Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 6 due Friday 11/3

Introductions

- One question or thought related to lecture last week (Model comparison, sequential variable selection, AIC, BIC)

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

These problems will deal with a dataset of country-level statistics from UNdata, Varieties of Democracy, and the World Bank.

2. The next three questions will run forwards, backwards, and both-direction variable selection procedures. Predict which model will have the highest R^2.

3. The following runs a forward variable selection procedure to predict log2 GDP per capita in 2010 from a country's urban population, its proportion of people 60+, its patents in force, its arable land, its energy supply, and its unemployment rate. The procedure starts with an intercept only model and uses an upper scope of all the two-way interaction terms for the variables listed above. The final model is shown along with the R^2 and AIC. How many coefficients are retained? Are the p-values reliable?

```
##                          Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)         6.3702857885 0.7964490837  7.998359 2.422931e-11
## Urban               0.0761795129 0.0123527713  6.166998 4.535709e-08
## `60+`               0.0881601105 0.0179549403  4.910075 6.139864e-06
## Energy              0.0321534985 0.0066009178  4.871065 7.104493e-06
## Unemployment        0.0903061807 0.0819156580  1.102429 2.742196e-01
## Urban:Energy       -0.0003171597 0.0000733135 -4.326076 5.171373e-05
## Urban:Unemployment -0.0022178644 0.0013383950 -1.657107 1.021736e-01
```

```
##      R2     AIC
##   0.816 191.123
```

4. The following runs a backwards variable selection procedure. The procedure starts with all the two-way interaction terms for the variables listed above. The final model is shown along with the R^2 and AIC. Note how the number of coefficients changes.

```
##                           Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)           6.483943e+00 1.004594e+00  6.4542921 2.253203e-08
## Urban                 4.312621e-02 1.523653e-02  2.8304488 6.345361e-03
## `60+`                 4.387424e-01 7.185831e-02  6.1056601 8.630022e-08
## Patents              -1.351123e-06 1.137438e-06 -1.1878647 2.396467e-01
## Arable               -1.517340e-02 2.453378e-02 -0.6184699 5.386460e-01
## Energy                1.224323e-02 3.340230e-03  3.6653852 5.309699e-04
## Unemployment          3.528566e-02 6.431629e-02  0.5486271 5.853316e-01
## Urban:`60+`          -1.359053e-03 9.414613e-04 -1.4435566 1.541528e-01
## `60+`:Arable         -5.066899e-03 1.466145e-03 -3.4559327 1.022681e-03
## `60+`:Energy         -6.507928e-04 2.069186e-04 -3.1451632 2.599751e-03
## `60+`:Unemployment   -1.091942e-02 4.104291e-03 -2.6604881 1.003367e-02
## Patents:Energy       -1.277402e-08 6.236921e-09 -2.0481294 4.500579e-02
## Patents:Unemployment  5.398933e-07 1.868535e-07  2.8893941 5.391595e-03
## Arable:Energy         4.483907e-04 1.011378e-04  4.4334636 4.098485e-05
## Arable:Unemployment   4.977333e-03 2.375911e-03  2.0949153 4.048104e-02
```

```
##      R2     AIC
##   0.864 184.493
```

5. The following runs a both-direction variable selection procedure. It starts with all the coefficients and has an upper bound of all interactions. The final model is shown along with the R^2 and AIC. How does this model compare to the ones above?

```
##                          Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)         6.3702857885 0.7964490837  7.998359 2.422931e-11
## Urban               0.0761795129 0.0123527713  6.166998 4.535709e-08
## `60+`               0.0881601105 0.0179549403  4.910075 6.139864e-06
## Energy              0.0321534985 0.0066009178  4.871065 7.104493e-06
## Unemployment        0.0903061807 0.0819156580  1.102429 2.742196e-01
## Urban:Energy       -0.0003171597 0.0000733135 -4.326076 5.171373e-05
## Urban:Unemployment -0.0022178644 0.0013383950 -1.657107 1.021736e-01
```

```
##      R2     AIC
##   0.816 191.123
```

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 predicting the proportion of arable land from the precipitation.

Here, we run k-fold cross validation 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?

k	polynomial	loess
10	12.712	12.047
20	12.427	11.625
50	11.714	11.158
100	10.975	9.849
200	10.040	9.136

From BIC to SSE

In a linear model, the Bayes Information Criterion has a nice interpretation as a trade-off between the sum of squares error $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ and the number of predictors p. In particular,

$$\text{BIC} = n\log(\text{SSE}/n) + (p+2)\log(n)$$

In this problem, we'll derive the result for ourselves.

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 extra sum of squares (SSE). (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 of the fit model. Recall that with $\hat{L}$ as the maximized likelihood function,

$$\text{BIC} = (p+2)\log(n) - 2\log(\hat{L})$$

AIC as hypothesis testing

Suppose we're performing stepwise variable selection on a linear model with coefficients $\beta_0, ..., \beta_k$ and we want to compare it to a model with $\beta_0, ..., \beta_{k+1}$. Here, we'll use

$$\text{AIC} = 2(p+2) - 2\log(\hat{L})$$

1. Recall from Stat 111 the likelihood ratio test: under the null $H_0 : \theta = \theta_0$ with $\hat{\theta}$ as the MLE, asymptotically
$$\Lambda(\vec{y}) = 2\left(\log L(\hat{\theta}, \vec{y}) - \log L(\theta_0, \vec{y})\right) \sim \chi^2_1$$
Write the equivalent null hypothesis and null test statistic distribution for testing whether $\beta_{k+1} = 0$.

2. Write the difference in AICs between the larger and smaller model in terms of log likelihoods. Feel free to refer to the log-likelihood as $\log L(\beta_{k+1}|\vec{y}, \hat{\beta}_0, ..., \hat{\beta}_k)$ (where β_{k+1} should be replaced with something in each AIC). If $\text{AIC}_2 - \text{AIC}_1 < 0$, what inequality does that give?

3. Under the null that $\beta_{k+1} = 0$, what is the asymptotic probability that $\text{AIC}_2 - \text{AIC}_1 < 0$?

4. If we view each step in the model selection as a likelihood test of whether $\text{AIC}_2 < \text{AIC}_1$, what is the α level we are using for each test?