Model Selection and Overfitting

Given a variable Y that we want to predict (linear or logistic regressions)
and a set of predictors $X_1, X_2, X_3, \ldots, X_k$

We ask the question: Which predictors should we use in our model?

Why don't we just use everything?

Thought: If a variable X_i is not a good predictor of Y, then the coefficient on that variable should be zero.

If this is true, p-value will be large, and we can remove it from the model.

Note that every p-value in the model is contextual to all other variables in the model.

It might be better to start from nothing and add variables only if they are significant.

Forward step selection but we can motivate with more than the p-values.

One metric we might consider is <u>likelihood</u>

We denote likelihood with

$$L(\text{Model}) = \text{likelihood}$$

Recall dnorm gives the likelihood of a value from normal dist.
likelihood is height of this function

$L(\text{Model}) = \text{likelihood}$

Likelihood only increases as more predictors are added to the model.

What we can do is rate a model based on likelihood, but penalize it for each added term.

2 options

Akaike information criterion (AIC)

$$AIC(\text{Model}) = -2\log(L(\text{Model})) + 2 \times \underbrace{k}_{\#\text{ of parameters in the model}}$$

Bayes information criterion (BIC)

$$BIC(\text{Model}) = -2\log(L(\text{Model})) + \log(n)\cdot k$$

These should be as small as possible, with the penalty term raising the value as too many terms are added to the model.

We use these to prevent <u>overfitting</u>.

<u>Overfitting</u> is a problem where too many predictors are used producing a model that <u>very accurately</u> describes the sample used to produce it, but cannot be generalized to new samples.

Model selection must be done in a way that avoids <u>overfitting</u>.