Infinite Estimates in Logistic Regression

Grinnell College

October 30, 2026

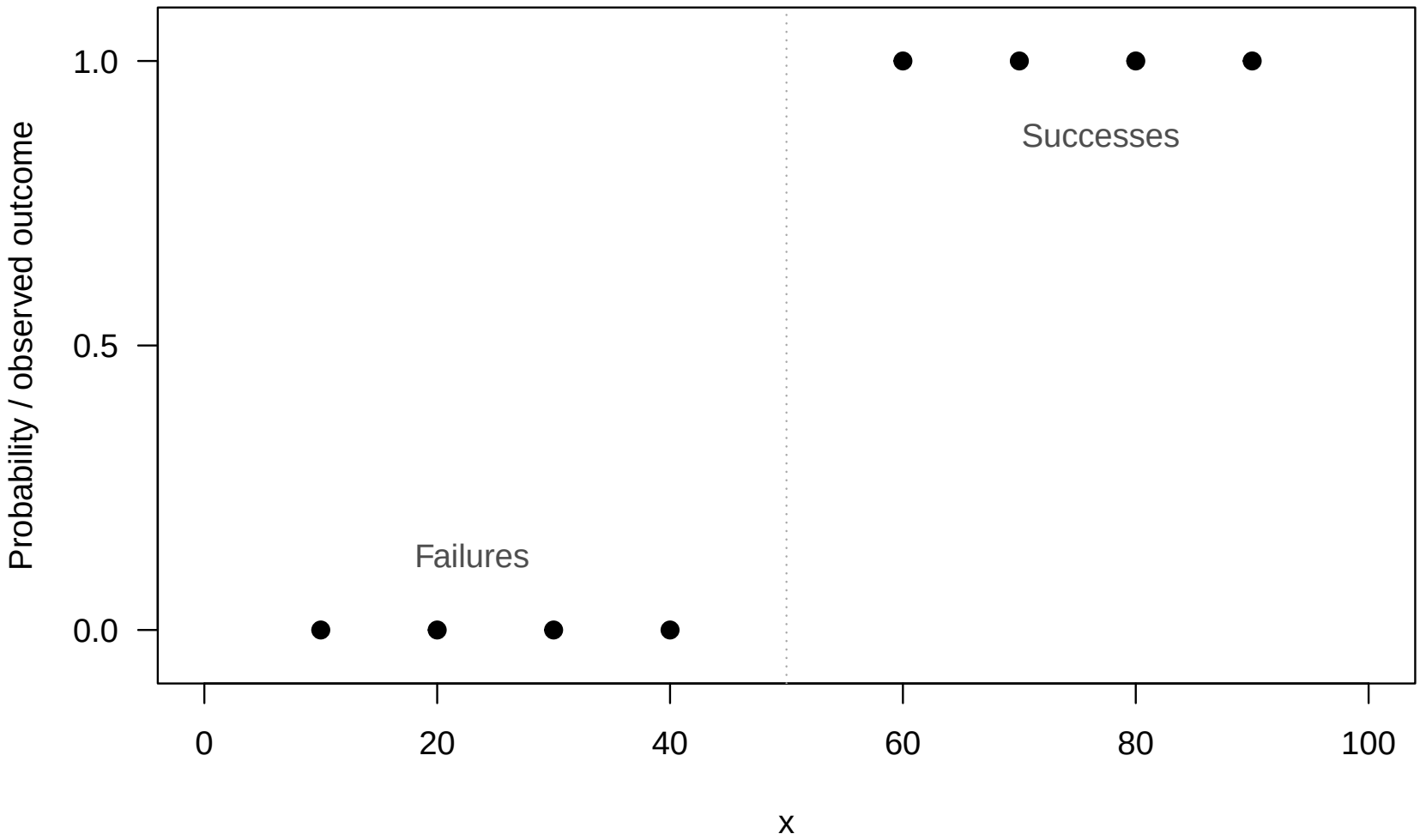
A Small Logistic Regression Example

Consider the following eight observations:

```
> x <- c(10, 20, 30, 40, 60, 70, 80, 90)
> y <- c(0, 0, 0, 0, 1, 1, 1, 1)
> toy <- data.frame(x, y)
> toy
   x y
1 10 0
2 20 0
3 30 0
4 40 0
5 60 1
6 70 1
7 80 1
8 90 1
```

Can we predict the outcome perfectly using a cutoff for x?

Complete Separation



All failures have $x \leq 40$ and all successes have $x \geq 60$. A cutoff at 50 separates the two outcomes.

The Logistic Curve as the Slope Increases

Recall the fitted probability:

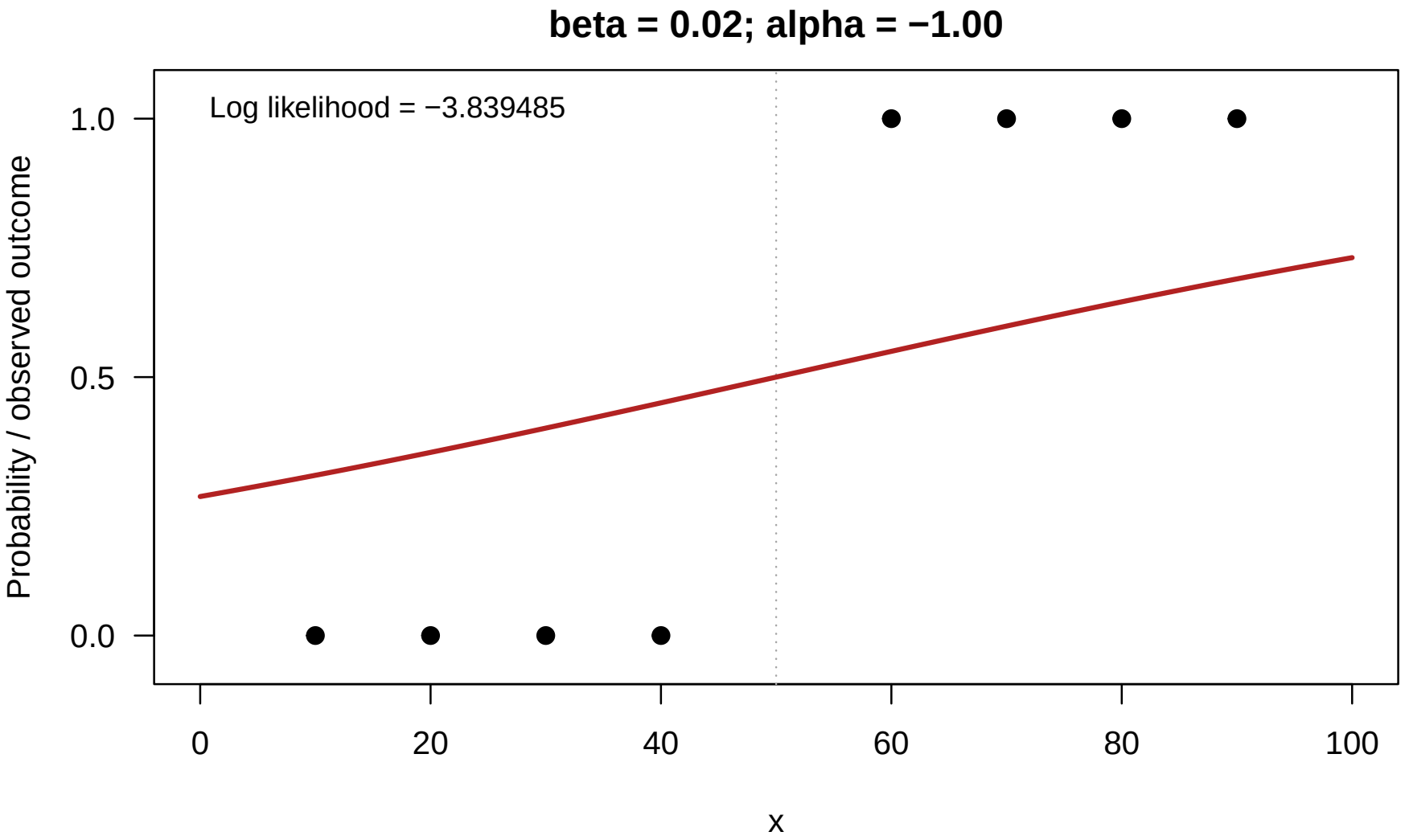
$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

To keep the midpoint at $x = 50$, set $\alpha = -50\beta$:

$$\text{logit}\{\pi(x)\} = \beta(x - 50).$$

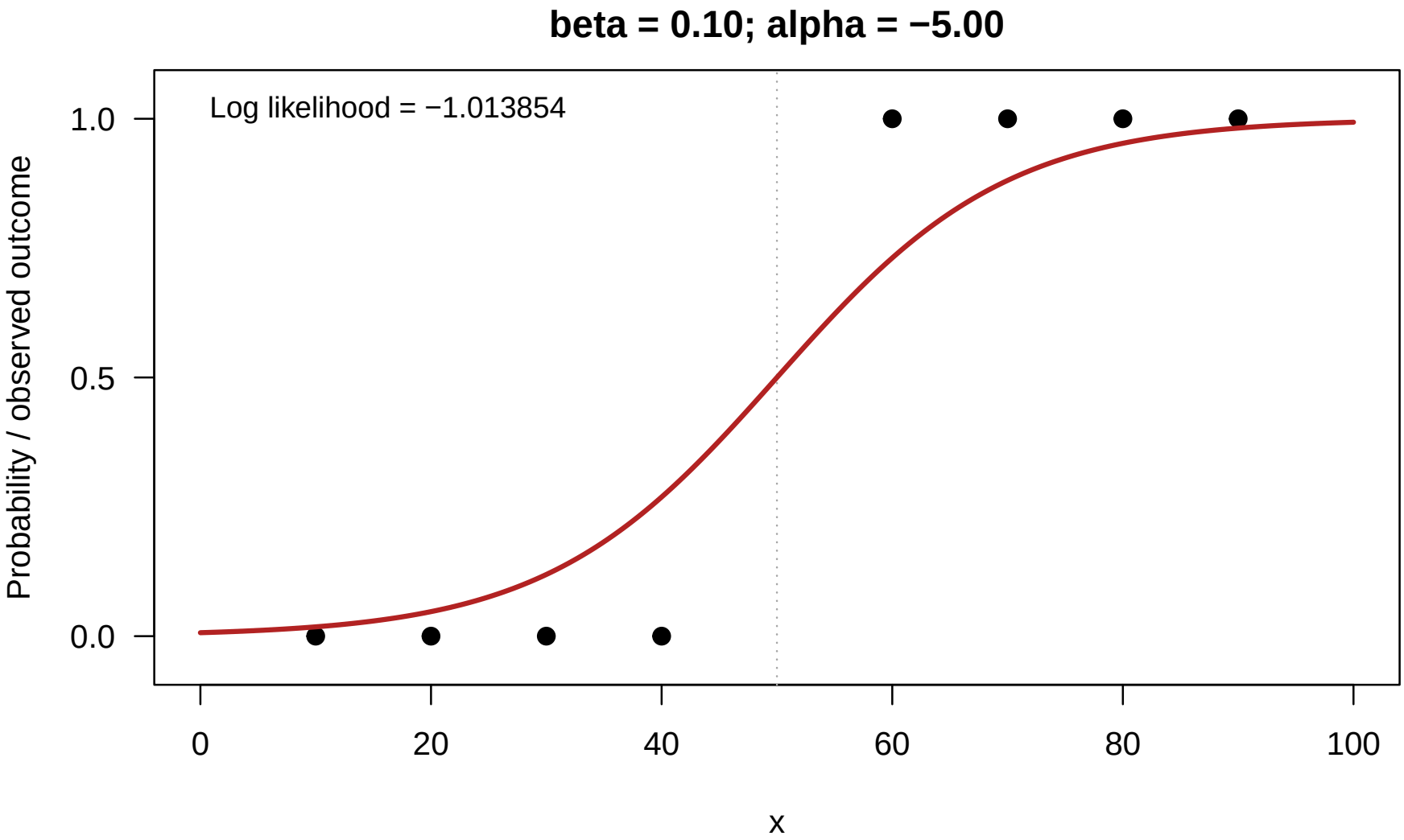
- $x = 50$: $\pi(x) = 0.5$ for every value of β.
- $x < 50$: increasing $\beta > 0$ pushes $\pi(x)$ toward 0.
- $x > 50$: increasing $\beta > 0$ pushes $\pi(x)$ toward 1.

Increasing the Slope: $\beta = 0.02$



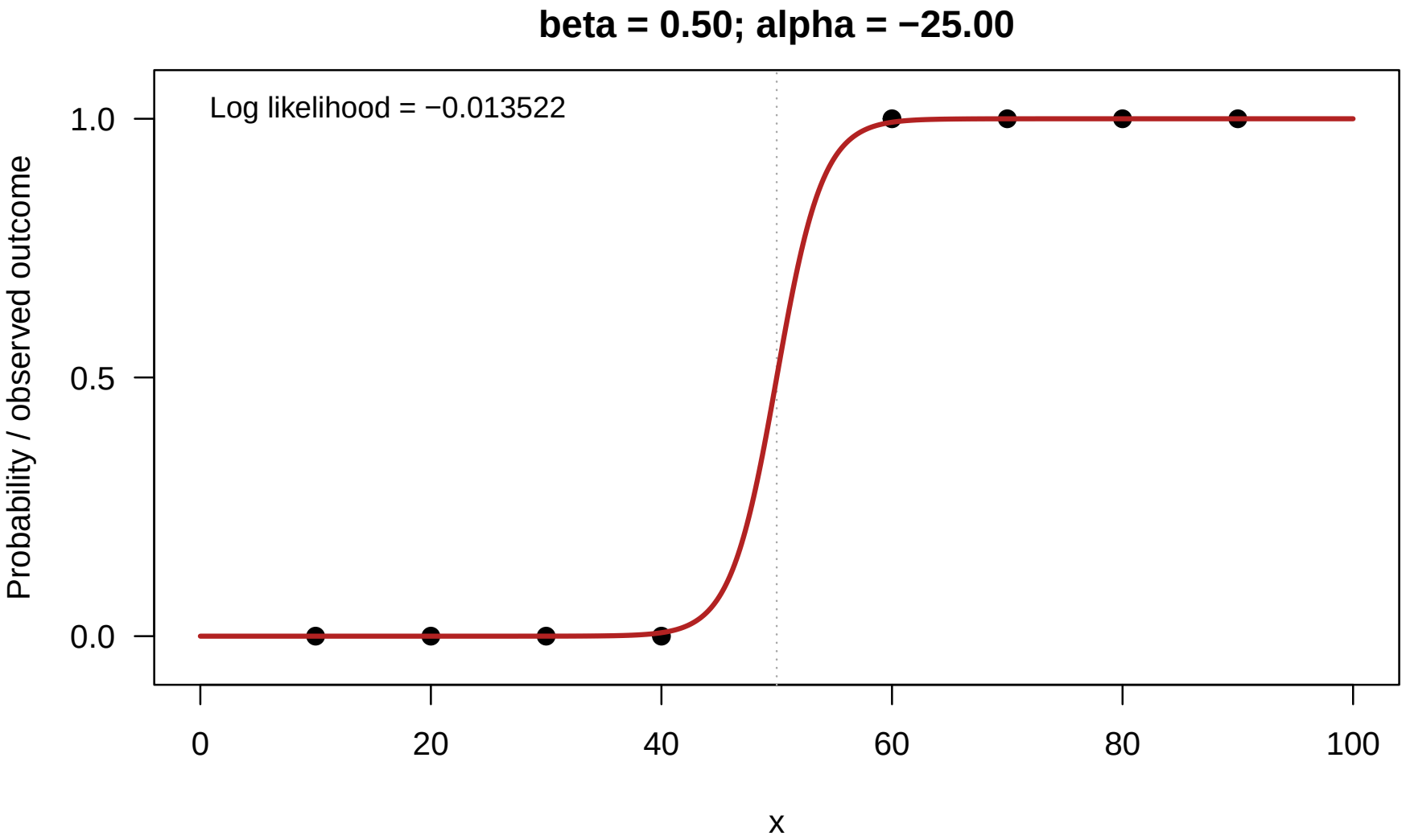
The curve assigns moderate probabilities to both outcomes.

Increasing the Slope: $\beta = 0.10$



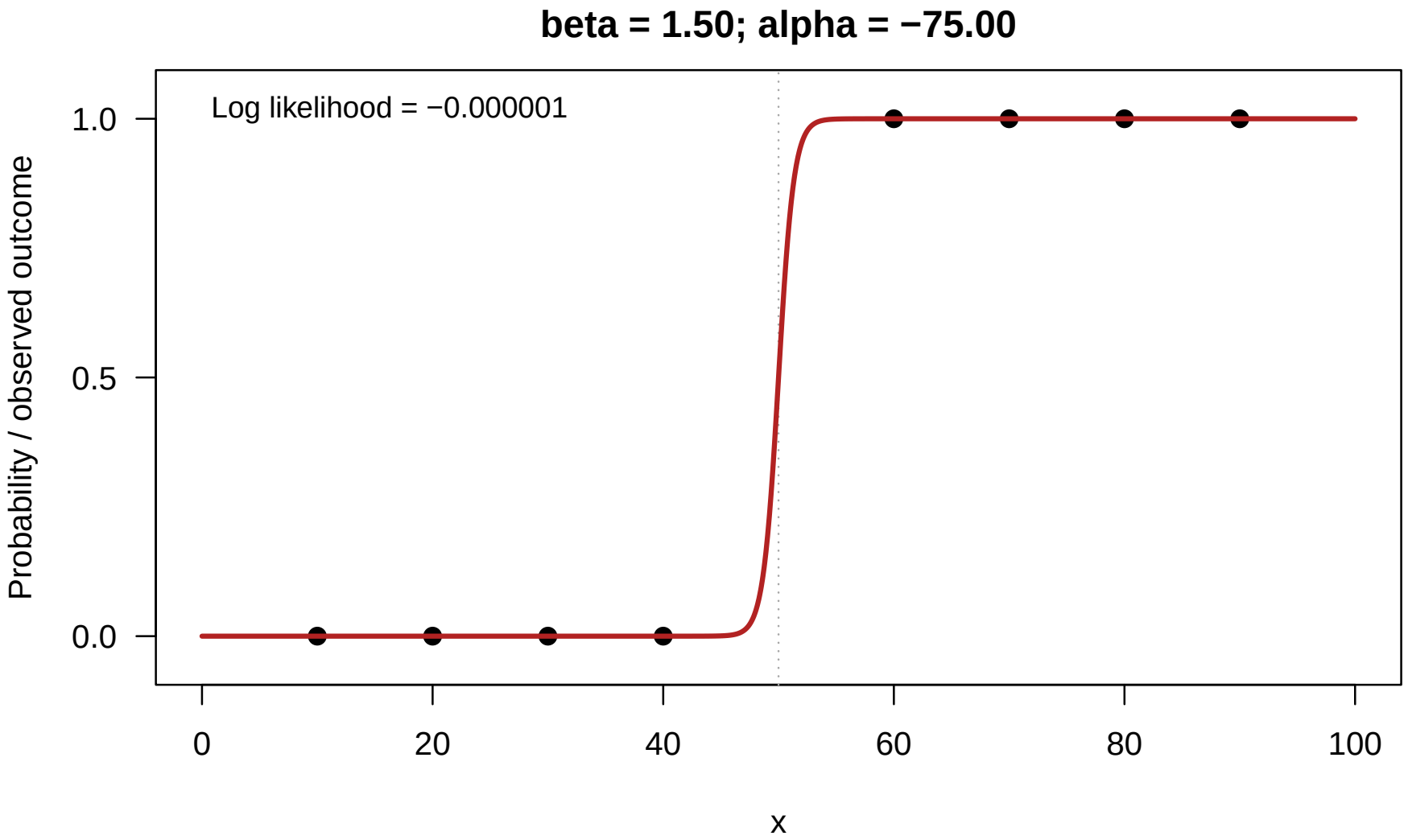
Each observed outcome receives a higher probability.

Increasing the Slope: $\beta = 0.50$



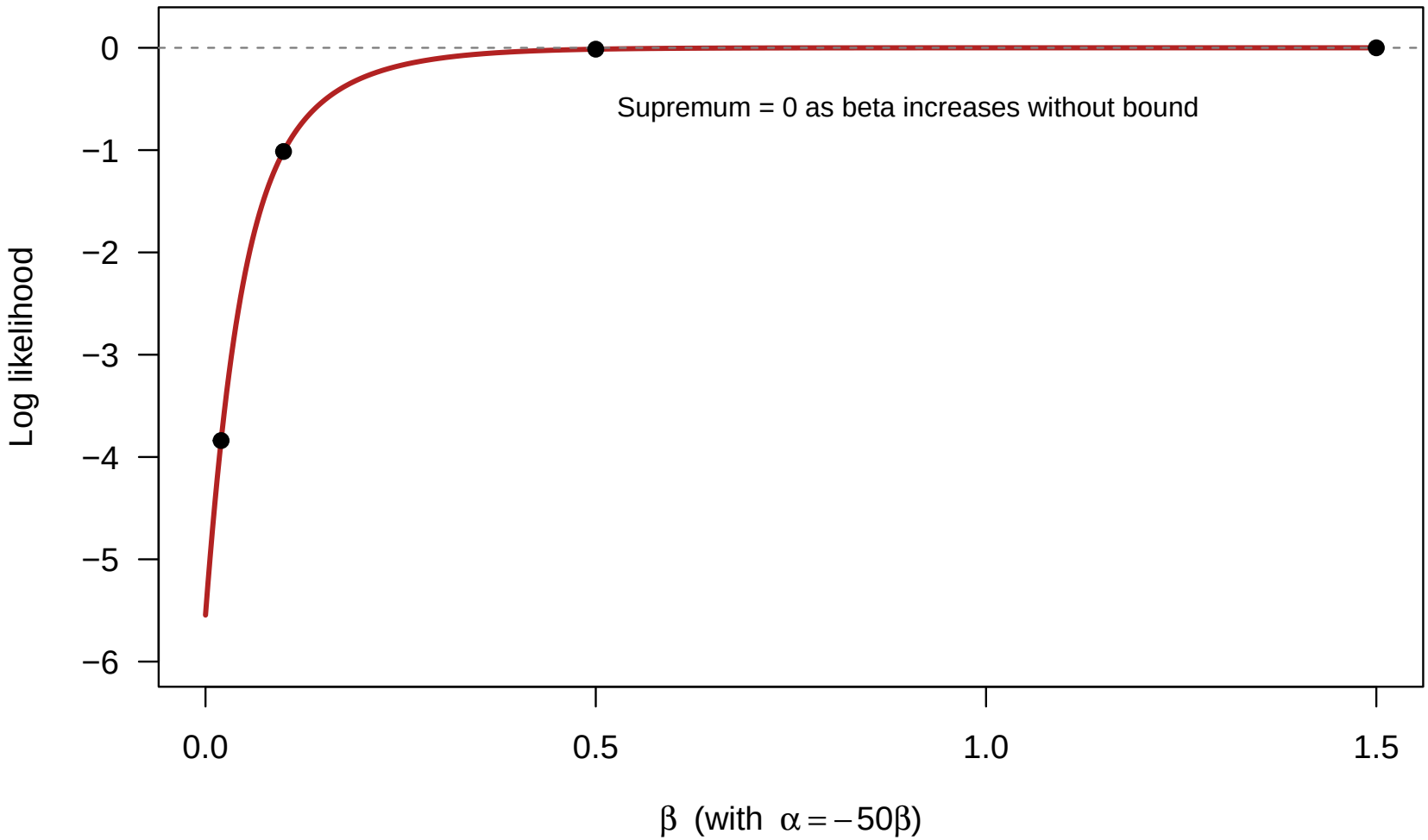
Predicted probabilities are now close to the observed zeros and ones.

Increasing the Slope: $\beta = 1.50$



The likelihood still improves, even though the curve looks almost like a step.

Likelihood Along This Sequence of Curves



The log likelihood approaches 0 as $\beta \to \infty$. No finite slope reaches the supremum.

An Infinite Maximum Likelihood Estimate

For these individual responses,

$$L(\alpha, \beta) = \prod_i \pi_i^{y_i}(1 - \pi_i)^{1-y_i}.$$

Along $\alpha = -50\beta$, increasing β makes every factor closer to 1.

- The likelihood approaches 1, but every finite logistic curve has probabilities strictly between 0 and 1.
- There is no finite maximizer: we describe the slope estimate as $\hat{\beta} = \infty$.
- The intercept also diverges along this path.

Fitting the Model in R

```
> fit <- glm(y ~ x, family = binomial, data = toy)
Warning: glm.fit: fitted probabilities numerically 0 or 1
    occurred
```

```
> summary(fit)
Call:
glm(formula = y ~ x, family = binomial, data = toy)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -118.16  296046.19       0        1
x                2.36    5805.94       0        1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 11.09035488895912  on 7  degrees of freedom
Residual deviance:  0.00000000021827  on 6  degrees of freedom
AIC: 4
Number of Fisher Scoring iterations: 25
```

The Wald Test

The reported test for the slope uses

$$z = \frac{\hat{\beta}}{\widehat{\text{SE}}(\hat{\beta})} \approx \frac{2.36}{5805.94} \approx 0.$$

- R stops at a finite numerical value, even though the ML estimate is infinite.
- The reported standard error is enormous.
- The resulting Wald p-value is approximately 1.

Does $p \approx 1$ make sense when x perfectly separates the outcomes?

A Likelihood-Ratio Test

Compare the intercept-only model to the model with x:

```
> fit_null <- glm(y ~ 1, family = binomial, data = toy)
> anova(fit_null, fit, test = "LRT")
Analysis of Deviance Table
Model 1: y ~ 1
Model 2: y ~ x
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         7       11.1
2         6        0.0  1     11.1  0.00087 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The deviance decrease is 11.09 on one degree of freedom, giving an approximate $p = 0.0009$. The likelihood-ratio test provides evidence of an association even though the Wald test fails. Its chi-square reference is a large-sample approximation.

Quasi-Complete Separation

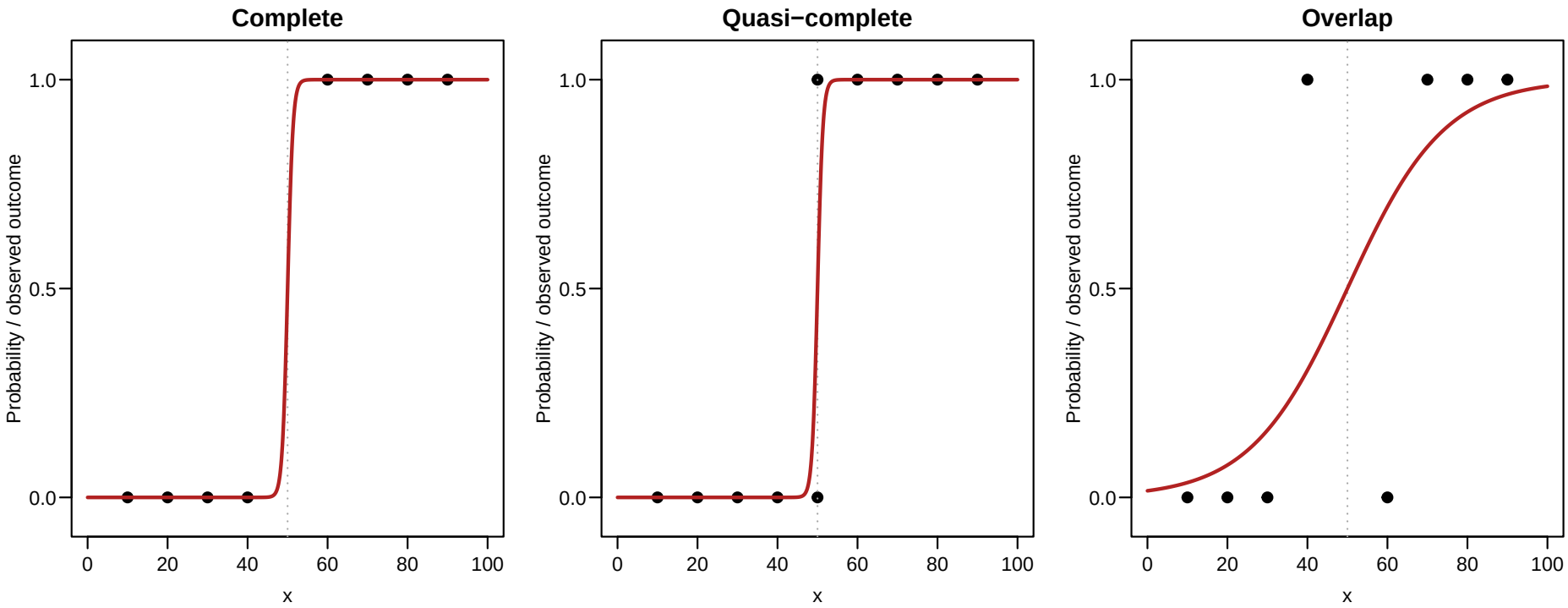
Now add two observations at the cutoff:

$$(x, y) = (50, 0) \qquad \text{and} \qquad (50, 1).$$

- Away from the cutoff, the outcomes still separate.
- Both outcomes occur at $x = 50$.
- Along $\alpha = -50\beta$, both cutoff observations keep fitted probability 0.5.
- The other observations are fit increasingly well as β increases.

The slope estimate is still infinite, but the limiting likelihood is below 1.

Separation and Overlap



For several predictors, a linear combination of the predictors may separate the outcomes even when no single predictor does so alone.

Separation with a Categorical Predictor

	$Y = 1$	$Y = 0$
$X = 1$	15	0
$X = 0$	12	8

$$\widehat{\text{OR}} = \frac{15 \times 8}{0 \times 12} = \infty, \qquad \hat{\beta} = \log \widehat{\text{OR}} = \infty.$$

- Both outcomes occur for $X = 0$, but only successes occur for $X = 1$.
- This is quasi-complete separation.

Endometrial Cancer Data

For 79 patients, HG is histology grade (0 = low, 1 = high). NV indicates neovasculation (0 = absent, 1 = present).

```
> Endo <- read.table("data/Endometrial.dat", header = TRUE)
> head(Endo)
  NV PI   EH HG
1  0 13 1.64  0
2  0 16 2.26  0
3  0  8 3.14  0
4  0 34 2.68  0
5  0 20 1.28  0
6  0  5 2.31  0
> xtabs(~ NV + HG, data = Endo)
   HG
NV   0  1
  0 49 17
  1  0 13
```

All 13 patients with NV = 1 have HG = 1. Which type of separation is this?

A Model for Histology Grade

The model also includes PI (pulsatility index) and EH (endometrium height):

$$\text{logit}\{P(\text{HG} = 1)\} = \beta_0 + \beta_{\text{NV}}\text{NV} + \beta_{\text{PI}}\text{PI} + \beta_{\text{EH}}\text{EH}.$$

- Increasing β_{NV} improves predictions for the NV $= 1$ patients.
- It leaves the predictions for NV $= 0$ patients unchanged when the other coefficients are held fixed.
- Hence $\hat{\beta}_{\text{NV}} = \infty$.

What do you expect the coefficient table to look like?

Endometrial Cancer Model in R

```
> endo_fit <- glm(HG ~ NV + PI + EH, family = binomial,
+                 data = Endo)
> summary(endo_fit)
Call:
glm(formula = HG ~ NV + PI + EH, family = binomial, data = Endo)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.3045     1.6373    2.63   0.0086 **
NV            18.1856  1715.7509    0.01   0.9915
PI            -0.0422     0.0443   -0.95   0.3413
EH            -2.9026     0.8456   -3.43   0.0006 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
AIC: 63.39
Number of Fisher Scoring iterations: 17
```

Testing the NV Effect

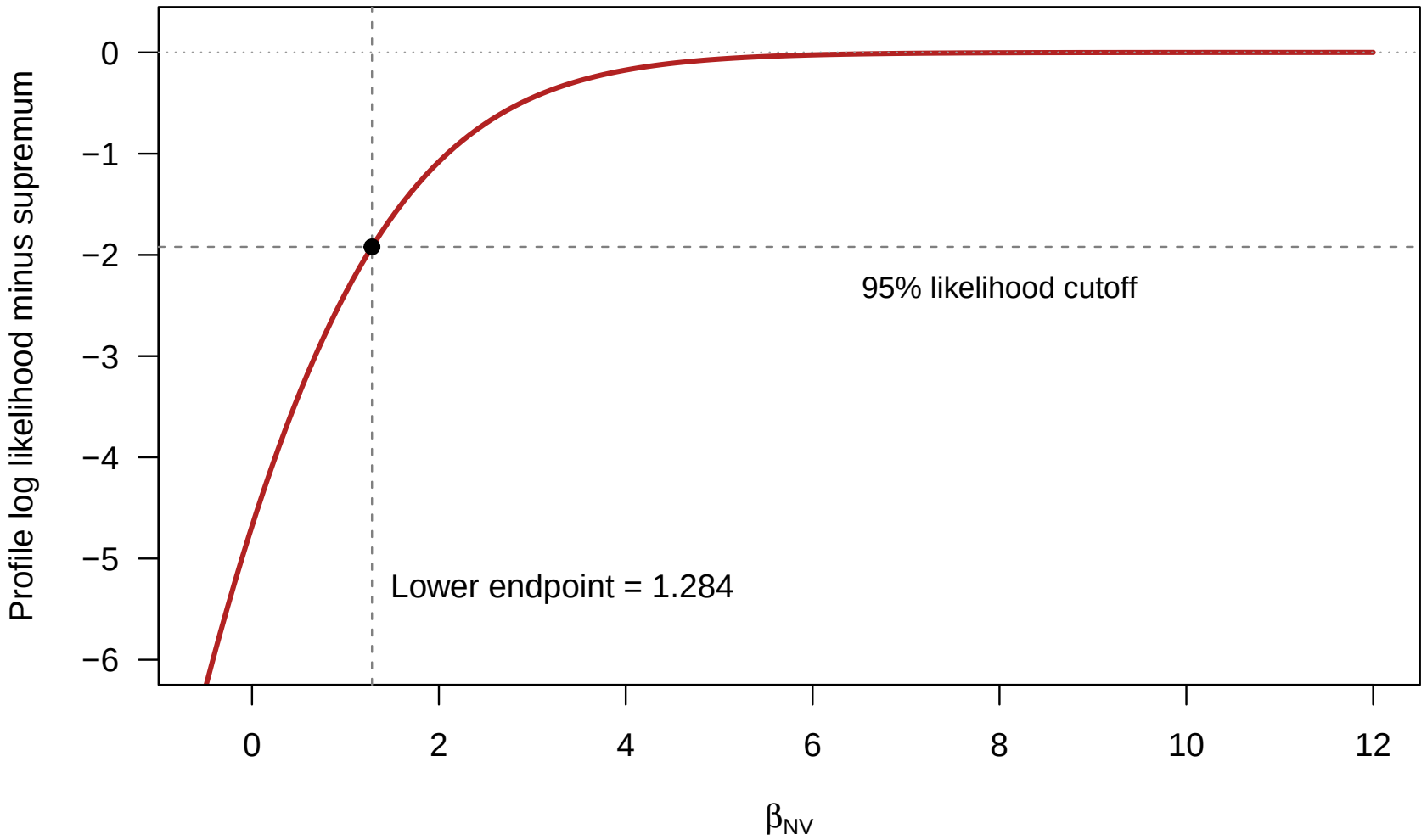
Compare models with and without NV, keeping PI and EH:

```
> endo_without_NV <- update(endo_fit, . ~ . - NV)
> anova(endo_without_NV, endo_fit, test = "LRT")
Analysis of Deviance Table
Model 1: HG ~ PI + EH
Model 2: HG ~ NV + PI + EH
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        76       64.8
2        75       55.4  1     9.36   0.0022 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wald p-value for NV was 0.9915. The likelihood-ratio p-value is about 0.0022. An infinite ML estimate does not prevent likelihood-based inference

about the effect.

Profile Likelihood for the NV Coefficient



For each fixed β_{NV}, refit the other coefficients. The profile approaches its supremum as $\beta_{\mathrm{NV}} \to \infty$.

A Profile Likelihood Confidence Interval

Values satisfying

$$2\{\ell_{\text{sup}} - \ell_{\text{profile}}(\beta_{\text{NV}})\} \leq \chi^2_{1,0.95} \approx 3.84$$

form an approximate 95% likelihood confidence interval.

```
> c(lower = ci_lower, upper = Inf)
 lower  upper
1.2841    Inf
> exp(c(lower = ci_lower, upper = Inf))
 lower  upper
3.6115    Inf
```

The interval for β_{NV} is about $(1.28, \infty)$, corresponding to an odds-ratio interval of about $(3.61, \infty)$, adjusting for PI and EH.

Recognizing Infinite Estimates

- Examine outcome counts within predictor categories and the arrangement of outcomes along numerical predictors.
- Watch for very large coefficients and standard errors, fitted probabilities close to 0 or 1, and fitting warnings.
- Numerical convergence can occur without a finite ML estimate.

A sample with only one outcome in a group does not establish that the population probability is exactly 0 or 1.

Inference and Finite Estimates

- Wald inference can be misleading when ML estimates are infinite.
- Likelihood-ratio tests and profile likelihood intervals can remain informative.
- Penalized likelihood and Bayesian methods can produce finite estimates under separation, using suitable penalties or priors.

Agresti introduces alternatives to ordinary maximum likelihood in Section 5.4.

Review

1. Why does increasing β keep improving the likelihood for the toy data?
2. Why do we also change α in the curve sequence?
3. How does quasi-complete separation differ from complete separation?
4. Why can a Wald test report $p \approx 1$ for a predictor that perfectly separates outcomes?
5. What can we conclude from a profile interval with an infinite upper endpoint?