# Variable Selection Approaches

BIOS 6611

CU Anschutz

Week 14

1. Variable Selection

2. Automatic Procedures

3. Model Averaging (Advanced Topic)

# Variable Selection

# Variable Roles

We have introduced various roles a variable may play in a model:

- Variable required to address the scientific question (e.g., the primary explanatory variable)
- Assess potential confounding and adjust for it (i.e., to provide a *valid* estimate for one or more regression coefficients of interest)
- Assess interaction/effect modification (e.g., biological reasoning or a question of interest)
- Assess mediation (i.e., need to assume a causal relationship between variables)
- Increase precision or to improve prediction (e.g., accounting for machine, study, or batch effects)

# A Few Approaches for Variable Selection

Some general guidelines that have been proposed include:

1. $p < \frac{n}{10}$ or $p < \frac{n}{15}$ or $p < \frac{n}{20}$
2. Always have a minimum of 10 degrees of freedom for your MSE
3. Use literature and context knowledge to eliminate unimportant variables.
4. Eliminate variables with distributions with limited variability (e.g., everyone is 25 years old).
5. Eliminate variables with a large percentage of missing data.
6. Perform statistical data reduction: clustering, means, PCA, combine into new variables (e.g., BMI), etc.
7. Choose some iterative process.

# Iterative Processes

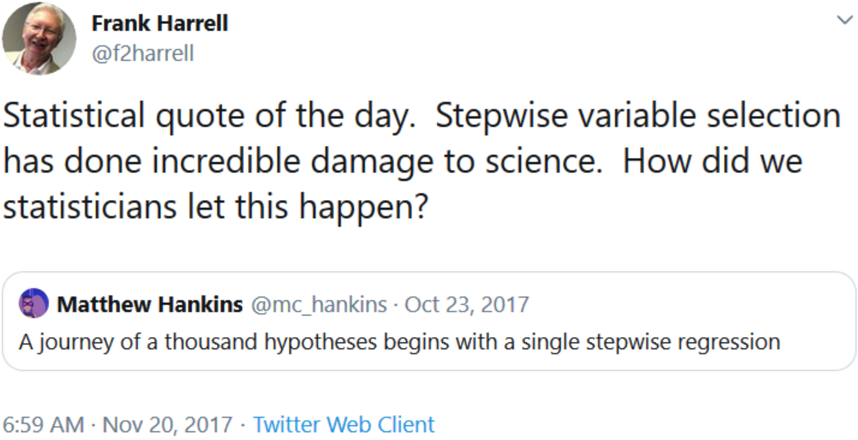
There are a few variations of iterative processes one could use:

- Run univariate analyses.
- Choose covariates for model selection with p-values $< 0.05$ (or 0.1, or 0.2, etc.)
- Run this set of variables through a stepwise model selection process (next section)

One challenge with these iterative approaches is that we do not think about polynomial trends, collinearity, or the possibility that a covariate could be insignificant univariately but significant in the multiple linear regression.

# Automatic Procedures

# A Warning



**Frank Harrell**
@f2harrell

Statistical quote of the day. Stepwise variable selection has done incredible damage to science. How did we statisticians let this happen?

**Matthew Hankins** @mc_hankins · Oct 23, 2017
A journey of a thousand hypotheses begins with a single stepwise regression

6:59 AM · Nov 20, 2017 · Twitter Web Client

*The automated variable selection procedures we will discuss need to be used carefully. They ultimately, if used as designed, remove any need for the user to use their mind and critically think!!*

# Forward Selection

Step 1. Start with a model that contains only an intercept and enter the variable most highly correlated with the dependent variable, if $p < 0.05$ (or other pre-determined $\alpha$).

Step 2. Calculate the Partial $F$ Test (or $t$ Test) for each variable not in the model based on a regression equation containing that variable and all other variables already in the model.

Step 3. Add the most significant variable if its p-value is less than some pre-selected value (e.g., an alpha level of 0.10).

Step 4. Repeat steps 2 and 3 until no additional variables can be added to the model (until all remaining variables are not significant).

*Note, all our selection algorithms can also use other criteria beyond p-values, such as model selection criterion, adjusted $R^2$, etc.*

# Backward Selection

Step 1. Start with all potential variables in the model.

Step 2. Calculate the Partial $F$ Test (or $t$ Test) for each variable in the model.

Step 3. Remove the least significant variable if its p-value is greater than some pre-selected value (e.g., an alpha level of 0.10).

Step 4. Re-compute the regression equation for the remaining variables and repeat steps 2 and 3 until all of the remaining variables are statistically significant.

# Stepwise Selection

Step 1. Start with a model that contains only an intercept and enter the variable most highly correlated with the dependent variable.

Step 2. Calculate the Partial $F$ Test (or $t$ Test) for each variable *not* in the model based on a regression equation containing that variable and all other variables already in the model. Add the most significant variable if its p-value is less than some pre-selected value (e.g., an alpha level of 0.10).

Step 3. Calculate the Partial $F$ Test (or $t$ Test) for each variable in the updated model. Remove the least significant variable if its p-value is greater than some pre-selected value (e.g., an alpha level of .15).

Step 4. Repeat steps 2 and 3 until no additional variables can be removed from or added to the model.

*The entry and exit criteria can be different. Note, alternatively, the stepwise procedure can begin with the full model.*

# Example

Let's identify the "optimal" model for the *Laryngoscope* data set from TSHS which examined two intubation techniques. We will use an outcome of ease of intubation with a subset of the available predictors in the dataset to implement each algorithm based on BIC (i.e., setting `k` to $log(n)$ for BIC, otherwise default is AIC with `k=2`).

```r
dat <- read.csv('Laryngoscope.csv')
dat <- dat[-which(is.na(dat$BMI) | is.na(dat$Mallampati)),]
n <- nrow(dat) # sample size

# note some variables are factors
dat$asa <- factor(dat$asa)
dat$Mallampati <- factor(dat$Mallampati)

# implement each algorithm + all subsets
lm_full <- lm(ease ~ age + gender + asa + BMI + Mallampati + Randomization + attempts
              + total_intubation_time + bleeding + view, data=dat)
lm_null <- lm(ease ~ 1, data=dat)

# trace=0 suppresses each step output
backward <- step(lm_full, direction='backward', k=log(n), trace=0)

# need to specify scope for full model for forward/stepwise
forward <- step(lm_null, direction = 'forward', scope = ~ age + gender + asa + BMI + Mallampati +
                  Randomization + attempts + total_intubation_time + bleeding + view, k=log(n), trace=0)

stepwise <- step(lm_null, direction = 'both', scope = ~ age + gender + asa + BMI + Mallampati +
                  Randomization + attempts + total_intubation_time + bleeding + view, k=log(n), trace=0)
```

In `SAS` we can specify `SELECTION` as `FORWARD`, `BACKWARD`, or `STEPWISE` to implement the algorithm in `PROC REG`.

# Example

We can see the resulting coefficients from each model below:

```r
coef(backward)
```

```
##   (Intercept)           age Randomization      attempts          view
##    18.2816117     0.4138817    14.1797457    24.0369055   -33.8467818
```

```r
coef(forward)
```

```
##           (Intercept) total_intubation_time                  view
##             39.237509              0.675404            -23.572445
```

```r
coef(stepwise)
```

```
##           (Intercept) total_intubation_time                  view
##             39.237509              0.675404            -23.572445
```

We see that while forward and stepwise agree, backward selection arrives at a different optimal model based on the BIC. If we did use any of these approaches, our next step should be to identify if these models make scientific sense and identify potential "manual" next steps.

# Model Averaging (Advanced Topic)

# Motivation

We know that different model selection approaches may not agree on the "optimal" model depending on the model selection criteria used or if an automated approach is employed.

Methods for model averaging have been developed that help us avoid having to select a single model for prediction or inference and instead weight the results from multiple models.

Depending on the analysis strategy, these can be frequentist (e.g., SuperLearner) or Bayesian (e.g., Bayesian model averaging).

Methods have also been developed based upon model selection criterion. Anderson and Burnham[1] describe one such approach.

We won't cover these approaches in BIOS 6611 beyond this section, but they are useful approaches to keep in mind as you move forward.

[1] Anderson, D., and K. Burnham. "Model selection and multi-model inference." Second ed. NY: Springer-Verlag 63 (2004).

# Model Averaging Example

A predictive model was developed to predict *Amblyomma americanum* populations in northeastern Missouri, with model averaging based on different model selection criterion.

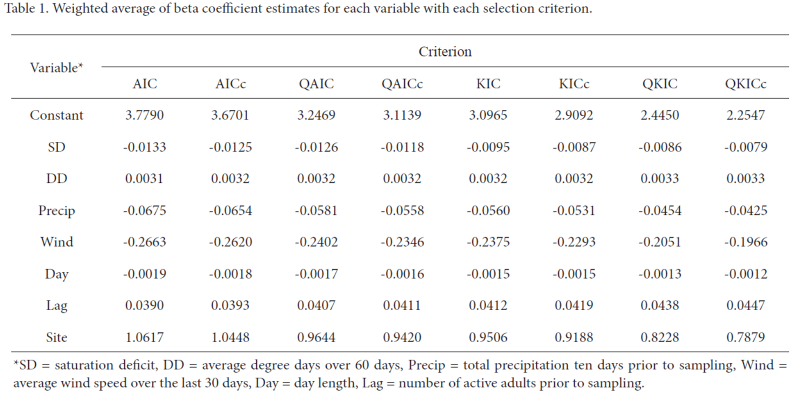


Table 1. Weighted average of beta coefficient estimates for each variable with each selection criterion.

| Variable* | Criterion | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AIC | AICc | QAIC | QAICc | KIC | KICc | QKIC | QKICc |
| Constant | 3.7790 | 3.6701 | 3.2469 | 3.1139 | 3.0965 | 2.9092 | 2.4450 | 2.2547 |
| SD | -0.0133 | -0.0125 | -0.0126 | -0.0118 | -0.0095 | -0.0087 | -0.0086 | -0.0079 |
| DD | 0.0031 | 0.0032 | 0.0032 | 0.0032 | 0.0032 | 0.0032 | 0.0033 | 0.0033 |
| Precip | -0.0675 | -0.0654 | -0.0581 | -0.0558 | -0.0560 | -0.0531 | -0.0454 | -0.0425 |
| Wind | -0.2663 | -0.2620 | -0.2402 | -0.2346 | -0.2375 | -0.2293 | -0.2051 | -0.1966 |
| Day | -0.0019 | -0.0018 | -0.0017 | -0.0016 | -0.0015 | -0.0015 | -0.0013 | -0.0012 |
| Lag | 0.0390 | 0.0393 | 0.0407 | 0.0411 | 0.0412 | 0.0419 | 0.0438 | 0.0447 |
| Site | 1.0617 | 1.0448 | 0.9644 | 0.9420 | 0.9506 | 0.9188 | 0.8228 | 0.7879 |

*SD = saturation deficit, DD = average degree days over 60 days, Precip = total precipitation ten days prior to sampling, Wind = average wind speed over the last 30 days, Day = day length, Lag = number of active adults prior to sampling.

Kaizer, A. M., Foré, S. A., Kim, H. J., & York, E. C. (2015). Modeling the biotic and abiotic factors that describe the number of active off-host *Amblyomma americanum* larvae. *Journal of Vector Ecology*, 40(1), 1-10.