Likelihood and Deviance

Recap

Definition: The *likelihood* $L(Model) = P(Data|Model)$ of a model is the probability of the observed data, given that we assume a certain model and certain values for the parameters that define that model.

Coin example: flip a coin 5 times, with $\pi_i = P(Heads)$

- Model: $Y_i \sim Bernoulli(\pi_i)$, and $\widehat{\pi}_i = 0.9$
- Data: $y_1, \ldots, y_5 = T, T, T, T, H$
- Likelihood: $L(\widehat{\pi}_i) = P(y_1, \ldots y_5 | \pi_i = 0.9) = 0.00009$

Recap

Maximum likelihood estimation: pick the parameter estimate that maximizes the likelihood.

Coin example: flip a coin 5 times, with $\pi_i = P(Heads)$

- Observed data: T, T, T, T, H
- Likelihood: $L(\widehat{\pi}_i) = (1 - \widehat{\pi}_i)^4(\widehat{\pi}_i)$
- Choose $\widehat{\pi}_i$ to maximize $L(\widehat{\pi}_i)$

Logistic regression

$$Y_i \sim Bernoulli(\pi_i) \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

Or in other words,

$$\pi_i = \frac{\exp\{\beta_0 + \beta_1 X_i\}}{1 + \exp\{\beta_0 + \beta_1 X_i\}}$$

- To fit this model, we need to obtain estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$
- Let's start by exploring the likelihood with this model

Logistic regression likelihood

$$Y_i \sim Bernoulli(\pi_i) \quad \pi_i = \frac{\exp\{\beta_0 + \beta_1 X_i\}}{1 + \exp\{\beta_0 + \beta_1 X_i\}}$$

Data: $(X_1, Y_1), \ldots, (X_n, Y_n)$

Likelihood:

Logistic regression log likelihood

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i))$$

Logistic regression log likelihood

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i \log\left(\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\right) + \\ \sum_{i=1}^{n} (1-Y_i) \log\left(1 - \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\right)$$

- Because we have two parameters, β_0 and β_1, the situation is more difficult
- The math to find MLEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is more complex than we will cover
- R calculates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for us

Logistic regression in R

Data: Grad application data

- `admit`: accepted to grad school? (0 = no, 1 = yes)
- `gre`: GRE score
- `gpa`: undergrad GPA
- `rank`: prestige of undergrad institution

Let's fit a logistic regression model with GPA as the predictor.

Logistic regression in R

Data: Grad application data

- `admit`: accepted to grad school? (0 = no, 1 = yes)
- `gre`: GRE score
- `gpa`: undergrad GPA
- `rank`: prestige of undergrad institution

Let's fit a logistic regression model with GPA as the predictor.

```
glm(admit ~ gpa, family = binomial, data = grad_app)
```

Logistic regression in R

```
glm(admit ~ gpa, family = binomial, data = grad_app)
```

```
##
## Call:  glm(formula = admit ~ gpa, family = binomial, dat
##
## Coefficients:
## (Intercept)          gpa
##      -4.358        1.051
##
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:          500
## Residual Deviance: 487      AIC: 491
```

What are $\widehat{\beta}_0$ and $\widehat{\beta}_1$?

Logistic regression in R

```
...
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:         500
## Residual Deviance: 487      AIC: 491
...
```

- For linear regression, the bottom part of the output usually contains things like R^2 and R^2_{adj} -- measures of how well the model fits the data.

> What quantity have we been exploring that allows us to evaluate how well the model fits the data?

Logistic regression in R

```
...
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:         500
## Residual Deviance: 487      AIC: 491
...
```

- For linear regression, the bottom part of the output usually contains things like R^2 and R^2_{adj} -- measures of how well the model fits the data.

> Does R report the likelihood of the fitted model?

Deviance

R reports the *deviance*, rather than the likelihood.

Deviance:

Deviance

```
...
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:         500
## Residual Deviance: 487     AIC: 491
...
```

Deviance $= -2 \log L = 487$

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_7.html

Class activity: deviance

```
glm(admit ~ gre, family = binomial, data = grad_app)
```

```
...
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:        500
## Residual Deviance: 486.1     AIC: 490.1
...
```

> What is the deviance of my fitted model?

Class activity: deviance

```
glm(admit ~ gre, family = binomial, data = grad_app)
```

```
...
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:           500
## Residual Deviance: 486.1      AIC: 490.1
...
```

> Which predictor (GRE or GPA) gives a model with a better fit?

Class activity: deviance

```
glm(admit ~ gre, family = binomial, data = grad_app)
```

```
...
## Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
## Null Deviance:          500
## Residual Deviance: 486.1     AIC: 490.1
...
```

> Which predictor (GRE or GPA) gives a model with a better fit?

GRE has a slightly smaller deviance (486.1 vs. 487), so GRE gives a slightly better fit.

Class activity: hypotheses

$$Y_i \sim Bernoulli(\pi_i) \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{GRE}_i$$

We want to know if there is actually a relationship between GRE score and grad school admission.

> How can I express this as null and alternative hypotheses about one or more model parameters?

Comparing deviances

```
...
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 486.06  on 398  degrees of freedom
...
```

499.98 = deviance for intercept-only model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0$$

486.06 = deviance for full model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \mathrm{GRE}_i$$

drop-in-deviance:

Comparing deviances

drop-in-deviance: $G =$ deviance for reduced model - deviance for full model = 13.92

Full model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{GRE}_i$

Reduced model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0$

> Why is G always ≥ 0?

Comparing deviances

drop-in-deviance: $G =$ deviance for reduced model - deviance for full model = 13.92

Full model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{GRE}_i$

Reduced model: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0$

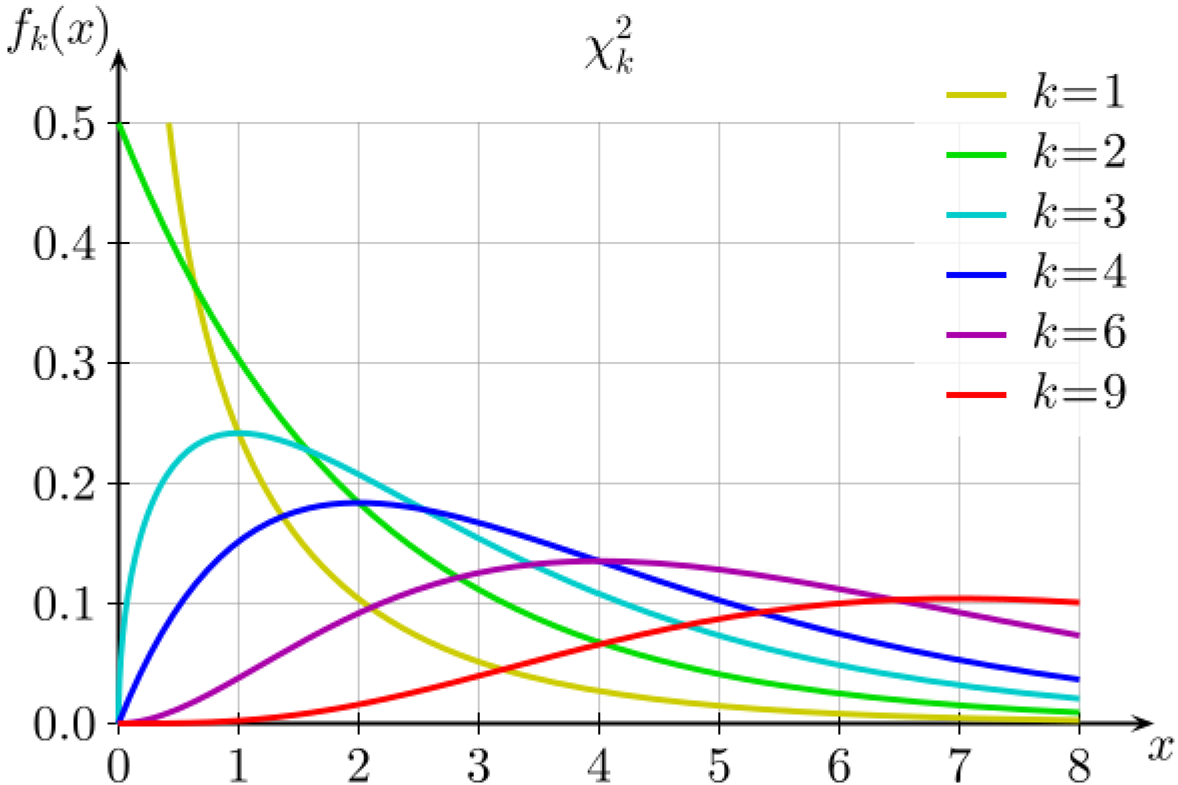
$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0$

> If H_0 is true, how unusual is $G = 13.92$?

χ^2 distribution

Under $H_0, G \sim \chi^2_{df_{\text{reduced}} - df_{\text{full}}}$

χ^2_k distribution: parameterized by degrees of freedom k



Computing a p-value

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{GRE}_i$$

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0$$

$G =$ deviance for reduced model - deviance for full model = 13.92 $\sim \chi^2_1$

```r
pchisq(13.92, df = 1, lower.tail=FALSE)
```

```
## [1] 0.0001907579
```

Concept check

> Our p-value is 0.0002. What is the most appropriate conclusion? Go to https://pollev.com/ciaranevans637 to respond.

(A) We reject the null hypothesis, since $p < 0.05$.

(B) We fail to reject the null hypothesis, since $p < 0.05$.

(C) The data provide strong evidence of a relationship between GRE score and the probability of admission to graduate school.

(D) The data do not provide strong evidence of a relationship between GRE score and the probability of admission to graduate school.

Likelihood ratio test for nested models