Introduction to model selection

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

27 April 2020

Goals for today

- Understand the the concept of bias-variance trade-offs
- Understand the use of null hypothesis tests for “in-sample” model selection
- Understand the use of an information criterion for model selection

Approximating the truth

In general, our goal is to approximate a true model $f(x)$ with an estimate $\hat{f}(x)$

In doing so, we estimate the model parameters from the data

Variability among models

Imagine we could repeat our model building process by gathering new data and fitting new models

Due to randomness in the underlying data sets, each of our models will have a range of predictions and associated errors

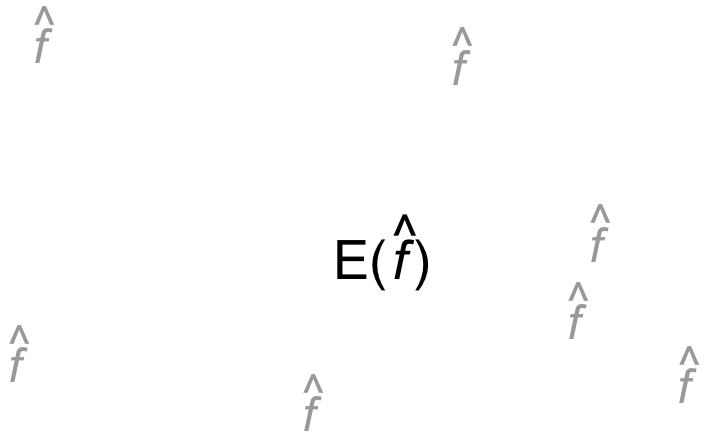
Model errors

The model errors arise from 2 sources:

1. Variance

How much do the predictions $\hat{f}(x)$ vary among our different models?

Variance in models



$$\mathrm{Var}(\hat{f}) = \frac{\sum(\hat{f}_i - \mathrm{E}(\hat{f}))^2}{n-1}$$

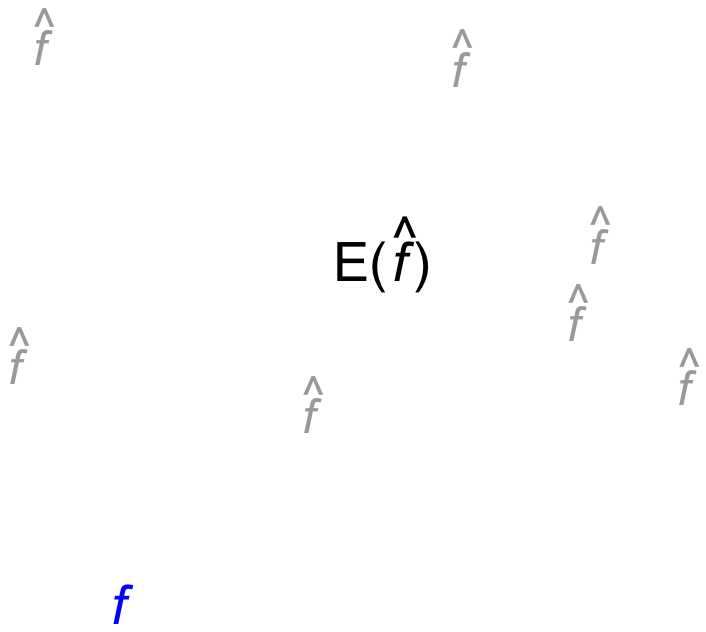
Model errors

The model errors arise from 2 sources:

1. Variance
2. Bias

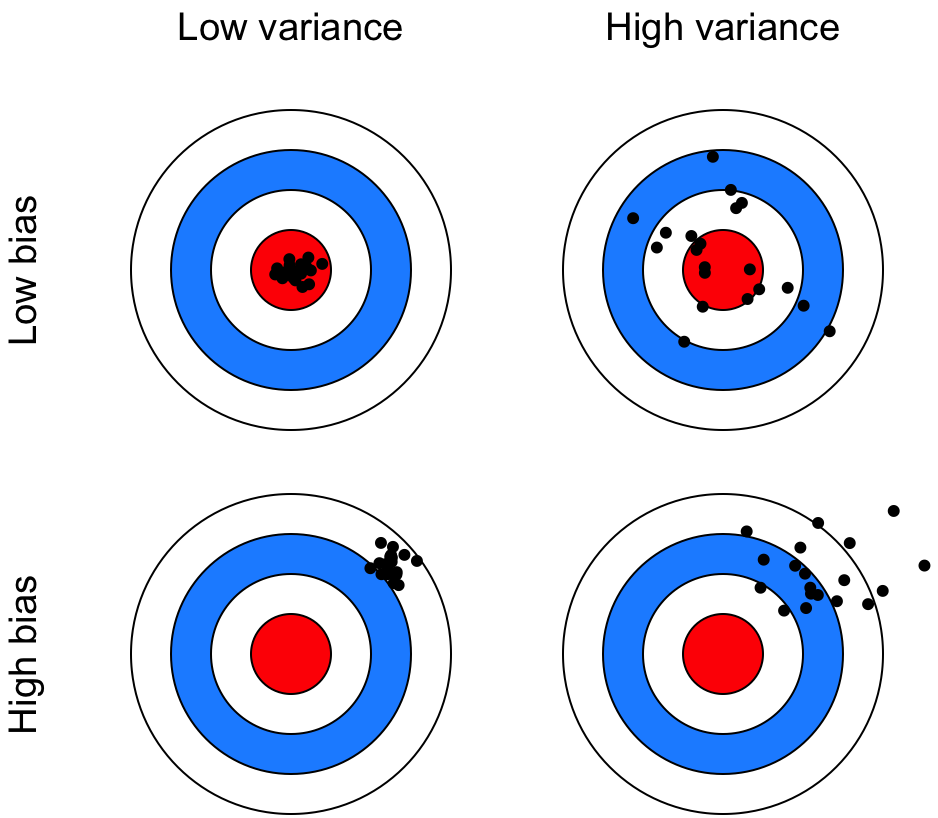
How close is the *expected* prediction of our model to the *true* value $f(x)$?

Bias in models

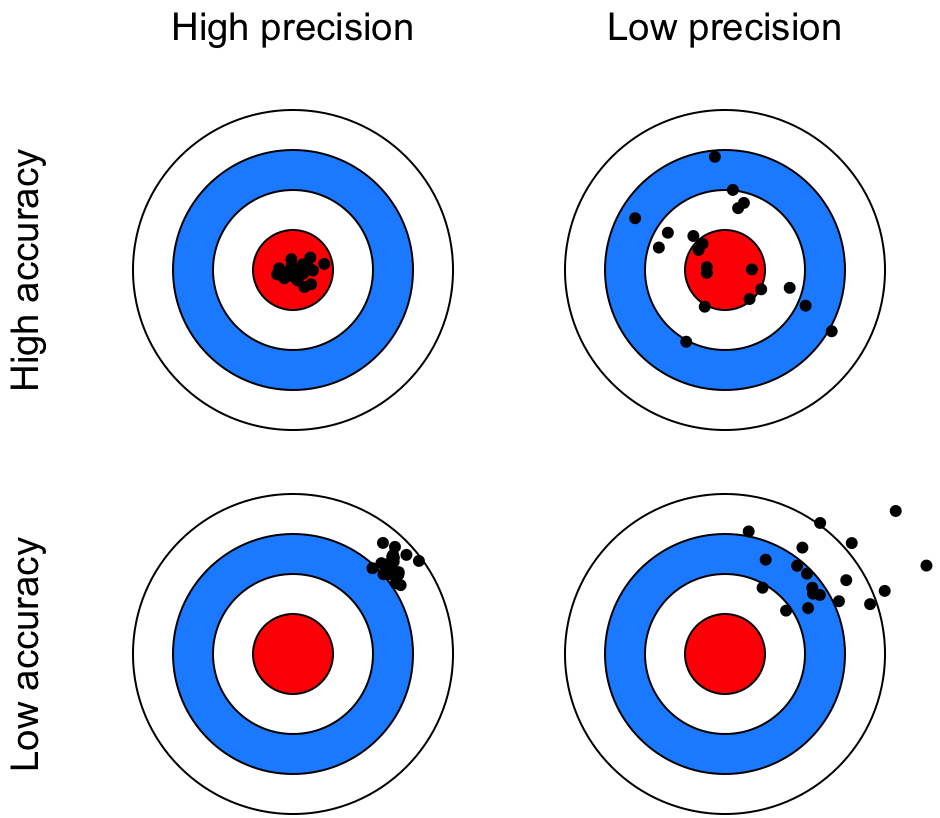


$$\mathrm{Bias}(\hat{f}) = f - \mathrm{E}(\hat{f})$$

Bias versus variance



Accuracy versus precision



Sum of squared errors

Recall that the squared difference between our model predictions and the observed values is the *sum of squared errors* (SSE)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}\mathbf{X}_i)^2$$

Mean squared error

Recall also that the *expectation* of the SSE is the *mean squared error*, which gives us an estimate of the variance in the e_i

$$MSE = \frac{SSE}{n-k} = \hat{\sigma}^2$$

Mean squared error

We can decompose the MSE into its bias and variance pieces

$$MSE = \text{Bias}^2 + \text{Var}(\hat{f}) + \sigma^2$$

Here σ^2 is the *irreducible error*

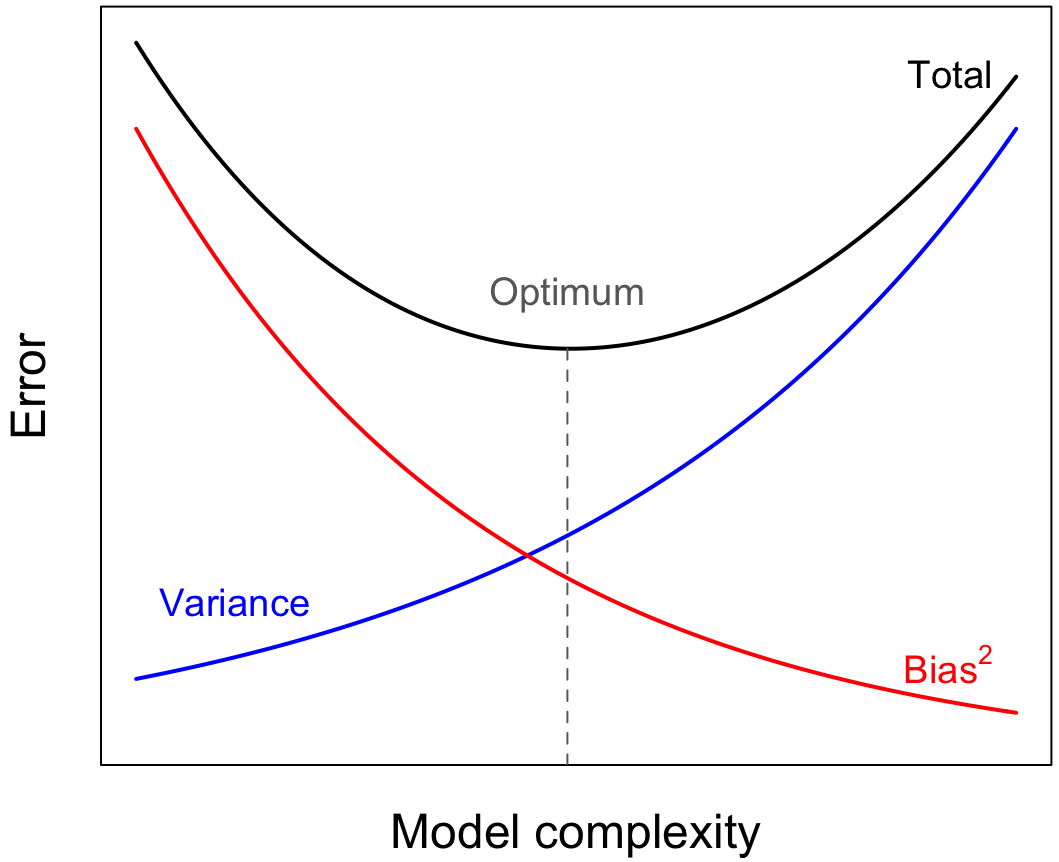
(See here for a full derivation)

Model errors

There is a trade-off between a model's ability to simultaneously minimize both bias and variance

- bias decreases as model complexity increases
- variance increases as model complexity increases

Bias-variance trade-off



Model complexity

How do we choose the right level of model complexity?

We want to *include* predictors x that

- have a strong relationship with y
- offer new info about y given other predictors

Model complexity

How do we choose the right level of model complexity?

We want to *exclude* predictors x that

- don't have a strong relationship with y
- offer the same info about y as other predictors (collinearity)

Model selection

Selecting among possible models begins with a *reasonable* set based on first principles

Model selection

Selecting among possible models begins with a *reasonable* set based on first principles

This set of models

- may represent different hypotheses about the process of interest
- may represent different combinations of predictors

Model selection

There are 2 general approaches to model selection:

1. In-sample

Uses the same information to fit and evaluate the model

Model selection

There are 2 general approaches to model selection:

1. In-sample
2. Out-of-sample

Uses different information to fit and evaluate the model

In-sample model selection

There are 2 general ways to do in-sample selection:

1. Null hypothesis testing
2. Regularization

Null hypothesis tests

We have already seen a variety of F-tests to test different models

For example, given this full model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i$$

we might test

$H_0 : \beta_1 = 0$ or $H_0 : \beta_2 = c$

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Null hypothesis tests

There are several methods for testing models in a *stepwise manner*:

- Forward

Sequentially add predictors to a model based on their p-value & re-test the model

Null hypothesis tests

There are several methods for testing models in a *stepwise manner*:

- Forward
- Backwards

Fit a model with *all* of the predictors, remove the one with the largest p-value & re-test the model

Null hypothesis tests

There are several issues with stepwise selection:

- The final model is chosen as if there was no uncertainty about it
- The one-at-a-time nature of adding/dropping predictors can miss the optimal model
- **Lots** of null hypothesis tests & choices about α
- No underlying theoretical basis for the approach
- These were once standard practice, but are rarely seen now

Likelihood ratio test

For *nested* models, we can make use of a *likelihood ratio test*, which compares the goodness of fit between 2 models

Likelihood ratio (LR)

Given the likelihoods from a full model $\mathcal{L}(y; \Theta)$ and a reduced model with fewer parameters $\mathcal{L}(y; \theta)$,

$$LR = \frac{\mathcal{L}(y; \theta)}{\mathcal{L}(y; \Theta)}$$

Likelihood ratio (LR)

Given the likelihoods from a full model $\mathcal{L}(y;\Theta)$ and a reduced model with fewer parameters $\mathcal{L}(y;\theta)$,

$$LR = \frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}$$

Because $\mathcal{L}(y;\theta) < \mathcal{L}(y;\Theta)$, this ratio will vary from

0 (data unlikely to have come from the reduced model) to
1 (data equally likely to have come from either model)

Likelihood ratio test

More formally, given the likelihoods from a full model $\mathcal{L}(y; \Theta)$ and reduced model $\mathcal{L}(y; \theta)$, the test statistic λ is given by

$$\begin{aligned} \lambda &= -2 \log(LR) \\ &= -2 \log\left(\frac{\mathcal{L}(y; \theta)}{\mathcal{L}(y; \Theta)} \right) \end{aligned}$$

Likelihood ratio test

The test statistic λ follows a Chi-squared distribution

$$\lambda \sim \chi^2_{(k_\Theta - k_\theta)}$$

where $df = k_\Theta - k_\theta$ is the difference in the number of parameters between the 2 models

Likelihood ratio test

Alternatively, the test can be expressed in terms of log-likelihoods

$$\begin{aligned}\lambda &= -2\log\left(\frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}\right) \\ &= -2\left[\log\mathcal{L}(y;\theta) - \log\mathcal{L}(y;\Theta)\right]\end{aligned}$$

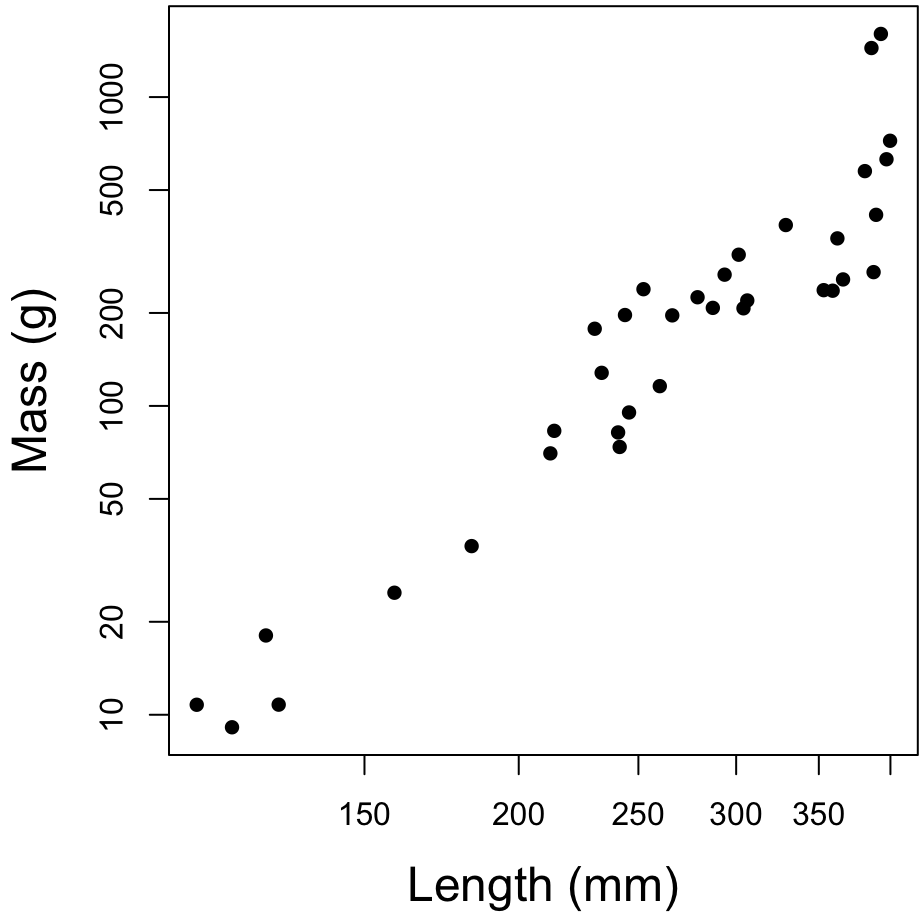
Likelihood ratio test

Our null hypothesis for this test is that the data were just as likely to have come from the reduced model

$$H_0 : y = f_\theta(x)$$
$$H_A : y = f_\Theta(x)$$

and we reject H_0 if $\lambda > \chi^2_{df}$

Linear models for size of fish



Linear models for size of fish

Two simple choices:

1. $\log_{10}(mass_i) = \alpha + e_i$
2. $\log_{10}(mass_i) = \alpha + \beta \log_{10}(length_i) + e_i$

Linear models for size of fish

Let's make some simple substitutions

$y_i = \log_{10}(mass_i)$ and $x_i = \log_{10}(length_i)$

so that

1. $y_i = \alpha + e_i$
2. $y_i = \alpha + \beta\, x_i + e_i$

Fit reduced model

```
## fit reduced model
m1 <- lm(L10_mass ~ 1)
faraway::sumary(m1)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 2.185557   0.094317  23.172 < 2.2e-16
##
## n = 35, p = 1, Residual SE = 0.55799, R-Squared = 0
```

```
## log-likelihood
(LL_1 <- logLik(m1))
```

```
## 'log Lik.' -28.73589 (df=2)
```

Fit full model

```
## fit full model
m2 <- lm(L10_mass ~ L10_length)
faraway::sumary(m2)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -5.77317    0.46626 -12.382 5.966e-14
## L10_length   3.29059    0.19237  17.106 < 2.2e-16
##
## n = 35, p = 2, Residual SE = 0.18031, R-Squared = 0.9
```

```
## log-likelihood
(LL_2 <- logLik(m2))
```

```
## 'log Lik.' 11.32497 (df=3)
```

Likelihood ratio test

$$\lambda = -2 \left[\log \mathcal{L}(y; \theta) - \log \mathcal{L}(y; \Theta) \right]$$

```
## test statistic
lambda <- as.numeric(-2 * (LL_1 - LL_2))
## degrees of freedom (ignoring sigma for both models)
df <- length(coef(m2)) - length(coef(m1))
## p-value
pchisq(lambda, df, lower.tail = FALSE)
```

```
## [1] 3.520383e-19
```

The p-value is *very* small so we reject H_0 and conclude that the data were unlikely to have come from the simple model

QUESTIONS?

Model selection for non-nested models

The likelihood ratio test only works for nested models

What can do we if model A is not nested within B?

Kullback-Leibler divergence

One can characterize the “distance” between 2 distributions (models) with the *Kullback-Leibler divergence*

Likelihood and K-L divergence

Let's return to our likelihood ratio

$$LR = \frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}$$

Likelihood and K-L divergence

For a set of data with independent samples $\{y_1, y_2, \ldots, y_n\}$, we can compute the likelihood ratio for all of the y_i by taking the product of the likelihood ratio over all i

$$LR = \prod_{i=1}^{n} \frac{\mathcal{L}(y_i; \theta)}{\mathcal{L}(y_i; \Theta)}$$

Likelihood and K-L divergence

As we saw for the likelihood, we can take the log of LR and work with a sum instead

$$LR = \prod_{i=1}^{n} \left(\frac{\mathcal{L}(y_i; \theta)}{\mathcal{L}(y_i; \Theta)} \right)$$

$$\Downarrow$$

$$\log LR = \sum_{i=1}^{n} \log \left(\frac{\mathcal{L}(y_i; \theta)}{\mathcal{L}(y_i; \Theta)} \right)$$

Likelihood and K-L divergence

We can normalize $\log LR$ for different sampling effort by dividing by n

$$\widehat{\log LR} = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\mathcal{L}(y_i; \theta)}{\mathcal{L}(y_i; \Theta)} \right)$$

Likelihood and K-L divergence

Let's now imagine we collect an infinite number of samples (we're going to be busy!) and consider what happens to the $\log LR$

$$\lim_{n\to\infty} \widehat{\log LR} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{\mathcal{L}(y_i;\theta)}{\mathcal{L}(y_i;\Theta)} \right)$$
$$= \mathrm{E}\left[\log\left(\frac{\mathcal{L}(y_i;\theta)}{\mathcal{L}(y_i;\Theta)} \right) \right]$$

Likelihood and K-L divergence

This expectation is known as the *Kullback-Leibler divergence*

$$D_{KL} = \mathrm{E}\left[\log\left(\frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}\right)\right]$$

K-L divergence

We can further decompose the K-L divergence as

$$
\begin{aligned}
D_{KL} &= \mathrm{E}\left[\log\left(\frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}\right)\right] \\
&= \mathrm{E}\left(\log \mathcal{L}(y;\theta) - \log \mathcal{L}(y;\Theta)\right) \\
&= \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\theta)\right)}_{\text{entropy}} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}}
\end{aligned}
$$

K-L divergence

We can further decompose the K-L divergence as

$$
\begin{aligned}
D_{KL} &= \mathrm{E}\left[\log\left(\frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}\right)\right] \\
&= \mathrm{E}\left(\log \mathcal{L}(y;\theta) - \log \mathcal{L}(y;\Theta)\right) \\
&= \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\theta)\right)}_{\text{entropy}} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}} \\
&= \text{constant} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}}
\end{aligned}
$$

An information criterion

In the early 1970s, Hirotugu Akaike figured out a connection between maximum likelihood and K-L divergence

- Imagine our data came from some *unknown* model f
- We have 2 candidate models, g_1 and g_2, for approximating f
- If we knew f, we could use the K-L divergence to measure the *information lost* when using g_1 and g_2 to approximate f
- Unfortunately, we do not know f, but Akaike found a way around this problem

An information criterion

Akaike showed that we can use *an information criterion* (AIC) based on the K-L divergence to measure the *relative* information lost when using g_1 versus g_2

$$D_{KL} = \text{constant} - \underbrace{\text{E}\left(\log \mathcal{L}(y; \Theta)\right)}_{\text{log likelihood}}$$

$$\Downarrow$$

$$AIC \approx 2D_{KL}$$

Akaike's information criterion

Specifically, *Akaike's information criterion* for a given model is

$$AIC = 2k - 2\log \mathcal{L}(y;\theta)$$

where k is the number of parameters in the model

Akaike's information criterion

Given a set of candidate models, the preferred model has the lowest AIC

$$AIC = 2k - 2\log \mathcal{L}(y;\theta)$$

- AIC rewards goodness of fit (as measured by the likelihood)
- AIC penalizes over-fitting (as measured by the number of parameters)
- Thus, AIC helps us prevent over-fitting by addressing the bias-variance trade-off

Bias in AIC

When the sample size n is small, AIC tends to select models that have too many parameters

To address this potential for over-fitting, a *corrected* form of AIC was developed

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

The additional penalty term goes to 0 as $n \to \infty$

Linear models for size of fish

Two simple choices:

1. $\log_{10}(mass_i) = \alpha + e_i$
2. $\log_{10}(mass_i) = \alpha + \beta \log_{10}(length_i) + e_i$

Fit the models in R

```
## fit intercept-only model
m1 <- lm(L10_mass ~ 1)
## fit intercept + slope model
m2 <- lm(L10_mass ~ L10_length)
## calculate AIC's
AIC(m1, m2)
```

```
##    df       AIC
## m1  2  61.47179
## m2  3 -16.64995
```

Model 2 has the lowest AIC and is therefore the most parsimonious

Bias-corrected AICc

```
## function for AICc
AICc <- function(AIC, n, k) {
  AIC + (2 * k^2 + k) / (n - k - 1)
}
## sample size
n <- 35
## number of parameters = intercept (+ slope) + sigma = 2 (3)
k1 <- 2; k2 <- 3
## AICc for model 1
AICc(AIC(m1), n, k1)
```

```
## [1] 61.78429
```

```
## AICc for model 2
AICc(AIC(m2), n, k2)
```

```
## [1] -15.97253
```

Model 2 still has the lowest AIC and is therefore the most parsimonious

Summary

We have seen 2 general approaches approaches to in-sample model selection

- F-tests and Likelihood-ratio tests for nested models
- AIC for both nested and non-nested models
- Only the latter helps us address the bias-variance trade-off