Class 9: Properties of MLEs and Logistic Regression

MFE 402

Dan Yavorsky

Last Class

Fully parametric models let us:

- Define the likelihood and log-likelihood
- Express our belief about the relative plausibility of different parameter values
- Find the parameter value(s) that maximize the likelihood (MLE)

MLE Examples:

- Single parameter models, solved analytically
- Multiple parameter models, solved analytically
- Found the MLEs for β and σ^2 for the Normal Linear Regression Model, analytically
- Found the MLEs $\hat{\beta}_{\text{MLE}}$ for the Logit and Probit Regression Models, numerically

Topics for Today

- Properties of ML Estimators
    - Invariant
    - Consistent
    - Asymptotically Normal
    - Asymptotically Efficient
- Logit Example

Properties of Maximum Likelihood Estimators

Invariance

Recall $\hat{\theta}_{\text{MLE}}$ maximizes the likelihood and log-likelihood functions $L_n(\theta)$ and $\ell_n(\theta)$:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L_n(\theta)$$

Suppose you wanted to find the MLE $\hat{\delta}_{\text{MLE}}$ for $\delta = h(\theta)$, where $h(\cdot)$ is a 1:1 function.

A convenient mathematical result about maximizers is that:

$$\hat{\delta}_{\text{MLE}} = h(\hat{\theta}_{\text{MLE}}) = \arg\max_{\delta = h(\theta)} L_n^*(\delta)$$

In other words, if $\delta = h(\theta)$ then $\hat{\delta}_{\text{MLE}} = h(\hat{\theta}_{\text{MLE}})$. This is called the invariance property.

Example: Invariance

Let $Y \sim N(\mu, \sigma^2)$ with μ known, and define $\delta = 1/\sigma^2$. Then

$$\ell_n(\delta) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\delta) - \frac{\delta}{2}\sum_{i=1}^{n}(Y_i - \mu)^2$$

The FOC is:

$$\frac{d}{d\delta}\ell_n(\delta) = \frac{n}{2\delta} - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2 = 0$$

Which has solution:

$$\hat{\delta}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n}(Y_i - \mu)^2} = \frac{1}{\hat{\sigma}^2_{\text{MLE}}}$$

Notice that, given σ^2_{MLE} we could have just substituted in: if $\delta = 1/\sigma^2$ then $\hat{\delta}_{\text{MLE}} = 1/\hat{\sigma}^2_{\text{MLE}}$

Consistency

MLEs are consistent: $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta \qquad \text{as} \quad n \to \infty$

Proof idea:

1. notice that scaling the log-likelihood function does not change its maximizer:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell_n(\theta) = \arg\max_{\theta} \frac{1}{n}\ell_n(\theta)$$

2. notice that the average log-likelihood function has a "familiar" form:

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\theta) \equiv \bar{\ell}_n(\theta)$$

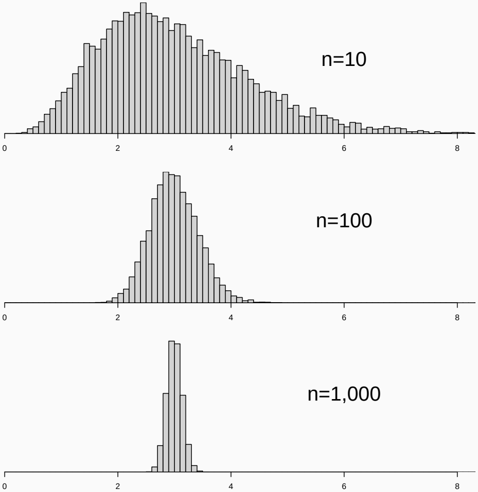
3. Then by the LLN: $\bar{\ell}_n(\theta) \xrightarrow{p} \mathbb{E}[\log f(Y|(\theta)] = \ell(\theta)$

4. And the maximizer of $\bar{\ell}_n(\theta)$ will converge in probability to true parameter value θ^*, which maximizes the expected log density $\ell(\theta)$

Example: Consistency Visualization

Let $Y \sim N(2,3)$ with μ known.
Plot empirical sampling distribution of $\hat{\sigma}^2_{\text{MLE}}$
for $n = 10$, 100, and 1000

```r
reps <- 10000
sig2_mle <- vector(length=reps)
for(n in c(10, 100, 1000)) {
    for(i in 1:reps) {
        y <- rnorm(n, mean=2, sd=sqrt(3))
        sig2_mle[i] <- sum((y-2)^2)/n
    }
    histogram(sig2_mle)
}
```



Score and Hessian

The Likelihood Score is the partial derivative of the log-likelihood function w.r.t. θ (or vector of partial derivatives if θ is a vector):

$$S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i|\theta)$$

which tells us how sensitive the log-likelihood is to the parameter vector.

The Likelihood Hessian is the negative second derivative of the log-likelihood function:

$$\mathscr{H}_n(\theta) = -\frac{\partial^2}{\partial\theta\,\partial\theta'}\ell_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\,\partial\theta'} \log f(Y_i|\theta)$$

which tells us the degree of curvature in the log-likelihood.

The Efficient Score

The Efficient Score is the derivative of the log-likelihood for a single observation, evaluated at the random vector Y and the true parameter value θ^*:

$$S = \frac{\partial}{\partial\theta} \log f(Y|\theta^*)$$

S plays important roles in asymptotic distribution and testing theory.

S is a random variable/vector because it is a function of the random variable/vector Y.

- The efficient score is mean-zero in expectation: $\mathbb{E}[S] = 0$
- The Fisher Information (a matrix when θ is a vector) is the variance of the efficient score:

$$\mathscr{I}_\theta = \text{Var}(S) = \mathbb{E}[SS'] - (\mathbb{E}[S])(\mathbb{E}[S])' = \mathbb{E}[SS']$$

Information Matrix Equality

The Expected Hessian equals the expectation of the Likelihood Hessian for a single observation:

$$\mathscr{H}_\theta = \text{-}\mathbb{E}\left[\frac{\partial^2}{\partial\theta\,\partial\theta'}\log f(Y|\theta)\right]$$

The **Information Matrix Equality** is the result that $\mathscr{H}_\theta = \mathscr{I}_\theta$

- ie, that the Expected Hessian is equal to the Fisher Information Matrix
- ie, that the curvature in the likelihood is equal to the variance of the efficient score

There's no intuition here.

It's a fascinating result that simplifies the asymptotic properties of MLEs.

Example: Score, Hessian, and Info Matrix

Suppose $Y \sim \text{Expon}(\lambda)$

- The density is $f(y|\lambda) = \lambda^{-1} \exp(-y/\lambda)$ (a different parameterization than last class)
- The log-density is $\log f(y|\lambda) = -\log(\lambda) - y/\lambda$
- The expectation is $\mathbb{E}[Y] = \lambda$
- The variance is $\text{Var}(Y) = \lambda^2$

Then:

- The efficient score is $S = \frac{d}{d\lambda} \log f(Y|\lambda) = -\frac{1}{\lambda} + \frac{Y}{\lambda^2}$
- Its expectation is $\mathbb{E}[S] = -\frac{1}{\lambda} + \frac{\lambda}{\lambda^2} = 0$
- Its variance is $\text{Var}(S) = \text{Var}\left(-\frac{1}{\lambda} + \frac{Y}{\lambda^2}\right) = \frac{\text{Var}(Y)}{\lambda^4} = \frac{1}{\lambda^2}$
- The expected hessian is $\mathscr{H}_\lambda = \mathbb{E}\left[-\frac{d^2}{d\lambda^2} \log f(Y|\lambda)\right] = -\frac{1}{\lambda^2} + 2\frac{\mathbb{E}[Y]}{\lambda^3} = \frac{1}{\lambda^2}$

$$\text{And so} \quad \mathscr{I}_\theta = \frac{1}{\lambda^2} = \mathscr{H}_\theta$$

Reminder: A Taylor Series Expansion

Suppose you want to find $f(b)$ and you know $f(a)$ for some function $f(\cdot)$ and points a and b.

The m^{th} degree **Taylor Series Polynomial** approximates $f(b)$ with a polynomial of degree m:

$$f(b) \simeq f(a) + \frac{1}{1!}f'(a)(b-a) + \frac{1}{2!}f''(a)(b-a)^2 + \frac{1}{3!}f'''(a)(b-a)^3 + \cdots + \frac{1}{m!}f^{(m)}(a)(b-a)^m$$

Many times, we'll use the first-order Taylor Series Expansion to approximate $f(b)$ with a linear function:

$$f(b) \simeq f(a) + f'(a)(b-a)$$

Asymptotic Normality

We will not have an explicit expression for the MLE in most models.

However, using a Taylor Series Expansion, the MLE can be approximated by (a matrix scale of) the sample average of the efficient scores:

$$0 = \frac{\partial}{\partial\theta}\bar{\ell}_n(\hat{\theta}) \simeq \frac{\partial}{\partial\theta}\bar{\ell}_n(\theta^*) + \frac{\partial^2}{\partial\theta\,\partial\theta'}\bar{\ell}_n(\theta^*)\left(\hat{\theta}_{\text{MLE}} - \theta^*\right)$$

Here $f(\cdot)$ is $\partial\bar{\ell}_n(\cdot)/\partial\theta$, b is $\hat{\theta}_{\text{MLE}}$, and a is θ^*. This can be re-written as

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*) \simeq \left(\underbrace{-\frac{\partial^2}{\partial\theta\,\partial\theta'}\bar{\ell}_n(\theta^*)}_{\xrightarrow{p}\mathscr{H}_\theta}\right)^{-1}\left(\underbrace{\sqrt{n}\frac{\partial}{\partial\theta}\bar{\ell}_n(\theta^*)}_{\xrightarrow{d}N(0,\mathscr{I}_\theta)}\right)$$

And so

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*) \xrightarrow{d} \mathscr{H}_\theta^{-1}N(0,\mathscr{I}_\theta) = N(0,\mathscr{H}_\theta^{-1}\mathscr{I}_\theta\mathscr{H}_\theta^{-1}) = N(0,\mathscr{I}_\theta^{-1})$$
$$\Rightarrow \hat{\theta_{\text{MLE}}} \sim N\left(\theta^*,\ \mathscr{I}_\theta^{-1}/n\right)$$

Cramer-Rao Lower Bound

Theorem: If $\tilde{\theta}$ is an unbiased estimator of θ, then $\mathrm{Var}(\tilde{\theta}) \geq \mathscr{I}^{-1}/n$

This is a famous result. In the class of unbiased estimators, the lowest possible variance is the inverse of the Fisher Information scaled by sample size.

We describe an estimator as being asymptotically Cramer-Rao Efficient if its asymptotic distribution attains the Cramer-Rao lower bound.

Maximum Likelihood Estimators are asymptotically Cramer-Rao Efficient!

Example: Cramer-Rao Lower Bound

Suppose $Y \sim \text{Expon}(\lambda)$ with density $f(y|\lambda) = \lambda^{-1} \exp(-y/\lambda)$

Then:

- The expected hessian is $\mathscr{H}_\lambda = \mathbb{E}\left[-\frac{d^2}{d\lambda^2} \log f(Y|\lambda)\right] = -\frac{1}{\lambda^2} + 2\frac{\mathbb{E}[Y]}{\lambda^3} = \frac{1}{\lambda^2}$
- Therefore, the information matrix is $\mathscr{I}_\theta = \frac{1}{\lambda^2}$
- The CRLB is $\mathscr{I}_\theta^{-1}/n = \lambda^2/n$

Last class we found the MLE $\hat{\lambda}_{\text{MLE}} = \bar{Y}$

- $\bar{Y}$ is an unbiased estimator of λ
- $\text{Var}(\bar{Y}) = \text{Var}(Y)/n = \lambda^2/n$
- Thus $\hat{\lambda}_{\text{MLE}}$ is Cramer-Rao efficient

Estimating the Asymptotic Variance

We have 3 ways to estimate the Hessian (ie, once inverted, the asymptotic variance of the MLE)

0. Expected Hessian Estimator

$$\hat{V}_0 = \hat{\mathscr{H}}_\theta^{-1} \quad \text{where} \quad \hat{\mathscr{H}}_\theta = \mathscr{H}_\theta(\hat{\theta})$$

1. Sample Hessian Estimator

$$\hat{V}_1 = \hat{\mathscr{H}}_\theta^{-1} \quad \text{where} \quad \hat{\mathscr{H}}_\theta = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\,\partial\theta'} \log f(Y_i|\hat{\theta}) = -\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta'}\ell_n(\hat{\theta})$$

2. Outer Product Estimator

$$\hat{V}_2 = \hat{\mathscr{I}}_\theta^{-1} \quad \text{where} \quad \hat{\mathscr{I}}_\theta = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial}{\partial\theta} \log f(Y_i|\hat{\theta})\right)\left(\frac{\partial}{\partial\theta} \log f(Y_i|\hat{\theta})\right)'$$

Asymptotic standard errors are constructed by taking the square roots of the diagonal elements of $n^{-1}\hat{V}$. When θ is a scalar, this is $s(\hat{\theta}) = \sqrt{\hat{V}/n}$.

Example: MLE Variance Estimates

Suppose $Y \sim \text{Expon}(\lambda)$ with density $f(y|\lambda) = \lambda^{-1}\exp(-y/\lambda)$

- The MLE is $\hat{\lambda}_{\text{MLE}} = \bar{Y}$
- 1st derivative of the log density is $\frac{d}{d\lambda}\log f(y|\lambda) = -1/\lambda + y/\lambda^2 = (y - \lambda)/\lambda^2$
- 2nd derivative of the log density is $\frac{d^2}{d\lambda^2}\log f(y|\lambda) = 1/\lambda^2 - 2y/\lambda^3$

$$\mathscr{H}_\lambda(\lambda) = 1/\lambda^2 \quad \Rightarrow \quad \mathscr{H}_\lambda(\hat{\lambda}) = 1/\bar{Y}^2$$

$$\hat{\mathscr{H}}_\lambda(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda^2} - 2\frac{Y_i}{\lambda^3} = -\frac{1}{\lambda^2} + 2\frac{\bar{Y}}{\lambda^3} \quad \Rightarrow \quad \hat{\mathscr{H}}_\lambda(\hat{\lambda}) = -\frac{1}{\bar{Y}^2} + 2\frac{\bar{Y}}{\bar{Y}^3} = 1/\bar{Y}^2$$

$$\hat{\mathscr{I}}_\lambda(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \lambda}{\lambda^2}\right)^2 \quad \Rightarrow \quad \hat{\mathscr{I}}_\lambda(\hat{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \bar{Y}}{\bar{Y}^2}\right)^2 = \frac{\hat{\sigma}_Y^2}{\bar{Y}^4}$$

Thus, $\hat{V}_0 = \hat{V}_1 = \bar{Y}^2$ and $\hat{V}_2 = \bar{Y}^4/\hat{\sigma}_Y^2$
$\Rightarrow s_0(\lambda) = s_1(\lambda) = \bar{Y}/\sqrt{n}$ and $s_2(\lambda) = \bar{Y}^2/(\hat{\sigma}_Y\sqrt{n})$

Example: MLE Variance Estimates in practice

Last class, we found the MLE for the logit model.

To find standard errors, use the `hessian=TRUE` argument in `optim()`:

```
ll <- function(beta, X, y) {
    pi_i <- 1 / (1 + exp(-1 * X %*% beta))
    ll <- sum(y*log(pi_i) + (1-y)*log(1-pi_i))
    return(ll)
}

out_logit <- optim(par=rep(0,k), fn=ll,
                   X=X, y=y, hessian=TRUE,   # <-- add hessian=TRUE
                   control=list(fnscale=-1))

est <- out_logit$par
se  <- sqrt(diag(-1*solve(out$hessian)))     # <-- need to multiply by -1
```

Confidence Intervals and Hypothesis Tests

Our familiar test statistic has an asymptotically standard-normal distribution:

$$T(\theta_0) = \frac{\hat{\theta}_{\text{MLE}} - \theta_0}{s(\hat{\theta}_{\text{MLE}})} \xrightarrow{d} N(0, 1)$$

As with OLS, we can use this to construct confidence intervals and hypothesis tests.

- Confidence intervals:
  $\hat{\theta}_{\text{MLE}} \pm c_{\alpha/2} s(\hat{\theta}_{\text{MLE}})$ where e.g. $c = 1.96$ for a 95% CI
- Hypothesis tests:
  reject $H_0 : \theta = \theta_0$ if $|T(\theta_0)| > c_{\alpha/2}$ where e.g. $c = 1.96$ for a test with 95% confidence

Summary of MLE Properties

Maximum Likelihood Estimators:

- Are invariant under 1:1 transformations, enables simplification
- Are consistent, as any decent estimator ought to be
- Are asymptotically normal, which enables inference
- Are asymptotically efficient, which is hard to beat
- Estimate the best-fitting model in the class of $f(Y|\theta)$ whether or not the model is correctly specified – analogous to how OLS is the best linear approximation to the true CEF (we did not discuss or prove this last one)

In finite samples, MLEs may be biased, not-normal distributed, and not efficient.

Logit Example

Logit Example: Intro

Suppose we have a binary dependent variable indicating default on a loan ($Y = 1$ is default) and a single covariate (X) indicating the borrower's FICO score (300-850, higher is better).

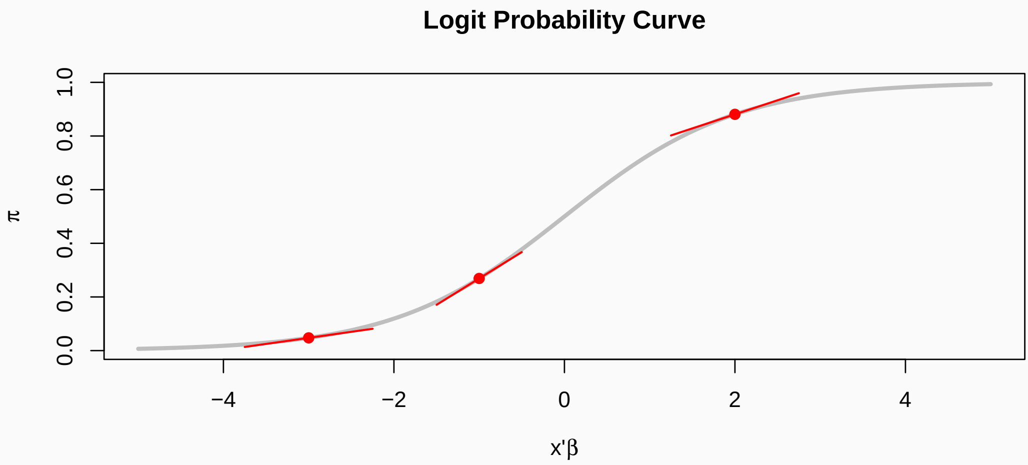
Recall that the CEF is the probability of default:

$$\mathbb{E}[Y|X] = 1 \times \mathbb{P}(Y = 1|X) + 0 \times \mathbb{P}(Y = 0|X) = \mathbb{P}(Y = 1|X) = \pi$$

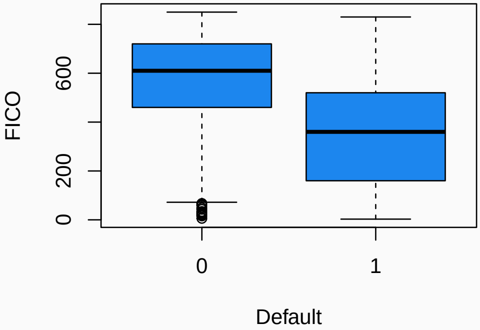
We will model the probability of default using the inverse of the logit link function:

$$\pi_i = \mathbb{P}(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(\text{-}(\beta_0 + \beta_1 X_i))}$$

Logit Example: Logit Curve

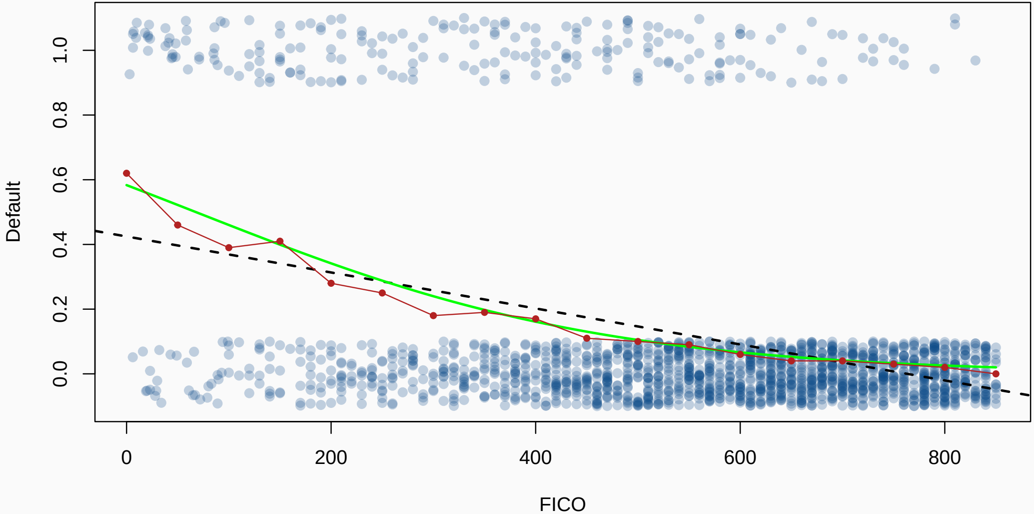


Logit Example: The Data



	FICO_group	prob_default
	<num>	<num>
1:	0	0.62
2:	50	0.46
3:	100	0.39
4:	150	0.41
5:	200	0.28
6:	250	0.25
7:	300	0.18
8:	350	0.19
9:	400	0.17
10:	450	0.11
11:	500	0.10
12:	550	0.09
13:	600	0.06
14:	650	0.04
15:	700	0.04
16:	750	0.03
17:	800	0.02
18:	850	0.00

Logit Example: Fitted Curves



Logit Example: `glm`

Fit the model using `glm()`; show coefficients and standard errors:

```
out <- glm(Default ~ FICO, data=default, family=binomial(link="logit"))
summary(out)$coefficients |> round(4)
```

```
            Estimate Std. Error  z value Pr(>|z|)
(Intercept)   0.3365     0.1684   1.9984   0.0457
FICO         -0.0050     0.0004 -13.3210   0.0000
```

Logit Example: `optim`

Fit the model using `optim()`; show coefficients and standard errors:

```r
X <- cbind(1, default$FICO)
y <- default$Default

ll <- function(beta, X, y) {
    pi_i <- 1 / (1 + exp(-1 * X %*% beta))
    sum(y*log(pi_i) + (1-y)*log(1-pi_i))
}
```

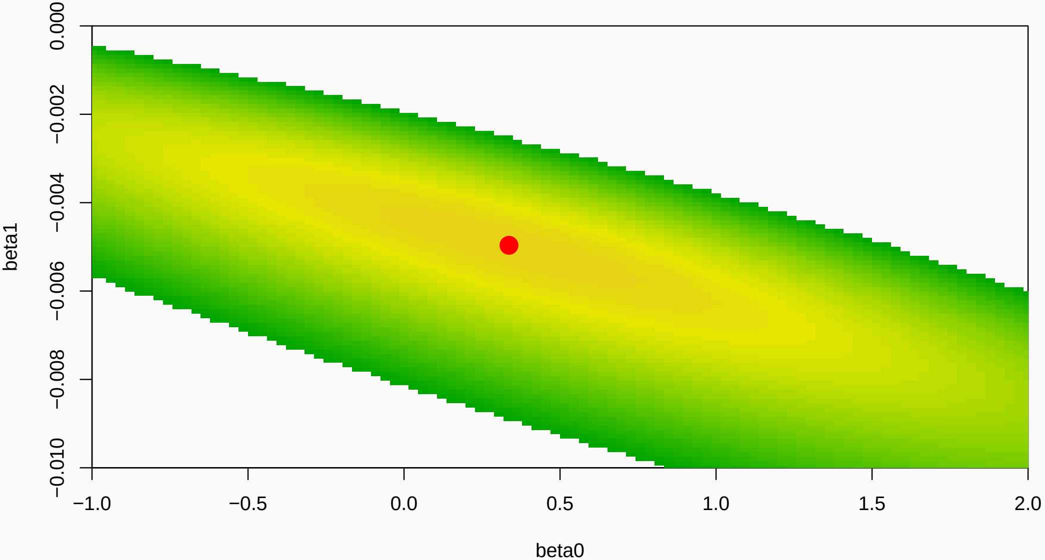
```r
out <- optim(par=c(0,0), fn=ll,
             X=X, y=y, hessian=T,
             control=list(fnscale=-1))

Hinv <- -1*solve(out$hessian)

cbind(coefs = out$par,
      sterr = sqrt(diag(Hinv))) |>
    round(4)
```

```
       coefs  sterr
[1,]  0.3364 0.1636
[2,] -0.0050 0.0003
```

Logit Example: Likelihood Surface



Logit Example: Interpreting Coefficients

The interpretation of the regression slope coefficients is the change in log odds:

"A unit increase in X_j is associated with a β_j increase in the log odds of $Y = 1$"

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

For binary outcome models, the typical quantity of interest (QOI) is the change in the probability $\mathbb{P}(Y = 1)$ with respect to a specific X variable:

$$\frac{\partial \mathbb{P}(Y = 1)}{\partial X_j} = \beta_j \times \mathbb{P}(Y = 1) \times (1 - \mathbb{P}(Y = 1))$$

To practically implement this, we can either:

1. Calculate $\mathbb{P}(Y = 1)$ for each observation and average the results
2. Calculate $\mathbb{P}(Y = 1)$ for the average value of the X's
3. Calculate the difference in $\mathbb{P}(Y = 1)$ for two values of X_j for each observation, and then average the resulting differences

Logit Example: Interpreting Coefficients

Suppose we had another variable X_2 in the model. The QOI for $X_1 = \text{FICO}$ is:

```r
# using observed values
X <- cbind(1, default$FICO, default$x2)
pr <- 1 / (1 + exp(-1 * X %*% out$par))
out$par[2] * mean(pr) * (1 - mean(pr))

# using average values
X <- cbind(1, mean(default$FICO),
              mean(default$x2))
pr <- 1 / (1 + exp(-1 * X %*% out$par))
out$par[2] * pr * (1 - pr)
```

```r
# direct calculation
X1 <- cbind(1, default$FICO, defualt$x2)
X2 <- cbind(1, default$FICO + 1, defualt$x2)

pr1 <- 1 / (1 + exp(-1 * X1 %*% out$par))
pr2 <- 1 / (1 + exp(-1 * X2 %*% out$par))
mean(pr2 - pr1)
```

Pseudo R-squared

Recall the Bernoulli density: $f(y|\pi) = \pi^y \times (1 - \pi)^{1-y}$

- Notice that an individual unit's contribution to the Likelihood is 1 if $\hat{\pi}_i = y_i$ (log-likelihood contribution is zero)
- Contribution to the Likelihood is between 0 and 1 if $0 < \hat{\pi}_i < 1$ (log-likelihood contribution is negative)

We define:

- $\log L_0$ is the log likelihood of the null model (intercept only)
- $\log L_M$ is the log likelihood of the full model (with predictors)

By comparing $\log L_0$ and $\log L_M$, we can measure the degree to which using the explanatory variables improves the predictability of Y

The pseudo R^2 is a measure of model fit that is analogous to the R^2 in linear regression:

$$R^2 = 1 - \frac{\log L_M}{\log L_0}$$

Next Time

Introduce Bayesian Inference

- Frequentist vs Bayesian Philosophy
- Bayes' Rule, again
- Prior, Likelihood, and Posterior
- Conjugate Priors
- MCMC sampling of the Posterior Distribution