10, 20, 50)) {
  # TODO: Run cross validation for the polynomial model
}

for (ncross in c(10, 20, 50)) {
  # TODO: Run cross validation for the LOESS model
}
```