EPSE 596: Correlational Designs and Analysis

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

Return to the idea of "best model"

- We have seen that there are multiple ways one could define the "best estimates" for our unknown regression model coefficients, notably OLS or ML (which coincide for simple models).
- Regardless, once we have these "best" estimates for the unknown population parameters, we can go about assessing the adequacy of our proposed model.
- **Big Question:** Just how does one define an "adequate" vs. an "inadequate" model?
- We have already seen that part of the answer to this question has to be that the model is *valid*; i.e., no obvious violations of the model assumptions. Easiest and more reliable way to check this is via *residual diagnostics*.

Return to the idea of "best model"

- But realize that there may be more than one reasonably valid model. So how to choose between them?
- In fact, we already saw this in Week 9 (optional) when working through our "mediation" examples.
- In that case, both the simple model

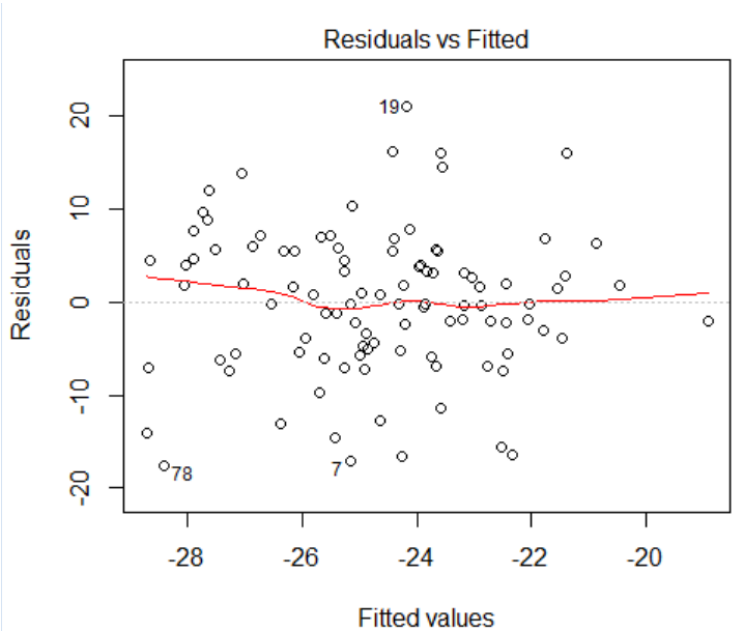
$$Y \sim N(\beta_0 + \beta_1 X_1, \sigma^2)$$

and the more complicated model

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 M_1 + \beta_3 X_2 + \beta_4 X_1 X_2, \sigma^2)$$

were valid models for the particular data under consideration.

Return to the idea of "best model"



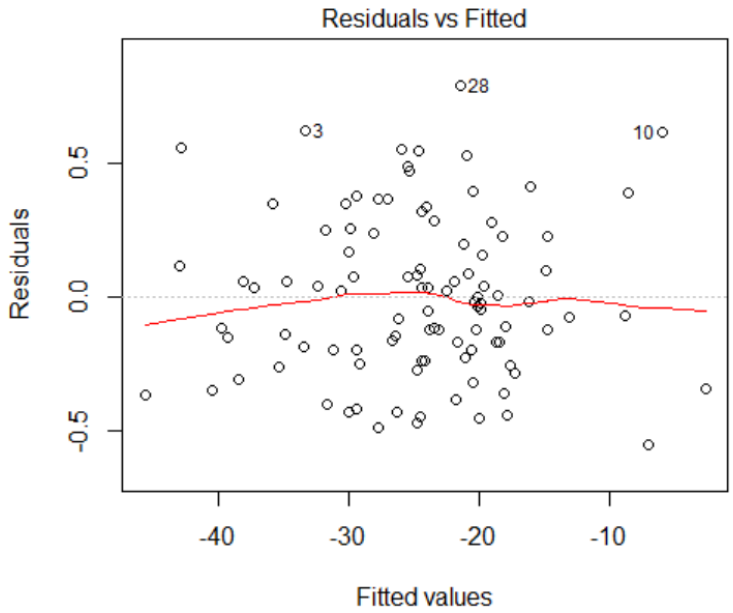
- Residual diagnostics for the simple model

$$Y \sim N(\beta_0 + \beta_1 X_1, \sigma^2)$$

look good.

- Notice the scale of this model's residuals.

Return to the idea of "best model"



- Residual diagnostics for the complicated model

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 M_1 + \beta_3 X_2 + \beta_4 X_1 X_2, \sigma^2)$$

look good.

- Notice the reduced scale of this model's residuals.

Return to the idea of "best model"

- We see a reduction (in this case, substantial) in unexplained/residual variance; so the complicated model seems to be a better one because it is simultaneously explaining more of the variation in the response and doing so validly.
- So if we have to choose between candidate (valid) models, should we always just choose the one with the smallest residuals, maybe quantified by the residual standard error $\hat{\sigma}$?
- For model 1, we have $\hat{\sigma} = 7.85$ and for model 2, we have $\hat{\sigma} = 0.31$.
- Why not make the model even more complicated though to try to explain more? E.g., consider model 3 specified by all two-way and three-way interactions between the explanatory variables:

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 M_1 + \beta_3 X_2 + \beta_4 X_1 X_2 + \beta_5 X_1 M_1 + \beta_6 M_1 X_2 + \beta_7 X_1 M_1 X_2, \sigma^2)$$

Model overfitting

- Intuitively, one might expect this model 3

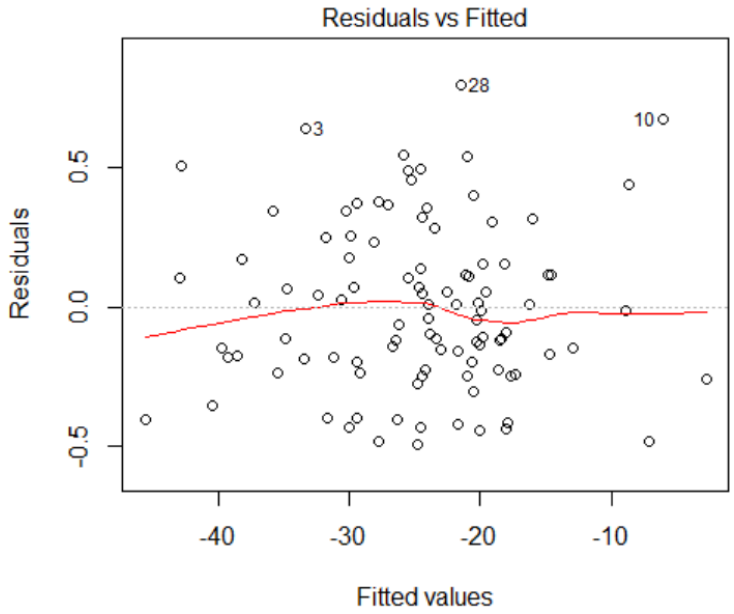
$$Y \sim N(\beta_0+\beta_1 X_1 + \beta_2 M_1 + \beta_3 X_2+ \\ \beta_4 X_1 X_2 + \beta_5 X_1 M_1 + \beta_6 M_1 X_2 + \beta_7 X_1 M_1 X_2, \sigma^2)$$

  to be even better than model 2 because it both:

  - *contains* model 2 as a special case (i.e., if/when $\beta_5 = \beta_6 = \beta_7 = 0$) and
  - allows for the possibility that we could explain more leftover variation in the response by the added terms in the model.

- However, in this case, we would be *overfitting* our model to the data; i.e., adding uninformative complications to the model. The model may still be *valid* in the sense of satisfying all assumptions, but it will no longer be optimally informative.

Model overfitting



- Residual diagnostics for the over-complicated model 3 still look good.
- Notice the scale of this model's residuals: basically the same as for model 2.
- However, no real *reduction* in the leftover/residual variation; in fact here, our estimate of σ doesn't really change: $\hat{\sigma} = 0.31$.

Model overfitting

- The more serious issue with overfitting becomes apparent when we actually examine the model estimates and their associated standard errors.
- First, examine the estimates for model 2:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.00996    0.86055   2.336   0.0216 *
x1           2.74015    0.21847  12.542   <2e-16 ***
m1           0.02545    0.06565   0.388   0.6992
x2           3.16355    0.21241  14.894   <2e-16 ***
x1:x2        0.95620    0.03284  29.121   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.306 on 95 degrees of freedom
```

Model overfitting

- Now examine the estimates for model 3:

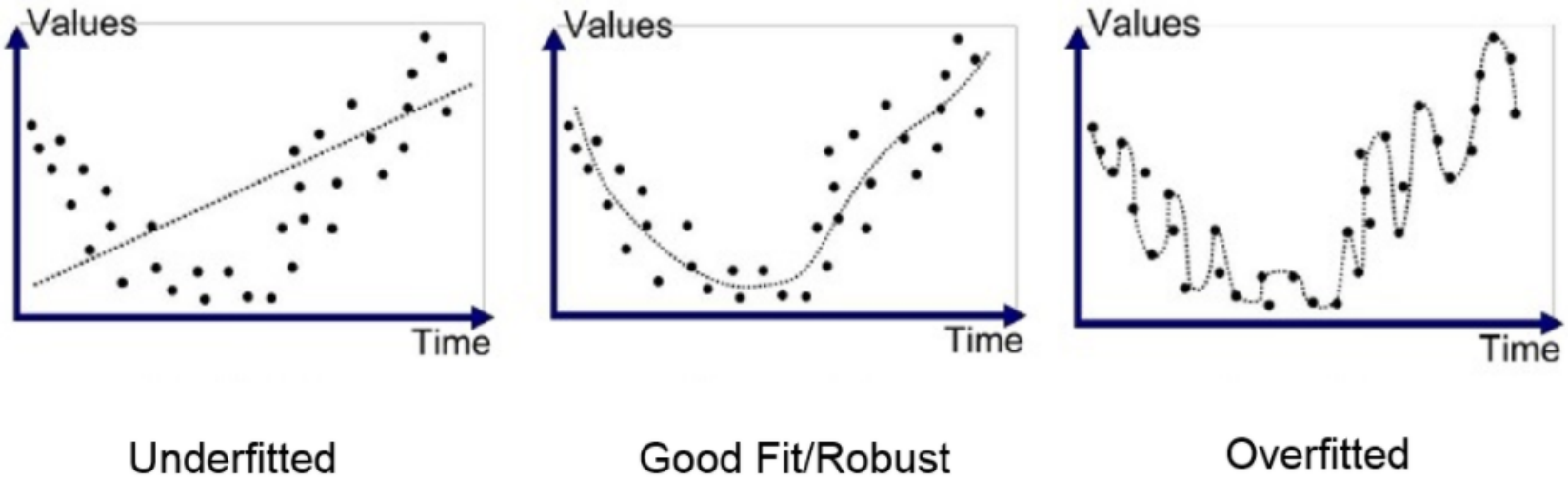
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.60839    1.13860   2.291   0.0243 *
x1           2.56552    0.27359   9.377 4.66e-15 ***
m1          -0.29832    0.28952  -1.030   0.3055
x2           3.22931    0.26561  12.158  < 2e-16 ***
x1:m1        0.06783    0.06176   1.098   0.2750
x1:x2        0.93272    0.04469  20.871  < 2e-16 ***
m1:x2       -0.05931    0.05763  -1.029   0.3061
x1:m1:x2     0.01180    0.01214   0.972   0.3337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.3081 on 92 degrees of freedom
```

- Notice the *inflated standard errors* on all coefficients from model 2.
- Inflation isn't too bad for this particular example, but in general, this is a bad thing: Inflated standard errors mean you lose the ability to estimate effects *precisely*. You may even lose the ability to tell they exist at all (loss of *power*).

Model building

- The classic picture is something like this:



- Finding the "best" model often requires finding this nice balance so that you are neither *underfitting* nor *overfitting* your model to the data.
- For our examples here, it has been easy to just look at the residuals and the reduction (or lack of reduction) in residual standard error to decide which model is "best." However, for more complicated modelling scenarios, some other tools may help.

Common tools to assess model "quality"

- There is a GIGANTIC collection of fit indices, information criteria, and other "quality" metrics that have been proposed and designed over the years to try to tell you if your model is doing an adequate job of explaining the data and/or if it is under/overfitting the data.
- There are so many different metrics, that many only exist in particular disciplines (e.g., psychology) or sub-disciplines (e.g. SEMs in psychology).
- In fact, in some applied disciplines, people seem to care more about their fit indices than about the actual validity of their models (similar to people obsessing over *reliability* of a test/scale and ignoring *validity*).
- I will talk through some of the most common ones you may encounter today, including: R^2 indices (saw these before), AIC/BIC.

Common tools to assess model "quality": R^2

There is a whole family of model fit indices called R^2 or r^2:

- Most classical is (multiple) R^2, the so-called *coefficient of determination*, defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2},$$

  where y_i are the observed responses, $\overline{y}$ is the ordinary average of these observed responses, and e_i are the model residuals (i.e., observed errors).

- When you fit a model with *no* predictors (i.e., intercept only model), then $R^2 = 0$.

- When your model perfectly matches your observations (i.e., no residual variation), then $R^2 = 1$.

Common tools to assess model "quality": R^2

- Classical interpretation: the closer R^2 is to 1, the better the model fit.
- In practice, *this is a useless criterion*.
- Quite common, with real data, to have decent models with small R^2 (say, around 0.2).
- Even more common to have *terrible, invalid* models with R^2 super close to 1.

Common tools to assess model "quality": R^2

- In particular, adding another term to your model with *automatically increase* R^2; so can artificially inflate your model "fit" by just making a stupid-complicated model. The "adjusted-R^2 index," R^2_{adj}, attempts to fix this problem.
- There are also many "pseudo-R^2 indices" that have been designed to quantify model fit in an analagous way for more not purely linear regression models (e.g., GLMs, nonparametric regressions).
- People have written papers for over 100 years about how terrible (or great) the R^2 family of statistics is; yet, it is still probably the most commonly reported statistic in a regression analysis, and R (and other software) happily spits it out for you regardless.
- **My advice:** Glance at the R^2_{adj}. If it is less than about 0.1, then your model is probably not explaining much of anything (there are exceptions to this though!). Otherwise, ignore all R^2 statistics.

More tools to assess model "quality": information criteria

There is a whole family of model "quality" tools based on *information criteria*:

- There are several rich and robust mathematical theories of *information* which play critical roles in computing, the study of dynamical systems, probability, and statistics.
- From out of this theory, comes a variety of statistical *information criteria* designed to simultaneously measure model fit and protect against overfitting.
- Most common varieties: AIC/AICc and BIC.
- These quantities can be *any real numbers*, positive or negative.
- Their scale is *meaningless* and somewhat arbitrary.
- The ICs are only useful as *relative* measures of model quality, e.g., is model 1 "better" than model 2? *Smaller* values (*not* in absolute value) indicate "better" models.

More tools to assess model "quality": information criteria

The AIC, Akaike's Information Criterion, is defined as

$$AIC = 2k - 2\ln(\widehat{L}),$$

where k is the number of estimated parameters in the model, and $\widehat{L}$ is the maximum value of the likelihood function generated by the model and the observed data.

- AIC can be any real number; smaller values (*not* in absolute value) indicate "better" model for the data.
- Critical to realize that AICs are only comparable between models that are fit *on the same dataset*.
- AICc is a simple adjustment that one often sees when sample sizes are small.
- AIC/AICc is ubiquitous in the ecological sciences, and very common in the social and health sciences.

More tools to assess model "quality": information criteria

The BIC, Bayesian Information Criterion, is defined as

$$BIC = k \ln(n) - 2 \ln(\widehat{L}),$$

where k is the number of estimated parameters in the model, n is the number of sample observations, and $\widehat{L}$ is the maximum value of the likelihood function generated by the model and the observed data.

- BIC can be any real number; smaller values (*not* in absolute value) indicate "better" model for the data.
- Critical to realize that BICs are only comparable between models that are fit *on the same dataset*.
- Called "Bayesian" because there is a very nice Bayesian interpretation of the BIC value. However, this interpretation is not necessary for implementation.
- BICs are super common in the health sciences, but also seen in the social and other sciences.

More tools to assess model "quality": information criteria

- Both the $AIC = 2k - 2\ln(\widehat{L})$ and the $BIC = k\ln(n) - 2\ln(\widehat{L})$ functionally operate the same way: smaller values (*not* in magnitude; i.e., the sign matters!) indicate "better" models.
- Notice both AIC and BIC penalize a model for having too many parameters (big k); i.e., they penalize overfitting. The BIC tends to penalize this more severely.
- Often, AIC and BIC will agree with which model is best when you compare different models; however, *they will not always agree*.
- In practice, look at both values. If both lead you to the same "best" model (among valid candidates), then great. If not, they are likely to lead you to quite similar models. If that doesn't happen, then you have probably really messed something up in your analysis or your theory.

Tools to assess model "quality"

Consider our different "mediation models" from Week 9. Model 1:

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.6214  -5.1843  -0.0976   5.4926  21.1616

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.3522     4.0644  -3.531 0.000632 ***
x1           -2.0154     0.7789  -2.588 0.011135 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.855 on 98 degrees of freedom
Multiple R-squared:  0.06395,   Adjusted R-squared:  0.0544
F-statistic: 6.695 on 1 and 98 DF,  p-value: 0.01114

> AIC(lm(y1~x1,data=dat))
[1] 699.9872
> BIC(lm(y1~x1,data=dat))
[1] 707.8027
```

- Note: The F-test that R spits out simply tests if there is evidence that the proposed model is explaining more variation in the response than just fitting an intercept-only model (i.e., no predictors). This test is virtually always significant in practice, and not practically useful for model selection.

Tools to assess model "quality"

Consider our different "mediation models" from Week 9. Model 2:

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.1343 -1.3483  0.0574  1.2062  4.3459

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.16136    1.38403   16.01   <2e-16 ***
x1          -9.28007    0.27001  -34.37   <2e-16 ***
m1           3.62217    0.09385   38.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.952 on 97 degrees of freedom
Multiple R-squared:  0.9428,    Adjusted R-squared:  0.9416
F-statistic: 798.9 on 2 and 97 DF,  p-value: < 2.2e-16

> AIC(lm(y1~x1+m1,data=dat))
[1] 422.5298
> BIC(lm(y1~x1+m1,data=dat))
[1] 432.9505
```

Tools to assess model "quality"

Consider our different "mediation models" from Week 9. Model 3:

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.6789 -0.3896 -0.0327  0.4190  2.1981

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1669     0.7176  36.464  < 2e-16 ***
x1           -2.2670     0.4224  -5.366 5.57e-07 ***
m1            0.1200     0.2055   0.584    0.561
x2            7.7697     0.4443  17.486  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9592 on 96 degrees of freedom
Multiple R-squared:  0.9863,    Adjusted R-squared:  0.9859
F-statistic:  2308 on 3 and 96 DF,  p-value: < 2.2e-16

> AIC(lm(y1~x1+m1+x2,data=dat))
[1] 281.3793
> BIC(lm(y1~x1+m1+x2,data=dat))
[1] 294.4052
```

Tools to assess model "quality"

Consider our different "mediation models" from Week 9. Model 4:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.55397 -0.20614 -0.02303  0.22690  0.79254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.00996    0.86055   2.336   0.0216 *
x1           2.74015    0.21847  12.542   <2e-16 ***
m1           0.02545    0.06565   0.388   0.6992
x2           3.16355    0.21241  14.894   <2e-16 ***
x1:x2        0.95620    0.03284  29.121   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.306 on 95 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9986
F-statistic: 1.722e+04 on 4 and 95 DF,  p-value: < 2.2e-16

> AIC(lm(y1~x1+m1+x2+x1:x2,data=dat))
[1] 53.85512
> BIC(lm(y1~x1+m1+x2+x1:x2,data=dat))
[1] 69.48614
```

Tools to assess model "quality"

Consider our overly complex model from today. Model 5:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.49449 -0.20463 -0.02657  0.17351  0.79965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.60839    1.13860   2.291   0.0243 *
x1           2.56552    0.27359   9.377 4.66e-15 ***
m1          -0.29832    0.28952  -1.030   0.3055
x2           3.22931    0.26561  12.158  < 2e-16 ***
x1:m1        0.06783    0.06176   1.098   0.2750
x1:x2        0.93272    0.04469  20.871  < 2e-16 ***
m1:x2       -0.05931    0.05763  -1.029   0.3061
x1:m1:x2     0.01180    0.01214   0.972   0.3337
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3081 on 92 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9985
F-statistic:  9708 on 7 and 92 DF,  p-value: < 2.2e-16

> AIC(lm(y1~x1*m1*x2,data=dat))
[1] 57.97577
> BIC(lm(y1~x1*m1*x2,data=dat))
[1] 81.4223
```

Notice that the AIC and BIC values both increased relative to Model 4, indicating *overfitting*.

Tools to assess model "quality"

My advice:

- Ignore all R^2 type statistics except for possible quick reality checks.
- AIC and/or BIC can be a mathematically coherent way of deciding on a "best" model (or a few "best" models), *but only after you have identified a candidate set of plausible and valid models first.*
- By mathematical definition, for real world data, there are *no* fit statistics of any kind that are going to tell you if a model is valid (i.e., satisifies the model assumptions) or not. You *must* perform residual diagnostics first and foremost.
- Naturally though, people love a recipe, so tend to just fit a few models and then pick the one with the lowest AIC (rampant in ecology). *This is terrible statistics and can lead to terrible science.* **You all know better than this!**

Automated model selection

In some domains, one has so much data (in the millions or more) and so many potential predictors of a phenomenon (hundreds or more), that our thoughtful approach to model building/selection is not feasible. E.g.,

- Anything that tracks or collects online (meta)-data.
- Lots of genetic work.
- Studying the interactions of neurons in the brain.
- Studying financial markets, real estate, etc.
- Speech and face recognition.

In these situations, *automated model selection* techniques are often employed, sometimes with great success. Most notably, researchers, adminstrators, granting agencies, and the public alike all "ooh" and "aah" over techniques in *machine learning*, *artificial intelligence*, and *big data*.

Automated model selection

My biases are now apparent, but there is nothing wrong with ML/AI techniques *when properly employed in the proper domains*. However:

- Automated model selection is not new. Early methods date back about 100 years.
- Automated model selection is the ultimate in unthinking recipes, so very attractive to researchers who do not want to take the time (or learn how) to thoughtfully construct reasonable statistical models.
- Automated model selection completely disguises what is functionally happening during model building/selection.
- ML/AI methods only work when one has a *ton* of data.
- For most applied research domains in the social, health, or ecological sciences, ML/AI techniques are inappropriate due to too small sample sizes and too robust theoretical and/or previously substantiated domain knowledge.

Automated model selection

The oldest form of automated model selection is forward/backward/stepwise model selection on p-values.

- For stepwise selection, one often starts adding predictors into the model and checking that each new predictor's estimated effect is statistically distinguishable from zero.
- If a predictor's effect becomes zero after adding a new term, remove that predictor.
- Continue ad nauseam.

There are a plethora of well established problems with this practice, including:

- It will lead you to different models depending on where you start.
- It mathematically biases all your effect estimates.
- It ignores model validity.

Automated model selection

- There are many variations on stepwise model selection, all of them bad.
- Thankfully, this technique is universally panned by statisticians nowadays, yet you still see it survive in a few applied domains.
- The ideas behind stepwise model building/selection have been adapted to some ML/AI problems (sometimes with success).
- **Do not perform any type of automated model selection** unless you are in a truly "big data" type situation *and* are working with a competent statistician.

Cross validation

- Another very popular type of model selection / fit assessment tool is to perform *cross validation*.
- This technique can work well in both small and big data situations.
- Basic idea:
  - Randomly select, say, 80% of your observed data (these are called the "training data"); fit your proposed model on these data.
  - Use the resulting fitted model to see how well it predicts the remaining 20% of your observed data (e.g., do about 95% of these hold-out data fall inside the 95% prediction bands for your fitted regression model?).
  - Compare different models by comparing how well they predict the "hold-out" or "validation" or "test" data.
  - This process is often iterated many times (K-fold cross validation) to create an average "best" fit.

Cross validation

- Many ways to measure how well a model "predicts" a new observation (remember our discussion of *prediction error/uncertainty* in Week 5).
- Most direct way is to compute a *mean squared prediction error* (MSPE):
  - Compute all the individual prediction errors $p_i = y_i - \hat{y}_i$ for all i indexing the "hold-out" validation dataset.
  - Square all these values and take the sample average.
- Then the model that yields the smallest MSPE is in this sense "best".
- Many other ways to quantify such *prediction error* of a model, just as there are many different ways to define a *best estimator* for an unknown model parameter.

Cross validation

- One could easily design an entire course that only examines different kinds of model fit/selection indices, criteria, and techniques.
- Easy to get overwhelmed with the options that are out there.
- And it is easy to get confused as different disciplines have a tendency to prefer their own indices, criteria, techniques.
- Regardless, what matters is the following:
  - Always use your domain knowledge when building models.
  - Outside of truly "big data" situations, avoid all automated model building/selection techniques.
  - Always check model assumptions (validity) by performing residual diagnostics.
  - With some possible valid models in hand, pick the best one by examining estimates of residual standard error, AICs or BICs, and/or prediction error. *Realize no one model may be best*.