From Model to Log Likelihood

Stephen Pettigrew

February 19, 2014

Outline

Outline

Big Picture

The class is all about how to do good research

Last week we talked about three theories of inference:
likelihood inference, Bayesian inference, and Neyman-Pearson hypothesis testing

For most of the remainder of the course, we'll be going down the rabbit hole of likelihood inference

Whether they acknowledge it or not, most of the quantitative research you've read in the past relied on likelihood inference, and specifically maximum likelihood estimation

If it's not using OLS and doesn't have a prior, it's probably MLE

Goals for tonight

By the end of tonight, you should feel comfortable with these three things:

1. Understand the IID assumption and why it's so important
2. Take a model specification (stochastic/systematic components) and turn it into a likelihood
3. Turn a likelihood function into a log likelihood and understand why logs should make us happy

Outline

Running Example

In the spirit of the Olympics, we're going to use data from curling at Sochi.

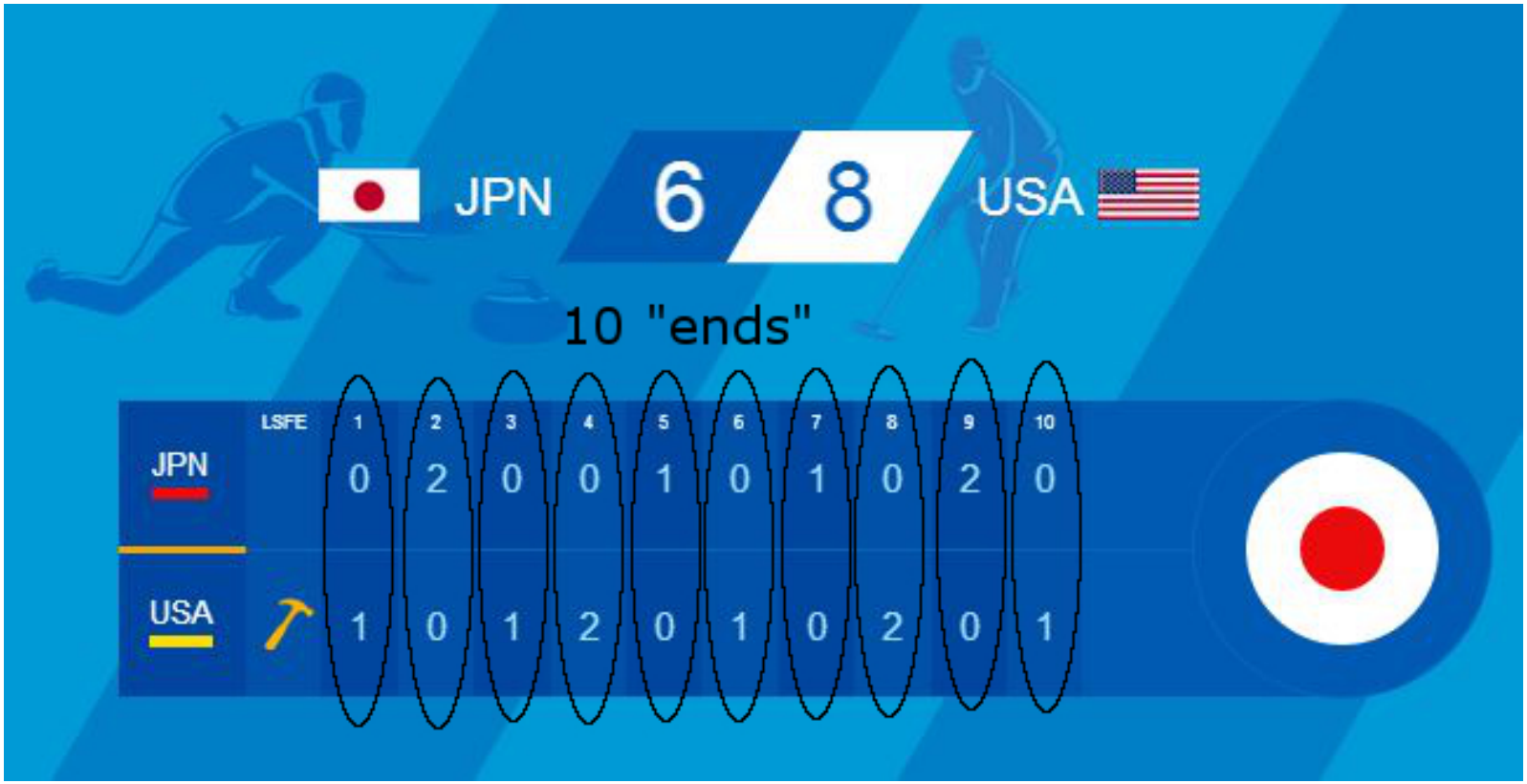
Specifically, we're going to model the number of points scored by the winning team in each of the ten "ends" of a curling match

Running Example

The ten data points for this game:

`c(1,0,1,2,0,1,0,2,0,1)`

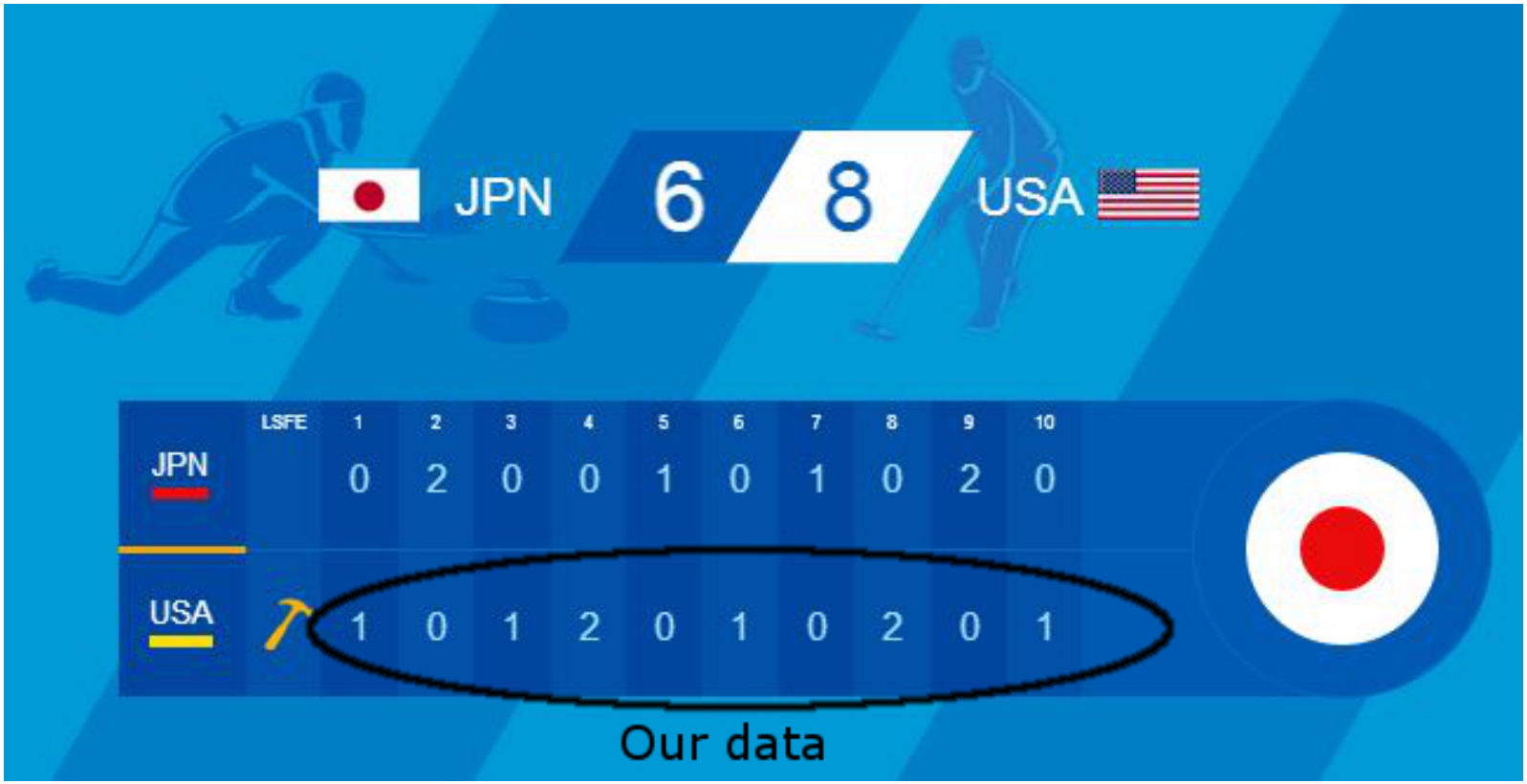
Running Example



The ten data points for this game:

`c(1,0,1,2,0,1,0,2,0,1)`

Running Example



The ten data points for this game:

`c(1,0,1,2,0,1,0,2,0,1)`

Writing out our model

Suppose we want to model these data as our outcome variable.

We'd first need to write out our model by writing out the stochastic and systematic components.

For the stochastic component, let's assume these data follow a Poisson distribution.

For the systematic component, let's keep things simple and leave out any explanatory variables. We'll only model an intercept term.

Writing out our model

Using the notation from UPM, this model could be written as:

$$y_i \sim Pois(\lambda_i)$$
$$\lambda_i = e^{\beta_0}$$

e^{β_0} because $\lambda > 0$ in the Poisson distribution.

If we had other covariates (such as team dummy variables) they would enter our model through the systematic component:

$$y_i \sim Pois(\lambda_i)$$
$$\lambda_i = e^{\beta_0 + \beta_1 x_i} = e^{X_i \beta}$$

For now though, we'll keep things easy and use the first model with an intercept with no covariates.

Outline

Probability statement with one data point

Imagine that we only had one datapoint (like last week's section) and that we knew the true value of λ_i

With those two pieces of information, we'd be able to calculate the probability of observing that datapoint given λ_i:

$$p(y_i|\lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

$$p(y_i = 1|\lambda_i = .86) = \frac{.86^1 e^{-.86}}{1!}$$

$$p(y_i = 1|\lambda_i = .86) = 0.3639$$

What if we have more than one observation?

Imagine that we only had two datapoints (1 and 0) and we still knew that $\lambda = .86$

What do we typically assume about those two datapoints when we do likelihood inference?

Independent and identically distributed (IID)

Identically distributed: We assume that the observations are drawn from the same distribution, conditional on the covariates.

Independence

Independent: Two events are independent if the occurrence (or non-occurrence) of one does not affect the probability of another.

More formally, A and B are independent if and only if $P(A \text{ and } B) = P(A)P(B)$

Why is this important? Because it makes it easy for us to write out our probability (or likelihood) function

Probability statement with two data points

If we know that $y_1 = 1$, $y_2 = 0$, and $\lambda_1 = \lambda_2 = .86$ then we can calculate the probability of observing both data points:

$$p(\mathbf{Y}|\lambda_i) = p(y_1 = 1|\lambda_i) \cdot p(y_2 = 0|\lambda_i)$$

$$p(\mathbf{Y}|\lambda_i) = \frac{\lambda_1^{y_1} e^{-\lambda_1}}{y_1!} \cdot \frac{\lambda_2^{y_2} e^{-\lambda_2}}{y_2!}$$

$$p(\mathbf{Y}|\lambda_i) = \prod_{i=1}^{2} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

$$p(\mathbf{Y}|\lambda_i = .86) = 0.3639 \cdot 0.4231$$

$$p(\mathbf{Y}|\lambda_i = .86) = 0.1540$$

Outline

Writing a likelihood function

What if we don't know the value of λ?

If we did, it probably wouldn't make for very interesting research.

Put another way, if we knew the value of our parameters, then we could analytically calculate $p(\mathbf{Y}|\lambda)$

But we don't know λ, so what we're really interested in is $p(\lambda|\mathbf{Y})$

And recall from lecture that:

$$P(\lambda|\mathbf{Y}) = \frac{P(\lambda)P(\mathbf{Y}|\lambda)}{\int P(\lambda)P(\mathbf{Y}|\lambda)d\lambda}$$

Writing a likelihood function

$$P(\lambda|\mathbf{Y}) = \frac{P(\lambda)P(\mathbf{Y}|\lambda)}{\int P(\lambda)P(\mathbf{Y}|\lambda)d\lambda}$$

By the likelihood axiom:

$$L(\lambda|\mathbf{Y}) \equiv k(\mathbf{Y})P(\mathbf{Y}|\lambda)$$

$$L(\lambda|\mathbf{Y}) \propto P(\mathbf{Y}|\lambda)$$

In a likelihood function, the data is fixed and without variance and the parameters are variables

Proportionality

Throughout the semester you're going to become intimately familiar with the proportionality symbol, $\propto$.

Why are we allowed to use it?
Why can we get rid of constants (like $k(\mathbf{Y})$) from our likelihood function?
Wouldn't our estimates of $\hat{\lambda}_{MLE}$ be different if kept the constants in?

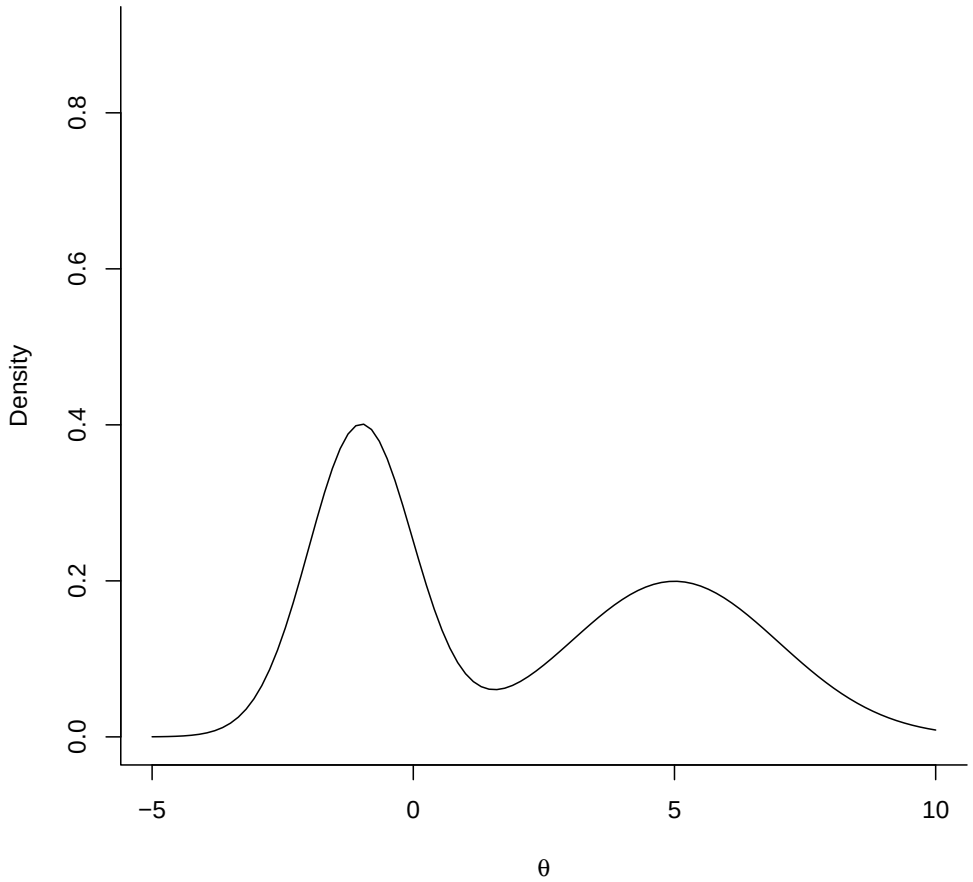
No.

Remember again from 8th grade when you learned about transformations of variables.

Multiplying a function by a constant vertically shrinks or expands the function, but it doesn't change the horizontal location of peaks and valleys.

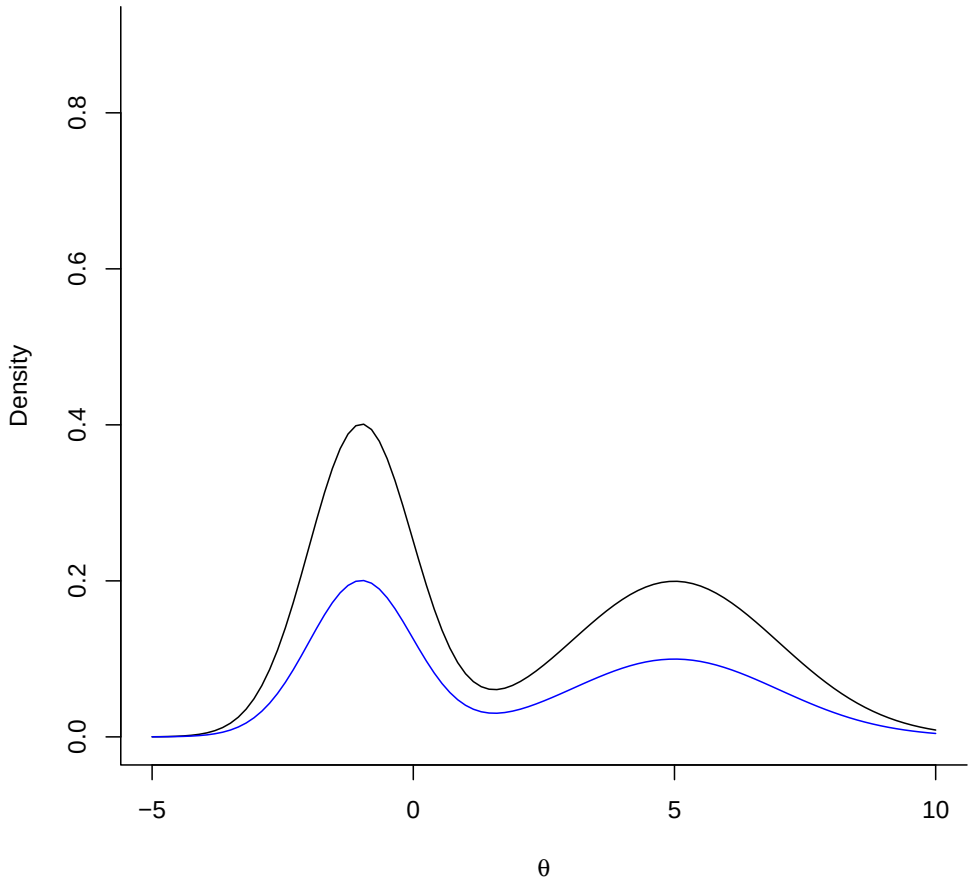
Proportionality

$$k(\mathbf{Y}) = 1$$



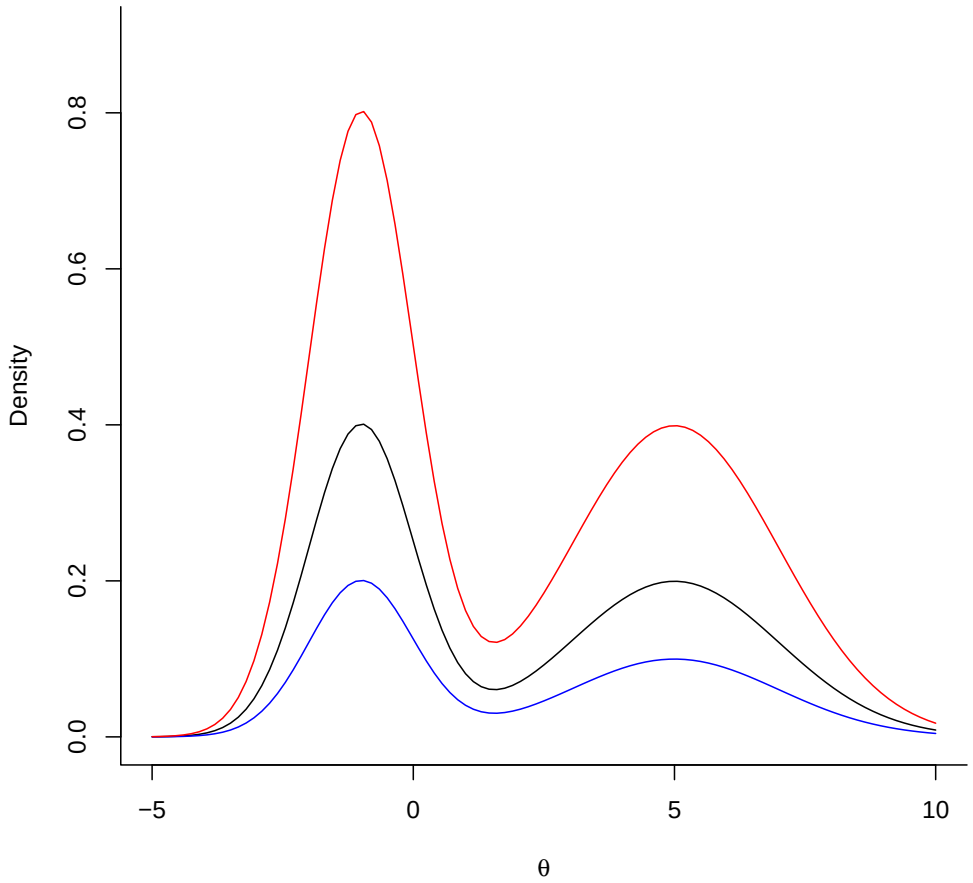
Proportionality

$$k(\mathbf{Y}) = .5$$



Proportionality

$$k(\mathbf{Y}) = 2$$



Proportionality

The likelihood function is not a probability density!

But it is *proportional* to the probability density, so the value that maximizes the likelihood function also maximizes the probability density.

Writing a likelihood function

$$L(\lambda_i|\mathbf{Y}) \propto P(\mathbf{Y}|\lambda_i)$$

$$L(\lambda_i|\mathbf{Y}) \propto \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

Get rid of any constants, i.e. any terms that don't depend on the parameters:

$$L(\lambda_i|\mathbf{Y}) \propto \prod_{i=1}^{n} \lambda_i^{y_i} e^{-\lambda_i}$$

Outline

Why do we use log-likelihoods?

When we begin to optimize likelihood functions, we *always* take the log of the likelihood function first. Why?

1. Computers don't handle very small decimals very well. If you're multiplying the probability of 10,000 datapoints together, you're going to get a decimal number that's incredibly tiny. Your computer could have an almost impossible time finding the maximum of such a flat surface.
2. The standard errors of maximum likelihood estimates are a function of the log-likelihood, not the likelihood.
   If you use R to optimize a likelihood function without taking the log, you are guaranteed to get the wrong standard error estimates.

8^{th} Grade Algebra Review: Logarithms

The log function can be thought of as an inverse for exponential functions.

$$y = \log_a(x) \iff a^y = x$$

In this class we'll always use the natural logarithm, $\log_e(x)$. Typically we'll just write $log(x)$, but you should always assume that the base is Euler's number, $e = 2.718281828$.

8^{th} Grade Algebra Review: Properties of logs

Five rules to remember:

1. $\log(xy) = \log(x) + \log(y)$
2. $\log(x^y) = y \log(x)$
3. $\log(x/y) = \log(x \cdot y^{-1}) = \log(x) + \log(y^{-1}) = \log(x) - log(y)$
4. Base change formula:

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$$

5. $\frac{\partial \log(x)}{\partial x} = \frac{1}{x}$
   Okay, that's probably not something you learned in 8th grade...

Writing out the log-likelihood

Recall that for a variable distributed Poisson(λ), the likelihood is:

$$L(\lambda_i|\mathbf{Y}) \propto \prod_{i=1}^{n} \lambda_i^{y_i} e^{-\lambda_i}$$

It's easy to get the log likelihood:

$$\ell(\lambda_i|\mathbf{Y}) \propto \log\left(\prod_{i=1}^{n} \lambda_i^{y_i} e^{-\lambda_i}\right)$$

Writing out the log-likelihood

$$\ell(\lambda_i|\mathbf{Y}) \propto \log\left(\prod_{i=1}^{n} \lambda_i^{y_i} e^{-\lambda_i}\right)$$

There's a proof of this step in the appendix of these slides:

$$\ell(\lambda_i|\mathbf{Y}) \propto \sum_{i=1}^{n} \log\left(\lambda_i^{y_i} e^{-\lambda_i}\right)$$

$$\ell(\lambda_i|\mathbf{Y}) \propto \sum_{i=1}^{n} \left(\log(\lambda_i^{y_i}) + \log(e^{-\lambda_i})\right)$$

$$\ell(\lambda_i|\mathbf{Y}) \propto \sum_{i=1}^{n} \left(y_i \log(\lambda_i) - \lambda_i\right)$$

What about the systematic components?

We still haven't accounted for the systematic part of our model!

Recall that our model was:

$$y_i \sim Pois(\lambda_i)$$
$$\lambda_i = e^{\beta_0}$$

Luckily, that's the easiest part to do. If our log-likelihood is:

$$\ell(\lambda_i|\mathbf{Y}) \propto \sum_{i=1}^{n} \Big(y_i \log(\lambda_i) - \lambda_i \Big)$$

We account for the systematic component by substituting in e^{β_0} every time we see a λ_i:

$$\ell(\beta_0|\mathbf{Y}) \propto \sum_{i=1}^{n} \Big(y_i \log(e^{\beta_0}) - e^{\beta_0} \Big) \propto \sum_{i=1}^{n} \Big(y_i \cdot \beta_0 - e^{\beta_0} \Big)$$

What about the systematic components?

If our model were more complicated, with covariates:

$$y_i \sim Pois(\lambda_i)$$
$$\lambda_i = e^{\beta_0 + \beta_1 x_i} = e^{\mathbf{X_i}\beta}$$

We account for them in the same way, by substituting in $e^{\mathbf{x_i}\beta}$ every time we see a λ_i:

$$\ell(\beta|\mathbf{Y}, \mathbf{X}) \propto \sum_{i=1}^{n} \left(y_i \log(e^{\mathbf{x_i}\beta}) - e^{\mathbf{x_i}\beta} \right)$$

$$\ell(\beta|\mathbf{Y}, \mathbf{X}) \propto \sum_{i=1}^{n} \left(y_i \cdot \mathbf{x_i}\beta - e^{\mathbf{x_i}\beta} \right)$$

How would we code this log-likelihood in R?

Next week Solé will show you up to use R to optimize log-likelihood functions. In order to do that you'll have to be comfortable coding the log-likelihood as a function.

$$\ell(\beta|\mathbf{Y},\mathbf{X}) \propto \sum_{i=1}^{n} \left(y_i \cdot \mathbf{x_i}\beta - e^{\mathbf{x_i}\beta} \right)$$

For this likelihood you'll have to write a function that takes a vector of your parameters ($\{\beta_0, \beta_1\}$), your outcome variable ($\mathbf{Y}$), and the matrix of your covariates ($\mathbf{X}$)

How would we code this log-likelihood in R?

$$\ell(\beta|\mathbf{Y},\mathbf{X}) \propto \sum_{i=1}^{n} \left(y_i \cdot \mathbf{x_i}\beta - e^{\mathbf{x_i}\beta} \right)$$

```
loglike <- function(parameters, outcome, covariates){

    cov <- as.matrix(cbind(1, covariates))

    xb <- cov %*% parameters

    sum(outcome * xb - exp(xb))
}
```

Questions?

Appendix

The switch the proportionality (and getting rid of constants) doesn't mess up the location of the maximum, but it does change the curvature of the function at the maximum.

And we use the curvature at the maximum to estimate our standard errors.

So doesn't getting rid of constants change the estimates we get of our standard errors?

Not exactly.

Appendix

As we'll talk about in the next couple weeks, the standard error of MLEs are, by definition,

$$Var(\hat{\theta}_{MLE}) = I^{-1}(\hat{\theta}_{MLE}) = \left[-E\left(\frac{\partial^2 \ell}{\partial \ell^2} \right) \right]^{-1}$$

$I^{-1}(\hat{\theta}_{MLE})$ is the inverse of the Fisher information matrix, evaluated at $\hat{\theta}_{MLE}$.

$\left[-E\left(\frac{\partial^2 \ell}{\partial \ell^2} \right) \right]^{-1}$ is the inverse of the negative Hessian matrix.

Notice that the Hessian is the expected value of the second derivative of the *log likelihood*, not the likelihood.

Appendix

Because the standard errors are a function of the log likelihood, it doesn't matter if we keep the constants in or not. They'll drop out when we differentiate anyway.
Here's an example with the Poisson likelihood:

$$L(\lambda|y) = k(y)\frac{\lambda^y e^{-\lambda}}{y!}$$

$$\ell(\lambda|y) = \log(k(y)) + y\log(\lambda) - \lambda - \log(y!)$$

$$\frac{\partial \ell}{\partial \lambda} = 0 + \frac{y}{\lambda} - 1 - 0$$

All the constants disappear when you take the first derivative of the log-likelihood, so it didn't matter if you kept them in the first place!

Which means that when we take out constants we'll get the same maximum likelihood estimates and the same standard errors on those estimates, regardless of whether we remove constants or not. Magic!

Appendix

When we switch from likelihoods to log-likelihoods we use the trick that

$$\log\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \log(x_i)$$

If x_i has 3 elements, then:

$$\log\left(\prod_{i=1}^{n} x_i\right) = \log(x_1 \cdot x_2 \cdot x_3)$$

And by our logarithm rules:

$$\log\left(\prod_{i=1}^{n} x_i\right) = \log(x_1) + \log(x_2) + \log(x_3)$$

$$\log\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \log(x_i)$$