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 $\hat{\sigma}^2$.

$$\prod_{i=1}^{n} \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}^2}\right]$$

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.)

$$\begin{aligned}
\log(\hat{L}) &= \sum_{i=1}^{n} -\log\left(\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}\right) - \log(\sqrt{2\pi}) - \frac{1}{2}\frac{(y_i - \hat{y}_i)^2}{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \\
&= \sum_{i=1}^{n} -\frac{1}{2}\log\left(\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}\right) - \log(\sqrt{2\pi}) - \frac{n}{2}\frac{(y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \\
&= -\frac{n}{2}\left[\log\left(\frac{\text{RSS}}{n}\right)\right] - n\log(\sqrt{2\pi}) - \frac{n}{2}
\end{aligned}$$

Since $n\log(\sqrt{2\pi})$ and $\frac{n}{2}$ do not depend on the predictors, we can drop these terms, so we end up with

$$-\frac{n}{2}\left[\log\left(\frac{\text{RSS}}{n}\right)\right]$$

Note that the part that actually mattered from the likelihood function was the normalizing constant!

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).

$$(p+1)\ln(n) + n\log\left(\frac{\text{RSS}}{n}\right)$$

In the original formulation of BIC, the first term should actually be $(p+2)\ln(n)$ because we are fitting p predictors, an intercept, and a residual standard error. However, this only changes the resulting BIC by a constant for all models of this type, so we (and R) drop it.

```
# Fit example model
countries <- read.csv("data/countries.csv")
model1 <-lm(wdi_araland~poly(wdi_precip, 2, raw = TRUE), countries)

# R's AIC
extractAIC(model1, k=log(length(model1$residuals)))[2]
```

```
## [1] 926.6248
```

```
# AIC by hand
n <- length(model1$residuals)
3 * log(n) + n * log(sum(model1$residuals^2)/n)
```

```
## [1] 926.6248
```

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?

For the adjusted R^2_{adj} to remain the same, we need:

$$1-(1-R^2_p)\frac{n-1}{n-p-1} = 1-(1-R^2_i)\frac{n-1}{n-i-1} \implies$$
$$R^2_i = 1-(1-R^2_p)\frac{n-i-1}{n-p-1}$$

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?

The difference in unadjusted R^2 required is

$$1-(1-R^2_p)\frac{n-(p+1)-1}{n-p-1} - R^2_p = (1-R^2_p)\left[1-\frac{n-p-2}{n-p-1}\right]$$

Since everything else is a multiplicative or additive constant, this is minimized when $\frac{n-p-2}{n-p-1}$ is largest, which is when $p=0$. In particular, since $R^2_0 = 0$, a single predictor needs to give an unadjusted R^2 of $1-\frac{n-2}{n-1} = \frac{1}{n-1}$ to give an adjusted R^2 of 0.

Likewise, the difference is maximized when $\frac{n-p-2}{n-p-1}$ is smallest, which happens when p is at the largest value for which the adjusted R^2 of the next p is defined: $p = n - 3$. For this p, the difference in R^2 needs to be

$$\frac{1}{2}(1 - R_p^2)$$

for the adjusted R^2 to be unchanged.

Also, note that for all p the required difference is maximized when R_p is 0.

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

Since the model with X_1 has the smallest residual sum of squares, we will test for it being a better predictive model than the intercept-only model. Our test statistic is

$$F = \frac{(7200 - 6600)/1}{6600/37} \approx 1.17$$

which we test using a $F_{1,37}$ distribution.

```
f_stat <- (7200 - 6600)/(6600/37)
1 - pf(f_stat, 1, 37)
```

```
## [1] 0.07470329
```

We get a p-value of $0.074 > 0.05$, so we fail to reject the null and conclude that X_1, X_2, and X_3 add no predictive power on their own.

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.

We expect the backwards variable selection procedure to give a model with the highest R^2 because it starts with the largest model and is therefore most likely to overfit. There is essentially a multiple hypothesis testing scenario taking place here in which the backwards variable selection procedure requires us to fail to reject the null far more times to end up with a model as small as that of the forward or both-direction procedures.

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.

```
lm1 <- lm(log(mad_gdppc) ~ 1, countries)
step_model_1 <- step(lm1,
                     scope = list(upper = formula(lm(log(mad_gdppc) ~ bi_fishes * bi_fungi * bi_mammals
                                                     bi_othinverts * bi_reptiles, countries))),
                     direction = "forward", trace = F)
summary(step_model_1)$coefficients
```

```
##                              Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)             9.156780e+00 1.564892e-01 58.513815 1.415772e-107
## bi_fungi                4.490360e-02 7.868532e-03  5.706731  5.689633e-08
## bi_mammals             -3.966940e-02 7.342030e-03 -5.403056  2.423709e-07
## bi_othinverts           6.216929e-03 1.459769e-03  4.258844  3.550071e-05
## bi_fishes               2.518018e-03 2.194244e-03  1.147556  2.529201e-01
## bi_fungi:bi_othinverts -2.571611e-04 6.627271e-05 -3.880346  1.538918e-04
## bi_mammals:bi_fishes    7.170603e-05 2.427654e-05  2.953717  3.629345e-03
## bi_fungi:bi_mammals     7.617022e-04 4.014075e-04  1.897578  5.961106e-02
```

```
summary(step_model_1)$r.squared
```

```
## [1] 0.4158964
```

```
AIC(step_model_1)
```

```
## [1] 450.4468
```

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.

```
step_model_2 <- step(lm(log(mad_gdppc) ~ bi_fishes * bi_fungi * bi_mammals * bi_molluscs *
                          bi_othinverts * bi_reptiles, countries),
                     scope = list(lower = formula(lm(log(mad_gdppc) ~ 1, countries))),
                     direction = "backward", trace = F)
summary(step_model_2)$coefficients
```

```
##                                                    Estimate   Std. Error
## (Intercept)                                     9.537384e+00 2.750873e-01
## bi_fishes                                       1.065477e-02 5.265896e-03
## bi_fungi                                        2.048858e-02 2.469698e-02
## bi_mammals                                     -1.066639e-01 2.030834e-02
## bi_molluscs                                    -4.259512e-04 9.732008e-03
## bi_othinverts                                   4.273362e-03 4.996886e-03
## bi_reptiles                                     8.441780e-03 2.379807e-02
## bi_fishes:bi_fungi                              1.756326e-06 5.949248e-04
## bi_fishes:bi_mammals                            1.478432e-04 2.423106e-04
## bi_fungi:bi_mammals                             8.331860e-03 4.101076e-03
## bi_fishes:bi_othinverts                        -1.637068e-05 7.152254e-05
## bi_fungi:bi_othinverts                          4.224546e-05 5.511920e-04
## bi_mammals:bi_othinverts                        4.678131e-04 2.514044e-04
## bi_molluscs:bi_othinverts                      -2.913071e-05 7.272467e-05
## bi_fishes:bi_reptiles                          -7.501769e-04 3.579829e-04
## bi_fungi:bi_reptiles                           -4.567867e-03 3.827641e-03
## bi_mammals:bi_reptiles                          3.009983e-03 6.469104e-04
## bi_molluscs:bi_reptiles                        -1.182918e-03 4.534948e-04
## bi_othinverts:bi_reptiles                      -2.557438e-04 3.664497e-04
## bi_fishes:bi_fungi:bi_mammals                  -1.231320e-04 6.719869e-05
## bi_fishes:bi_fungi:bi_othinverts               -6.228704e-06 8.061064e-06
## bi_fishes:bi_mammals:bi_othinverts             -2.239350e-06 1.685264e-06
## bi_fungi:bi_mammals:bi_othinverts              -9.269786e-05 8.416823e-05
## bi_fishes:bi_fungi:bi_reptiles                  2.068335e-04 8.318510e-05
## bi_fungi:bi_mammals:bi_reptiles                -1.988221e-04 6.718364e-05
## bi_fishes:bi_othinverts:bi_reptiles             4.522570e-06 3.314888e-06
## bi_fungi:bi_othinverts:bi_reptiles              1.024596e-04 6.974791e-05
## bi_mammals:bi_othinverts:bi_reptiles           -1.256087e-05 2.891824e-06
## bi_molluscs:bi_othinverts:bi_reptiles           8.828719e-06 3.189834e-06
## bi_fishes:bi_fungi:bi_mammals:bi_othinverts     1.710768e-06 7.800937e-07
## bi_fishes:bi_fungi:bi_othinverts:bi_reptiles -1.981382e-06 8.254462e-07
##                                                    t value     Pr(>|t|)
## (Intercept)                                    34.670394648 3.630078e-68
## bi_fishes                                       2.023354304 4.505566e-02
## bi_fungi                                        0.829598651 4.082635e-01
## bi_mammals                                     -5.252221238 5.860064e-07
## bi_molluscs                                    -0.043768067 9.651554e-01
## bi_othinverts                                   0.855205085 3.939872e-01
## bi_reptiles                                     0.354725327 7.233615e-01
## bi_fishes:bi_fungi                              0.002952181 9.976490e-01
## bi_fishes:bi_mammals                            0.610139159 5.428186e-01
## bi_fungi:bi_mammals                             2.031627549 4.419838e-02
## bi_fishes:bi_othinverts                        -0.228888366 8.193099e-01
```

```
## bi_fungi:bi_othinverts                          0.076643823 9.390229e-01
## bi_mammals:bi_othinverts                        1.860799063 6.499740e-02
## bi_molluscs:bi_othinverts                      -0.400561679 6.893904e-01
## bi_fishes:bi_reptiles                          -2.095566541 3.803077e-02
## bi_fungi:bi_reptiles                           -1.193389589 2.348573e-01
## bi_mammals:bi_reptiles                          4.652859825 7.858661e-06
## bi_molluscs:bi_reptiles                        -2.608448389 1.014379e-02
## bi_othinverts:bi_reptiles                      -0.697896178 4.864697e-01
## bi_fishes:bi_fungi:bi_mammals                  -1.832357421 6.915202e-02
## bi_fishes:bi_fungi:bi_othinverts               -0.772689983 4.410872e-01
## bi_fishes:bi_mammals:bi_othinverts             -1.328782790 1.862121e-01
## bi_fungi:bi_mammals:bi_othinverts              -1.101340302 2.727528e-01
## bi_fishes:bi_fungi:bi_reptiles                  2.486425289 1.415149e-02
## bi_fungi:bi_mammals:bi_reptiles                -2.959382605 3.654922e-03
## bi_fishes:bi_othinverts:bi_reptiles             1.364320579 1.747882e-01
## bi_fungi:bi_othinverts:bi_reptiles              1.468998524 1.442124e-01
## bi_mammals:bi_othinverts:bi_reptiles           -4.343581682 2.771295e-05
## bi_molluscs:bi_othinverts:bi_reptiles           2.767766997 6.455771e-03
## bi_fishes:bi_fungi:bi_mammals:bi_othinverts     2.193028554 3.005706e-02
## bi_fishes:bi_fungi:bi_othinverts:bi_reptiles -2.400377452 1.777329e-02
```

```
summary(step_model_2)$r.squared
```

```
## [1] 0.5326062
```

```
AIC(step_model_2)
```

```
## [1] 460.1131
```

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.

```
step_model_3 <- step(lm(log(mad_gdppc) ~ bi_fishes + bi_fungi + bi_mammals + bi_molluscs +
                          bi_othinverts + bi_reptiles, countries),
                     scope = list(lower = formula(lm(log(mad_gdppc) ~ 1, countries)),
                                  upper = formula(lm(log(mad_gdppc) ~ bi_fishes * bi_fungi * bi_mammals
                                                     bi_molluscs * bi_othinverts * bi_reptiles,
                                                  countries))),
                     direction = "both", trace = F)
summary(step_model_3)$coefficients
```

```
##                              Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)             9.154028e+00 1.565081e-01 58.4891545 4.507408e-107
## bi_fishes               2.205808e-03 2.277696e-03  0.9684383  3.343440e-01
## bi_fungi                4.946267e-02 7.467700e-03  6.6235478  5.532211e-10
## bi_mammals             -3.962537e-02 7.992185e-03 -4.9580143  1.860876e-06
## bi_othinverts           6.137801e-03 1.484601e-03  4.1343110  5.831148e-05
## bi_reptiles             9.413612e-04 5.120060e-03  0.1838575  8.543671e-01
## bi_fishes:bi_mammals    7.355148e-05 2.516199e-05  2.9231188  3.988243e-03
## bi_fungi:bi_othinverts -2.627610e-04 6.934266e-05 -3.7893127  2.162646e-04
## bi_fungi:bi_reptiles    7.797251e-04 4.059308e-04  1.9208324  5.660051e-02
```

```
summary(step_model_3)$r.squared
```

```
## [1] 0.4185932
```

```
AIC(step_model_3)
```

```
## [1] 451.6925
```

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

The forward procedure found a model with the lowest AIC, so it produced the best model. Step-wise variable selection is prone to getting stuck in local minima, so it is possible for different starting models to converge on different final models.

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?

```
library(caret)
set.seed(139)

poly_model <- vector(length = 3)
i <- 1
for (ncross in c(10, 20, 50)) {
  poly_model[i] <- train(wdi_araland ~ poly(wdi_precip, 2, raw = TRUE), data = countries,
      method = "lm",
      trControl = trainControl(method = "cv", number = ncross),
      na.action = na.omit)$results[2]
  i <- i + 1
}

loess_model <- vector(length = 3)
i <- 1
for (ncross in c(10, 20, 50)) {
  loess_model[i] <- train(wdi_araland ~ wdi_precip, data = countries,
      method = "gamLoess", tuneGrid = expand.grid(span = 1,degree=1),
      trControl = trainControl(method = "cv", number = ncross),
      na.action = na.omit)$results[3]
  i <- i + 1
}

cbind(poly_model, loess_model)
```

```
##      poly_model loess_model
## [1,] 12.8573    12.9569
## [2,] 12.39758   12.77906
## [3,] 11.90762   12.52899
```

The polynomial model performs better than the LOESS model for all k. The RMSE decreases for higher k because a smaller proportion of the data is reserved for testing, so more data is used to fit the model. (A higher k makes cross validation take longer though, which can be problematic for larger models.)