Overview of Maximum Likelihood Estimation

Andrew Zieffler

Conceptual Overview

For which parameter values are the data most likely?

- ★ Given the **data and model**, what are the most likely value of the parameters
- ★ Maximum likelihood (ML) provides framework for answering this question
- ★ ML provides results which have attractive properties and are favorable for inference

Optimization Criterion

- ★ Optimization criterion provides **basis for computing estimates** of parameters
  - ✓ Optimization criterion for OLS is SSE
  - ✓ SSE is optimal at its minimum (least) value
  - ✓ Fixed effects estimates chosen to minimize SSE

- ★ Optimization criterion for ML is **likelihood function** or **deviance function**
  - ✓ Fixed effects (and other) estimates chosen to minimize likelihood function

$$\boxed{L\left(x_1, x_2, x_3, \ldots, x_n; \theta\right)}$$

Advantages of ML

- ML yields a **global fit index** for the model on top of the parameter estimates
  - ✓ Index is minimum of the deviance function
  - ✓ Can be used to compare models

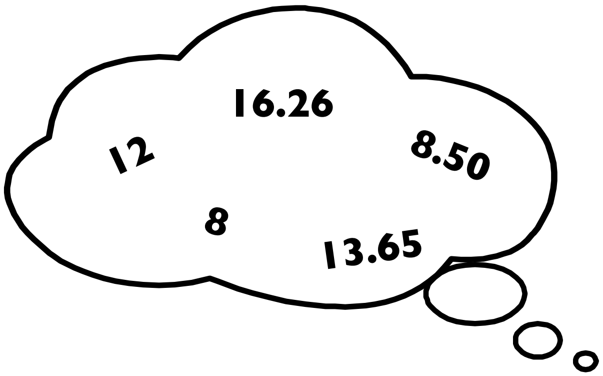
- Estimators produced under ML have **desirable large-sample (asymptotic) properties**
  - ✓ Consistent, asymptotically normally distributed
  - ✓ With small samples this may not hold

- ★ ML estimates are **always** approximate
  - ✓ Samples are finite
  - ✓ Often have missing data

- ★ Approximations imply researchers **should not** get too hung up on rigid rules of practice
  - ✓ Use of 2(SE) vs 1.96(SE)
  - ✓ Inflexible cutoffs (0.05) are not warranted

- ★ Notes only consider **regular problems**
  - ✓ ML solutions are possible
  - ✓ Assumption of sample data coming from hypothetical, infinitely sized population (repeated sampling scenario)

Example 1: Estimate the Mean



Given the data and the model, what is the most likely value of the mean?

- ★ Obviously data is given–it exists in data frame
- ★ What is the "given" model?
  - ✓ Traditional **marginal mean equation** along with the **enhancement** of an **explicit probability model**

- ★ Probability model specified before the analysis
  - ✓ Gives probability for **all possible samples** for the parameters in the model
  - ✓ Inferences made according to this model after sample data selected

- ★ Inferences based on likelihood (not probability)
  - ✓ Likelihood supplies **order of preference** (plausibility) among possible parameter values **given** the data and model
  - ✓ Denoted *Lik*

* ML begins with probability model for each and every observed score

$$y_i = \beta_0 + \epsilon_i$$

In the marginal mean model this is a LM with distributional assumption on the errors

where ε_i are normally distributed with $\mu_\varepsilon = 0$ and $\sigma^2_\varepsilon > 0$

$$\epsilon_i \sim \mathbb{N}(\mu_\epsilon, \sigma_\epsilon^2)$$

Express the LM in Terms of the Probability Model

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \cdot e^{\left[-\frac{(\epsilon_i - \mu_\epsilon)^2}{2\sigma_\epsilon^2}\right]}$$

where π = 3.14159...

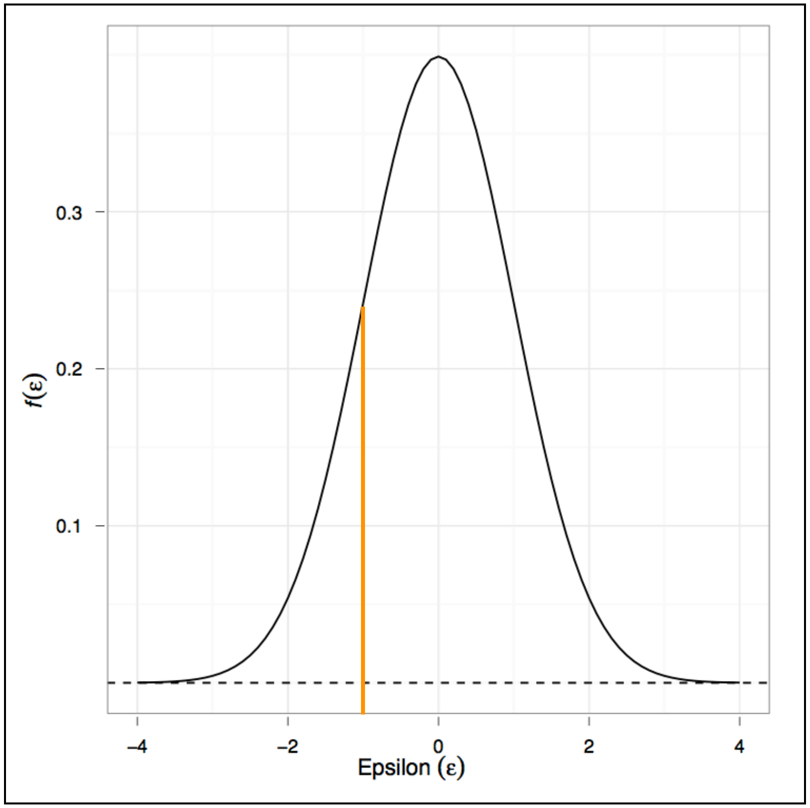
With the usual assumption that $\mu_\varepsilon = 0$ this simplifies to

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \cdot e^{\left[-\frac{\epsilon_i^2}{2\sigma_\epsilon^2}\right]}$$

Probability model associates different probability densities with different individual errors

Suppose $\mu_\varepsilon = 0$, $\sigma_\varepsilon^2 = 1$ and $\varepsilon_i = -1$ then the probability density is given by $f(\varepsilon_i)$

$$f(-1) = \frac{1}{\sqrt{2 \cdot \pi \cdot 1}} \cdot e^{\left[-\frac{(-1)^2}{2 \cdot 1}\right]}$$
$$= 0.2419707$$



The `dnorm()` function computes probability densities from a normal distribution in R.

```
> dnorm(-1, mean = 0, sd = sqrt(1) )

[1] 0.2419707
```

Note it takes the SD rather than the variance of the distribution

Likelihood Function

- ★ Basis for all inference in ML estimation
- ★ Consists of probability function for each and every potential observation
- ★ For LM probability function defined for each ε_i
  - **Assumption of independence** allows us to multiply each probability function to obtain **joint density**

$$Lik = [f(\epsilon_1)] \cdot [f(\epsilon_2)] \cdot [f(\epsilon_3)] \cdot \ldots \cdot [f(\epsilon_N)]$$

$$Lik = \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\epsilon_1^2}{2\sigma_\epsilon^2}\right) \cdot \cancel{\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}} exp\left(-\frac{\epsilon_2^2}{2\sigma_\epsilon^2}\right) \cdot$$

$$\cancel{\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}} exp\left(-\frac{\epsilon_3^2}{2\sigma_\epsilon^2}\right) \cdot \ldots \cdot \cancel{\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}} exp\left(-\frac{\epsilon_N^2}{2\sigma_\epsilon^2}\right)$$

★ Substitute the function for the normal probability density in for each term in the likelihood function. Note that *exp*(a) = e*ᵃ*.

$$Lik = \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\epsilon_1^2}{2\sigma_\epsilon^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(-\frac{\epsilon_2^2}{2\sigma_\epsilon^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(-\frac{\epsilon_3^2}{2\sigma_\epsilon^2}\right) \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} exp\left(-\frac{\epsilon_N^2}{2\sigma_\epsilon^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\sum_{i=1}^{N} \epsilon_i^2}{2\sigma_\epsilon^2}\right)$$

Likelihood function

Easier to take natural logarithm of both sides of equation

$$\ln(Lik) = \ln\left(\left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^N exp\left(-\frac{\sum_{i=1}^N \epsilon_i^2}{2\sigma_\epsilon^2}\right)\right)$$

Do the math to convince yourself it reduces to this

$$= -\frac{N}{2}\cdot\ln\left(2\pi\sigma_\epsilon^2\right) - \frac{1}{2\sigma_\epsilon^2}\cdot\sum_{i=1}^N \epsilon_i^2$$

Log-Likelihood function

Can multiply both sides by –2

$$-2\ln(Lik) = N \cdot \ln\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} \epsilon_i^2$$

-2(log-likelihood)
is called the **deviance**

We can write ε_i as $y_i - \beta_0$

$$\text{deviance} = N \cdot \ln\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} (y_i - \beta_0)^2$$

The model is **embedded in the specified probability function**, by way of the deviance function. Now the deviance function can be minimized.

Intuitive Idea of Minimization

- ⋆ Goal is to minimize the value of the deviance
- ⋆ Assume that the parameters of the error variance is known, and that the **only unknown** parameter to be estimated is the intercept

$$\sigma_\varepsilon^2 = 15$$

- ⋆ Given the model (with known parameter values) and the data, object is to choose "best" value for the slope
- ⋆ "Best" value here means that after substituting values in to the deviance function equation, the smallest deviance possible has been obtained
- ⋆ In ML theory, this is the estimate (value) of β_0 that is **maximally likely** given the data and the model–hence maximum likelihood

★ Everything in the deviance function is known, except for β₀

$$\text{deviance} = \boxed{N} \cdot \ln\left(2\pi \boxed{\sigma_\epsilon^2}\right) + \frac{1}{\boxed{\sigma_\epsilon^2}} \cdot \sum \left(\boxed{\text{wage}_i} - \beta_0\right)^2$$

5

15

i	wage
1	12.00
2	8.00
3	16.26
4	13.65
5	8.50

★ Substituting in the values

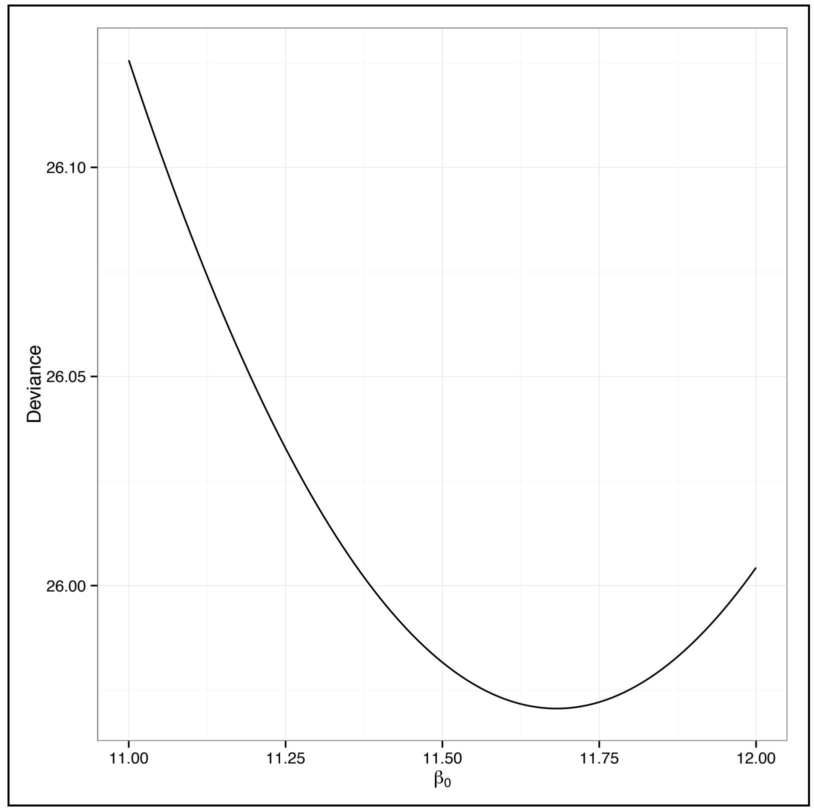
i	wage
1	12.00
2	8.00
3	16.26
4	13.65
5	8.50

$$\text{deviance} = 5 \cdot \ln(2 \cdot \pi \cdot 15) + \frac{1}{15} \cdot \Big[(12 - \beta_0)^2 + (8 - \beta_0)^2 + (16.26 - \beta_0)^2 + (13.65 - \beta_0)^2 + (8.5 - \beta_0)^2 \Big]$$

Given we want the Smallest Deviance what Value should we Choose for β_0?

- Pick several candidate values for β_0 and substitute them into the equation. Solve for deviance. Choose the one with the smallest deviance value.
- For example, consider candidate values for β_0 between 11 and 12

β_0	deviance
11.00	26.12564
11.01	26.12113
11.02	26.11668
11.03	26.11230
⋮	⋮



Carry Out a Grid Search in R

```
## Create the function
> dev <- function(b0) {
     5 * log(2 * pi * 15) + 1/15 * ( (12 - b0)^2 +
     (8 - b0)^2 + (16.26 - b0)^2 +
     (13.65 - b0)^2 + (8.5 - b0)^2 )
     }

## Try function
> dev(0)

[1] 71.46031
```

Carry Out a Grid Search in R

```
## Generate values for b0
> new <- data.frame(
    b0 = seq(from = 11, to = 12, by = 0.01)
    )

## Generate the deviance values and store the results
> library(plyr)
> new <- mdply(new, dev)


## Change the name of the second column
> names(new)[2] <- "deviance"
```

Carry Out a Grid Search in R

```r
## Plot deviance vs. b0
> ggplot(data = new, aes(x = b0, y = deviance)) +
    geom_line() +
    theme_bw() +
    xlab(expression(beta[0])) +
    ylab("Deviance")

## Arrange from smallest to largest deviance
> head(arrange(new, deviance))
```

Use Calculus

- ★ The derivative of the deviance function can be analytically computed and set equal to zero. Solving for β_0 will give the value that minimizes the deviance
- ★ Avoids exhaustive search
- ★ In more complex models (e.g., LMER) estimation is not so straightforward.
  - ✓ Exhaustive search methods are required (numerical analysis)
- ★ Once the deviance function becomes more complex, the multidimensional form of the tangent line most be discovered

$$\frac{\partial}{\partial \beta_0} \quad -2n \ln\left(\frac{1}{\sqrt{2\pi}}\right) + (x_1 - \beta_0)^2 + (x_2 - \beta_0)^2 + \ldots + (x_n - \beta_0)^2$$

Do you remember the chain rule?

$$0 - 2\,(x_1 - \beta_0) - 2\,(x_2 - \beta_0) - \ldots - 2\,(x_n - \beta_0)$$

Find the root of the derivative.

Set = 0 $\quad \boxed{0 = -2\,[(x_1 - \beta_0) - (x_2 - \beta_0) - \ldots - (x_n - \beta_0)]}$

Divide by -2 $\quad 0 = (x_1 - \beta_0) - (x_2 - \beta_0) - \ldots - (x_n - \beta_0)$

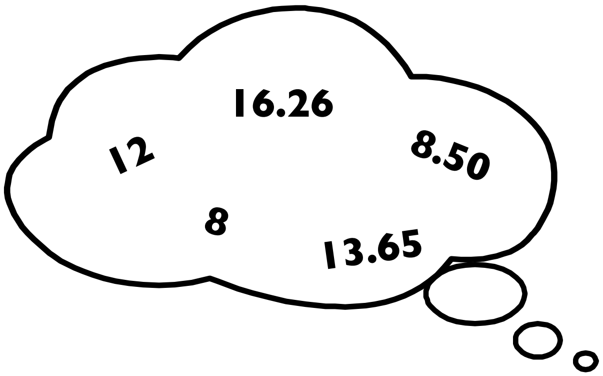
Combine 'like' terms $\quad 0 = (x_1 + x_2 + \ldots + x_n) - n\beta_0$

Add *n*β_0 to both sides $\quad n\beta_0 = (x_1 + x_2 + \ldots + x_n)$

Divide both sides by *n* $\quad \boxed{\beta_0 = \frac{\sum x_i}{n}}$

Sample mean is the estimate that minimizes the deviance.

Example 2: Estimate the Mean and Std. Dev.



Given the data and the model, what is the most likely value of the mean and standard deviation?

★ Model and data are still the same, but now we don't assume a specific value for σ

$$\text{deviance} = N \cdot \ln\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum \left(y_i - \beta_0\right)^2$$

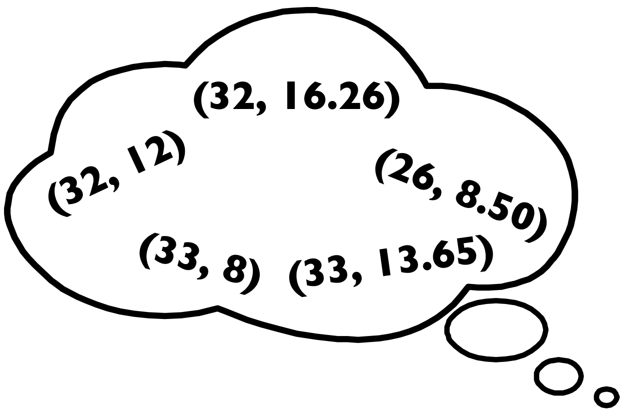
Carry Out a Grid Search in R

```
## Create the function
> dev <- function(b0, s) {
     5 * log(2 * pi * s^2) + 1/s^2 * ( (12 - b0)^2 +
     (8 - b0)^2 + (16.26 - b0)^2 +
     (13.65 - b0)^2 + (8.5 - b0)^2 )
     }

## Try function
> dev(b0 = 0, s = 1)

[1] 740.1495
```

Carry Out a Grid Search in R

```
## Generate independent search grids for b0 and b1
> b0 = seq(from = 11, to = 12, by = 0.01)
> s = seq(from = 12, to = 13, by = 0.01)

## Create combined search grid
> new <- expand.grid(b0 = b0, s = s)

## Generate the deviance values and store the results
> new <- mdply(new, dev)


## Change the name of the second column
> names(new)[3] <- "deviance"

## Arrange from smallest to largest deviance
> head(arrange(new, deviance))
```

Example 3: Regression (Estimate Intercept Only)



Given the data and the model, what is the most likely value of the unknown slope parameter

- Data now include *x* and *y* values
- Model is now a traditional regression model (conditional mean model)

$$y_i = \beta_0 + \beta_1(\text{age}_i) + \epsilon_i \qquad \epsilon_i \sim \mathbb{N}(\mu_\epsilon, \sigma^2_\epsilon)$$

Embed the LM in the Probability Model

We can write ε_i as $y_i - \beta_0 - \beta_1(x_i)$

$$\text{deviance} = N \cdot \ln\left(2\pi\sigma_\epsilon^2\right) + \frac{1}{\sigma_\epsilon^2} \cdot \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 \cdot x_i)^2$$

The model is **embedded in the specified probability function**, by way of the deviance function. Again, the deviance function can be minimized.

- ★ Assume that the parameters of the error variance and the intercept are known, and that the **only unknown** parameter to be estimated is the slope

$\beta_0 = -4 \qquad \sigma_\varepsilon^2 = 15$

★ Everything in the deviance function is known, except for β1

$$deviance = \boxed{N} \cdot \ln\left(2\pi \boxed{\sigma_\epsilon^2}\right) + \frac{1}{\boxed{\sigma_\epsilon^2}} \cdot \sum_{i=1}^{N} \left(\boxed{\text{wage}_i} - \boxed{\beta_0} - \beta_1 \cdot \boxed{\text{age}_i}\right)^2$$

5

15

–4

i	wage	age
1	12.00	32
2	8.00	33
3	16.26	32
4	13.65	33
5	8.50	26

★ Substituting in the values

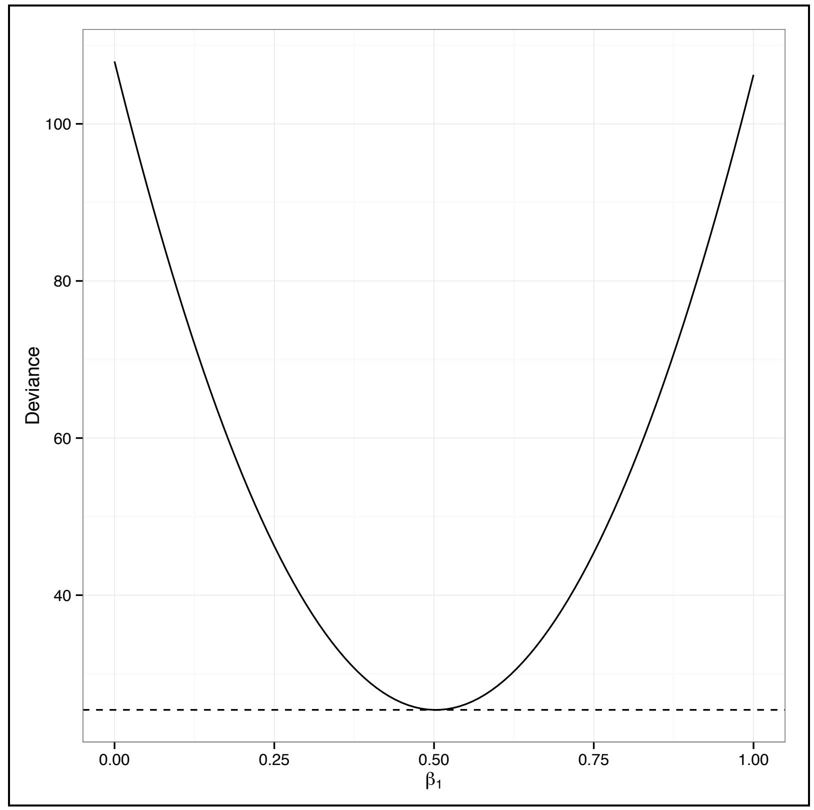
i	wage	age
1	12.00	32
2	8.00	33
3	16.26	32
4	13.65	33
5	8.50	26

$$\begin{aligned}\text{deviance} = 5 \cdot \ln\left(2 \cdot \pi \cdot 15\right) + \frac{1}{15} \cdot \Big[&(12 + 4 - \beta_1 \cdot 32)^2 \\ &+ (8 + 4 - \beta_1 \cdot 33)^2 + (16.26 + 4 - \beta_1 \cdot 32)^2 \\ &+ (13.65 + 4 - \beta_1 \cdot 33)^2 + (8.5 + 4 - \beta_1 \cdot 26)^2 \Big]\end{aligned}$$

Given we want the Smallest Deviance what Value should we Choose for β_1?

- Pick several candidate values for and substitute them into the equation. Solve for deviance. Choose the one with the smallest deviance value.
- For example, consider candidate values for β_1 between 0 and 1

β_1	**deviance**
0.0	107.94564
0.1	104.69330
0.2	101.50631
0.3	98.38468
⋮	⋮



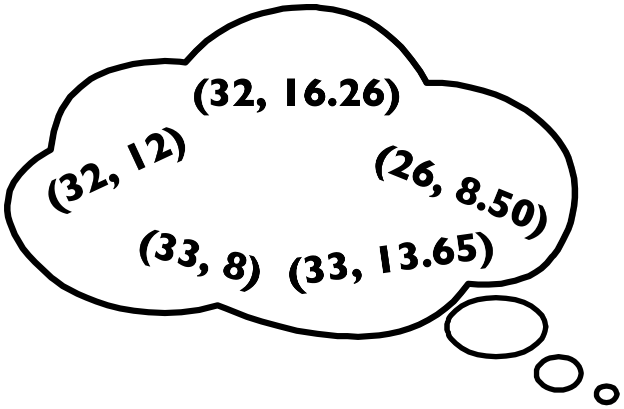
Carry Out a Grid Search in R

```
## Create the function
> dev <- function(b1) {
     5 * log(2 * pi * 15) + 1/15 * ( (12 + 4 - b1 * 32)^2 +
     (8 + 4 - b1 * 33)^2 + (16.26 + 4 - b1 * 32)^2 +
     (13.65 + 4 - b1 * 33)^2 + (8.5 + 4 - b1 * 26)^2 )
     }

## Try function
> dev(0)

[1] 107.9456
```

Carry Out a Grid Search in R

```
## Generate values for b1
> new <- data.frame(
    b1 = seq(from = 0, to = 1, by = 0.01)
    )

## Generate the deviance values and store the results
> library(plyr)
> new <- mdply(new, dev)


## Change the name of the second column
> names(new)[2] <- "deviance"
```

Example 4: Regression (Estimate Intercept & Slope)



Given the data and the model, what is the most likely value of the unknown slope parameter and unknown intercept parameter?

- ★ Data:
- ★ Model:

Your turn!

- ★ Assume that the parameters of the **error variance is known**, and that both the **intercept and the slope are unknown** parameters to be estimated

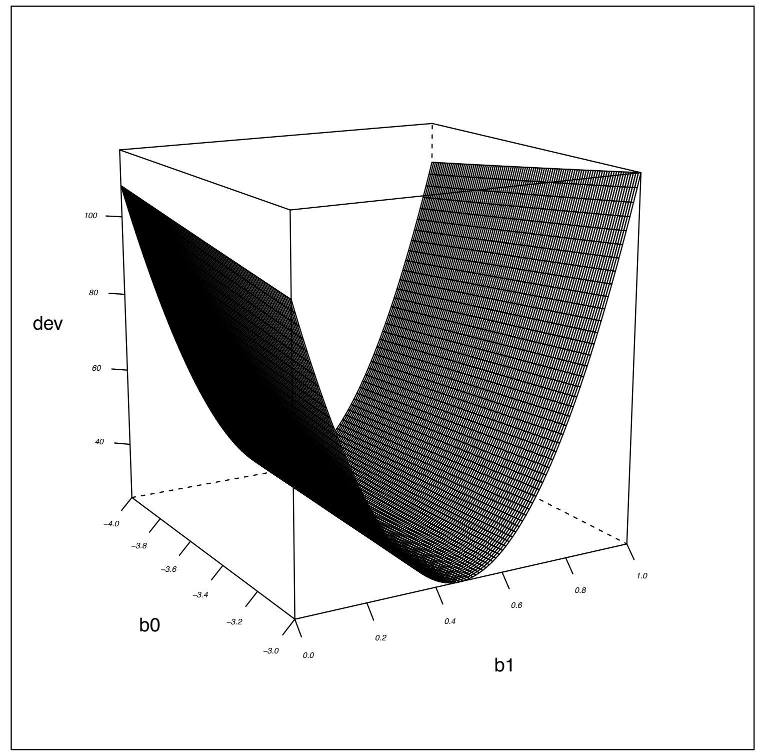
$$\sigma_\epsilon^2 = 49$$

- ★ Substituting into the deviance

$$\begin{aligned}\text{deviance} = 5 \cdot \ln\left(2 \cdot \pi \cdot 15\right) \frac{1}{15} \cdot \Big[&\left(12 - \beta_0 - \beta_1 \cdot 32\right)^2 \\ &+ \left(8 - \beta_0 - \beta_1 \cdot 33\right)^2 + \left(16.26 - \beta_0 - \beta_1 \cdot 32\right)^2 \\ &+ \left(13.65 - \beta_0 - \beta_1 \cdot 33\right)^2 + \left(8.5 - \beta_0 - \beta_1 \cdot 26\right)^2\Big]\end{aligned}$$

- ★ With two unknowns, the deviance function is a plane in three-dimensional space ($x = \beta_0$, $y = \beta_1$, z = deviance)
- ★ Candidate values for β_0 and β_1 are considered simultaneously
- ★ Search grid includes (–4, –3) for β_0 and (0, 1) for β_1 in increments of 0.01

β_0	β_1	deviance
–4.00	0	107.9456
–3.99	0	107.8411
–3.98	0	107.7367
–3.97	0	107.6323
⋮	⋮	⋮
–3.92	0.50	25.39207
⋮	⋮	⋮



Minimum deviance occurs at floor of the plot (horizontal plane on which the graph rests)

Carry Out a Grid Search in R

```
## Create the function
> dev <- function(b0, b1) {
     5 * log(2 * pi * 15) + 1/15 * ( (12 - b0 - b1 * 32)^2 +
     (8 - b0 - b1 * 33)^2 + (16.26 - b0 - b1 * 32)^2 +
     (13.65 - b0 - b1 * 33)^2 + (8.5 - b0 - b1 * 26)^2 )
     }

## Try function
> dev(b0 = 0, b1 = 0)

[1] 71.46031
```

Carry Out a Grid Search in R

```
## Generate values for b1
> new2 <- expand.grid(
    b0 = seq(from = -4, to = -3, by = 0.01),
    b1 = seq(from = 0, to = 1, by = 0.01)
    )

## Generate the deviance values and store the results
> new2 <- mdply(new2, dev)


## Change the name of the third column
> names(new)[2] <- "deviance"

## Order the data frame by deviance
> arrange(new2, dev)
```

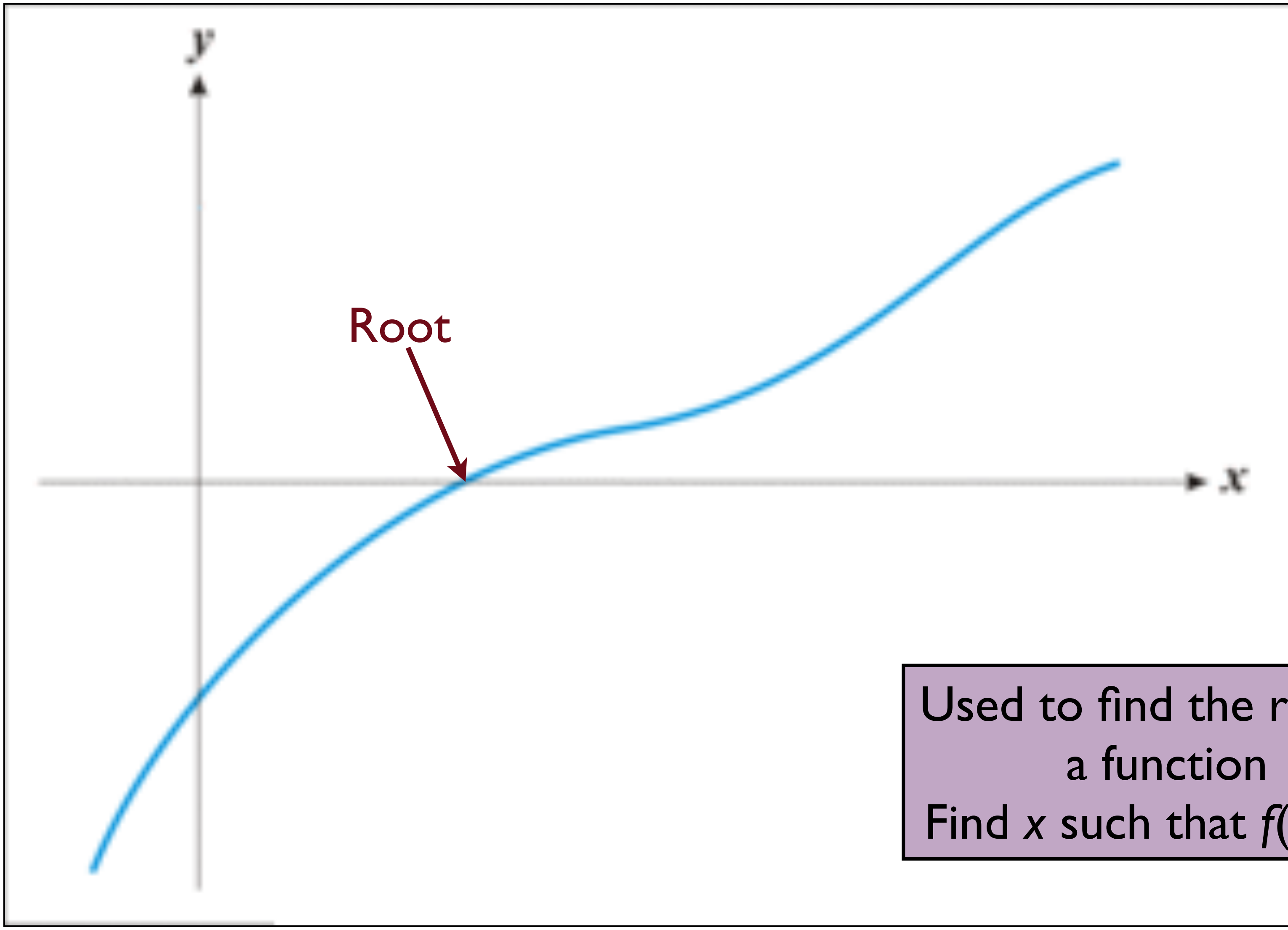
★ If the error variance is also unknown, **all three parameters are simultaneously estimated** and the deviance is again minimized.

```
## Fit linear model
> lm.1 <- lm(wage ~ age, data = wg)

## Compute log-likelihood
> logLik(lm.1)

'log Lik.' -12.28932 (df=3)

## Compute deviance
> -2 * logLik(lm.1)[1]

24.57864
```

Exhaustive Search vs. Numerical Methods

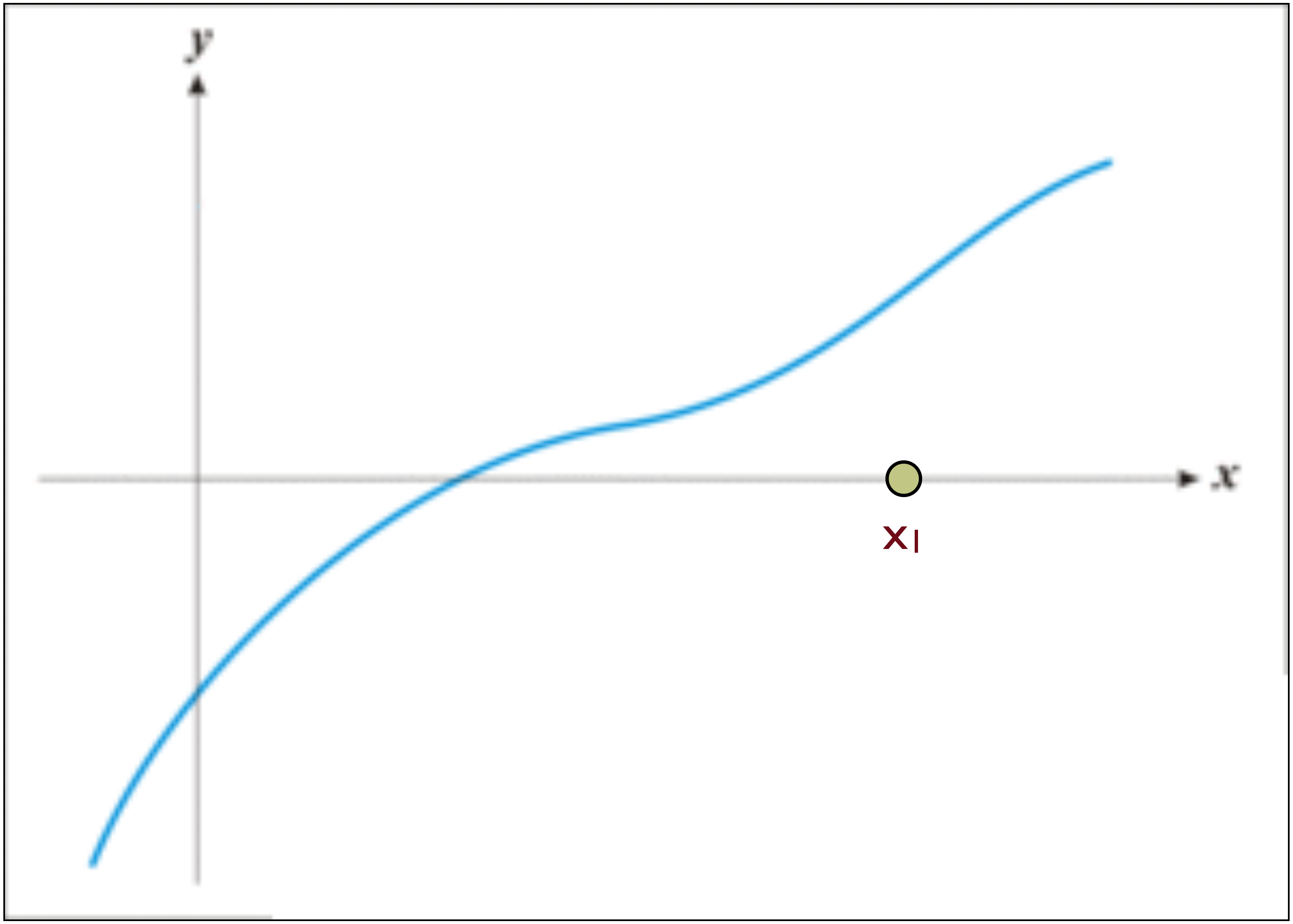
- ★ The exhaustive searches carried out in the examples were convenient
  - ✓ Ranges were used where the ML estimates were known to reside
  - ✓ In practice, these are not known
  - ✓ Increments are usually finer than 0.01 in practice
- ★ Computing time and memory becomes a valuable resource
- ★ Numerical methods, based on calculus are usually employed
  - ✓ Newton-Raphson method is most common algorithm
  - ✓ Deviance function assumed to be smooth and continuous with only one minimum (regularity assumption)
- ★ Functions such as `lmer()` generally combine both methods
  - ✓ Numerical methods are used to get in the neighborhood of the minimum deviance
  - ✓ Once more limited search space has been defined an iterative method can hone in on the minimum deviance until it is "good enough"

Newton-Raphson Method

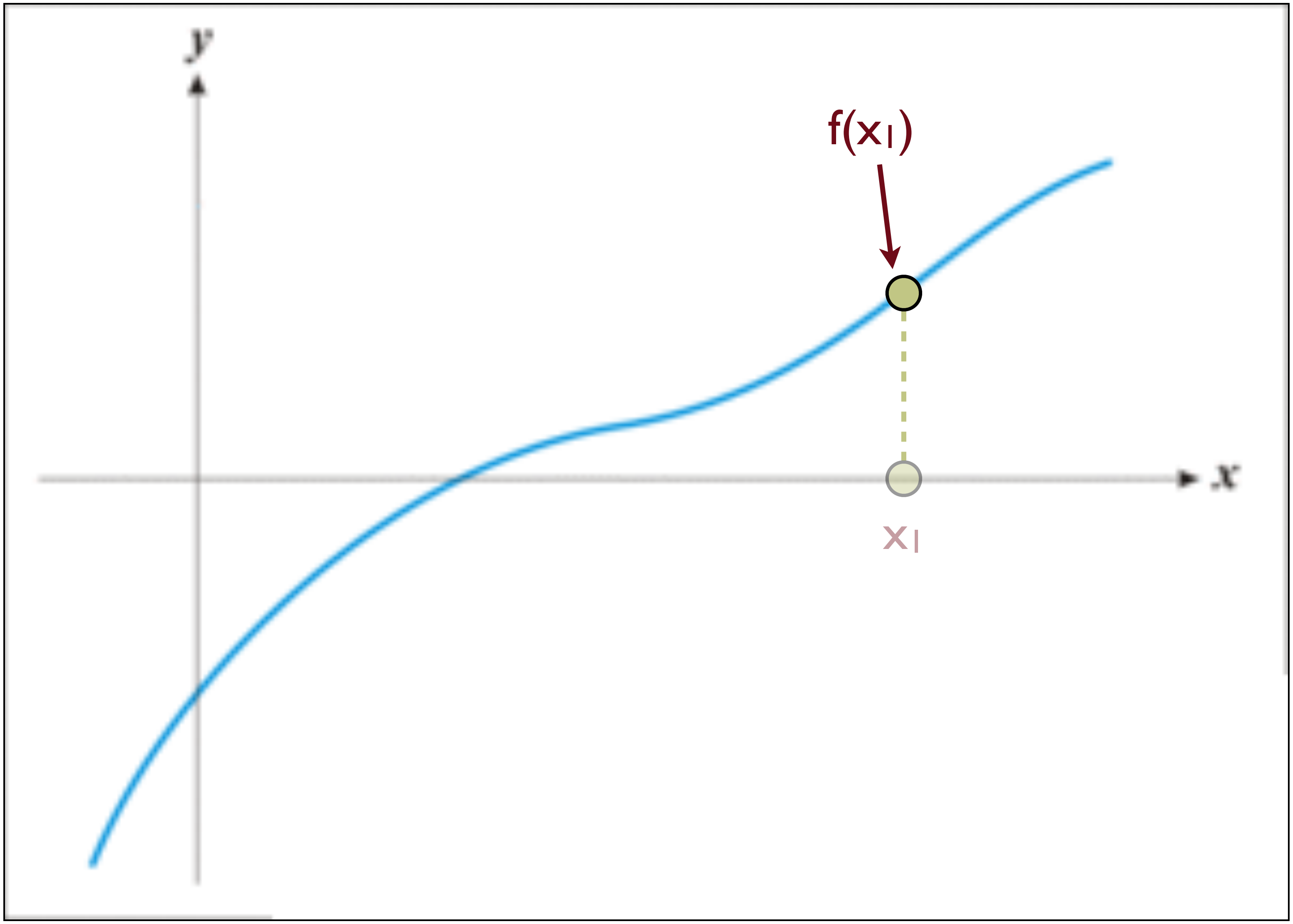


Used to find the root of a function
Find x such that $f(x) = 0$

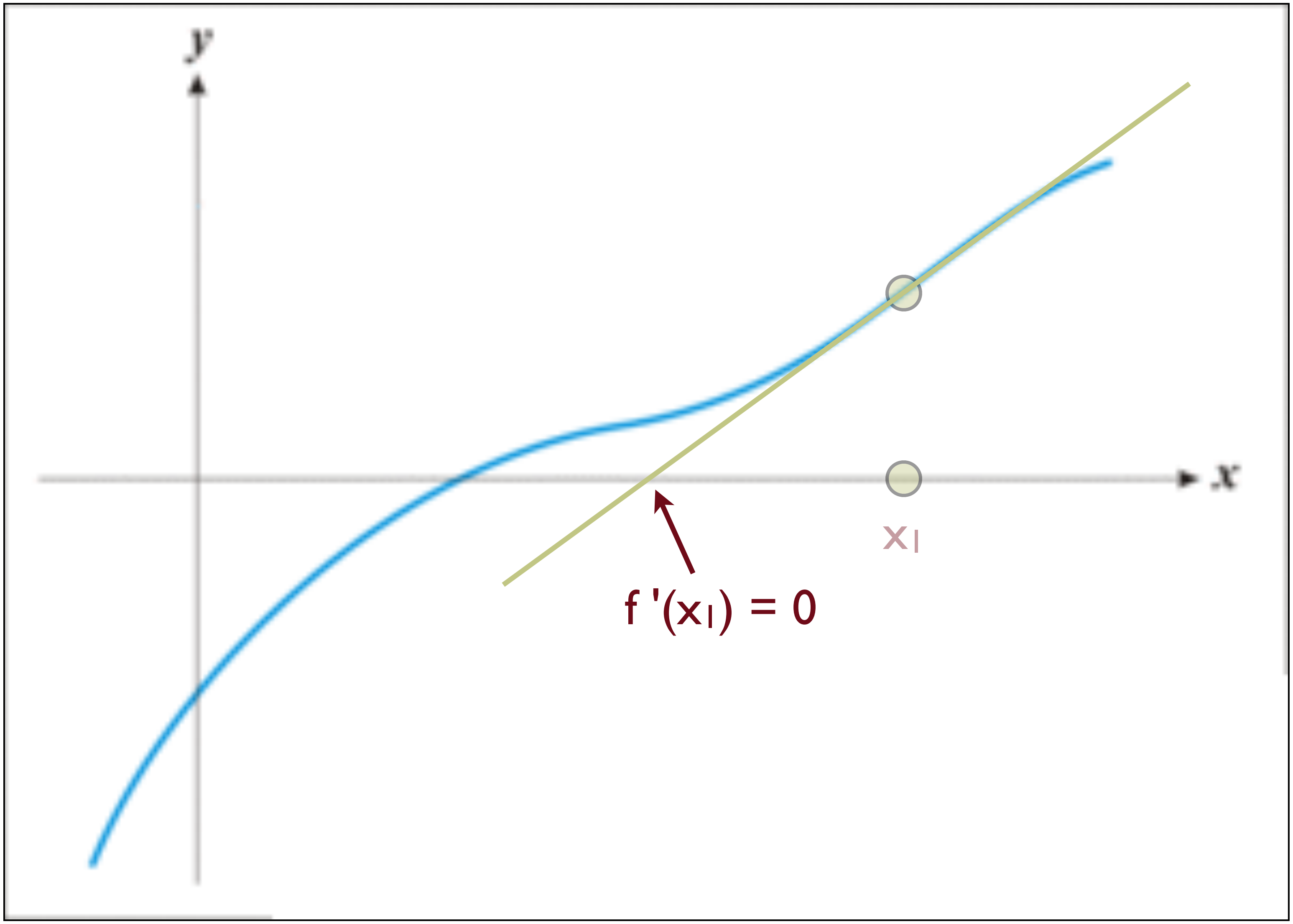
Choose some value x



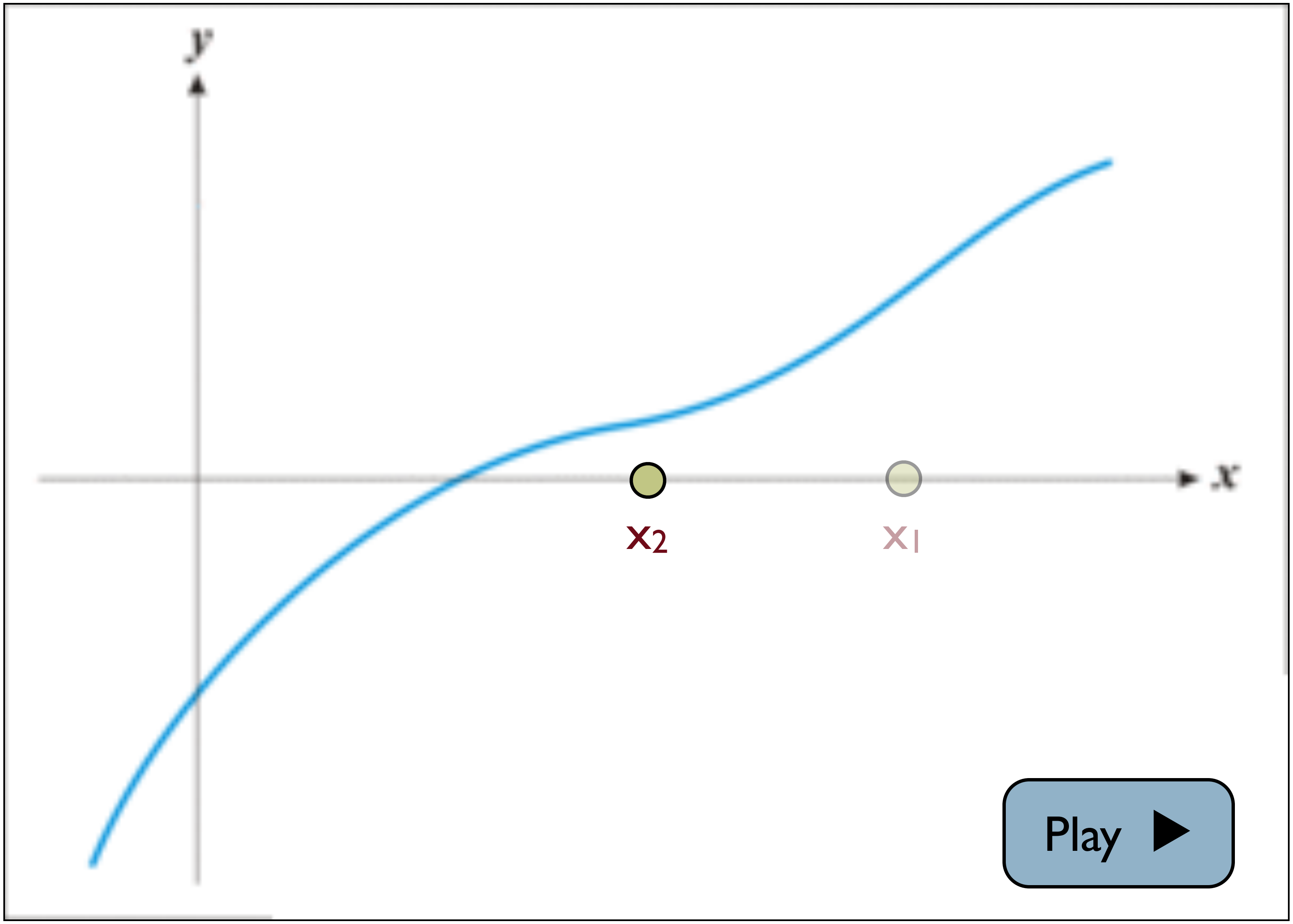
Compute f(x)

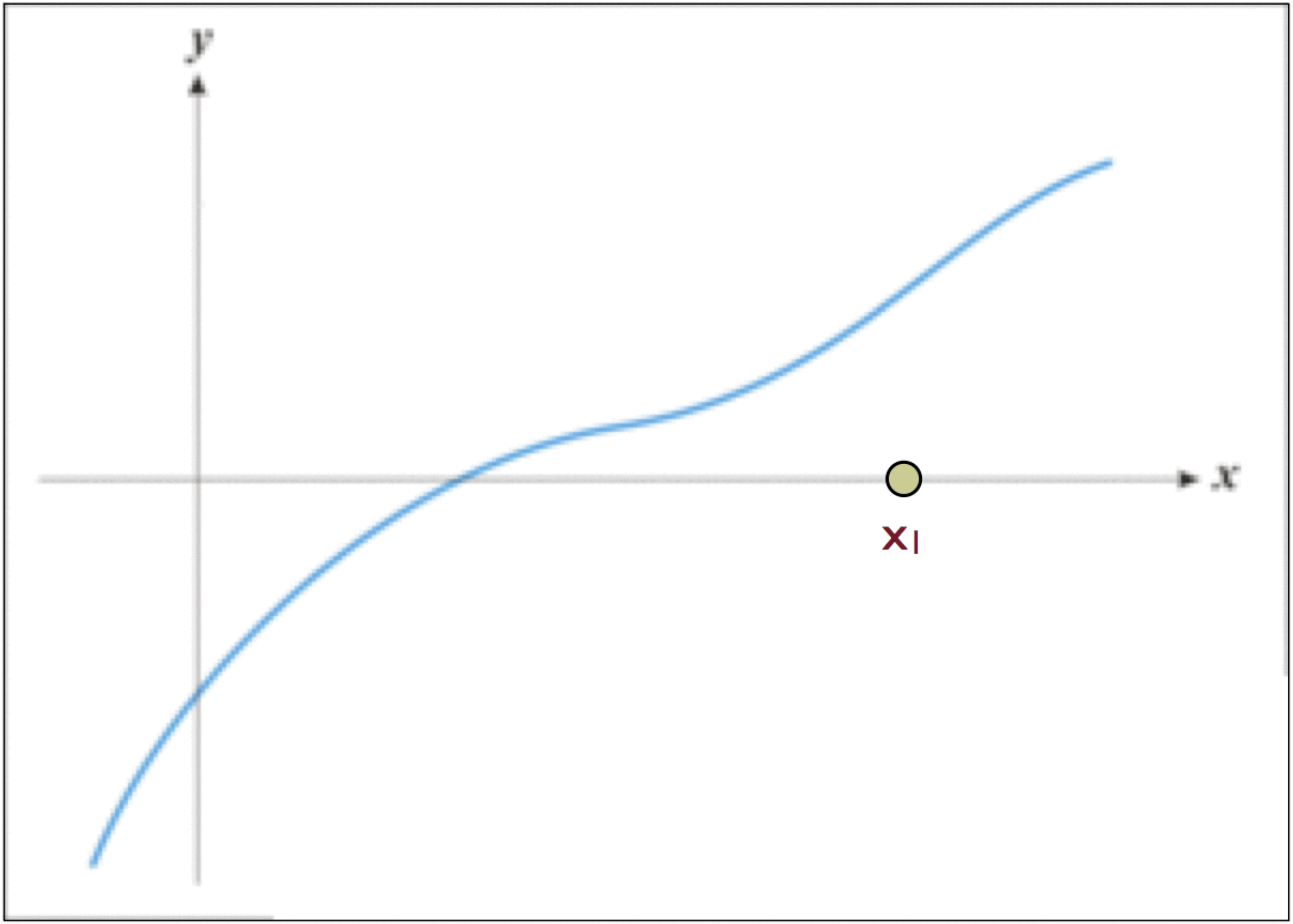


Solve $f'(x) = 0$



This will become the next value for *x*





y
x
x₁

Restricted Maximum Likelihood

★ It can be shown that the ML estimator for the **error variance** is

$$\hat{\sigma}^2_{\epsilon,ML} = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{N}$$

★ This is a **biased estimator** of the population variance

✓ Underestimates the population value in repeated sampling

★ Square root of error variance is **residual standard error**

✓ Printed in `summary()` output of `lm()`

- ★ To correct the bias, a different denominator is used
  - ✓ Called **restricted maximum likelihood (REML)** estimator

$$\hat{\sigma}^2_{\epsilon,REML} = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{N - p - 1}$$

- ★ Estimation without correction for bias is called *full ML* or just ML

where *p* is the number of predictors

- ★ When sample size is large, ML and REML results are similar
- ✓ As sample size increases without bound results converge
- ✓ Most ML results used for inference based on large-sample theory
- ★ REML is correction for variances
- ✓ Nature of correction depends on fixed effects structure of model
- ✓ Nested models can therefore differ not only in fixed effects, but also in correction terms
- ✓ REML should not be used for comparing models

- ★ Estimation **without correction** for bias is called **full ML** or just ML

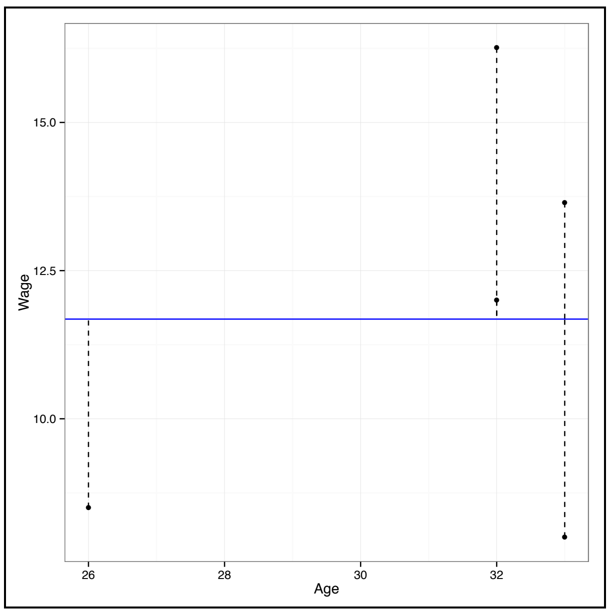
- ★ When sample size is large, ML and REML results are similar
  - ✓ As sample size increases without bound results converge
  - ✓ Most ML results used for inference based on large-sample theory

- ★ REML is correction for variances
  - ✓ Nature of correction depends on fixed effects structure of model
  - ✓ Nested models can therefore differ not only in fixed effects, but also in correction terms
  - ✓ REML should not be used for comparing models

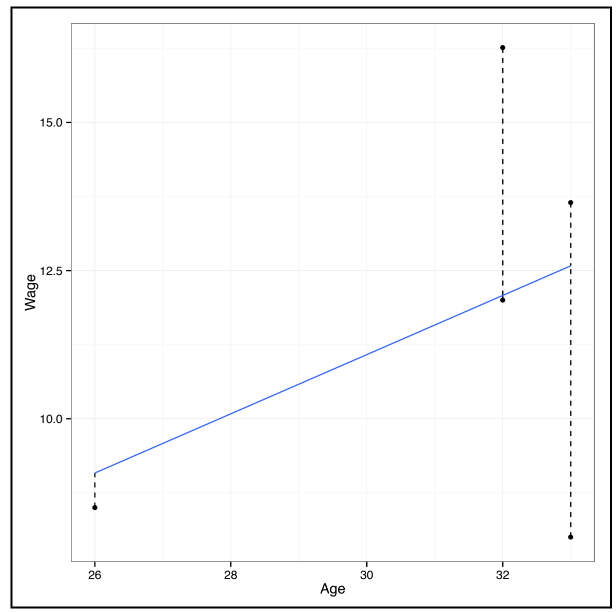
Comparing Models

- ★ Deviance is solid foundation for comparing models
  - ✓ Basis for AIC and LRT

$$y_i = \beta_0 + \epsilon_i$$



$$y_i = \beta_0 + \beta_1(\text{age}) + \epsilon_i$$



- ★ Residuals are larger for the intercept-only model
- ★ Slope model fits better

★ Using the ML estimator

$$\text{deviance} = N \cdot \ln\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + \frac{1}{\boxed{\hat{\sigma}^2_{\epsilon,ML}}} \cdot \sum \epsilon_i^2$$

$$\hat{\sigma}^2_{\epsilon,ML} = \boxed{\frac{\sum \hat{\epsilon}_i^2}{N}}$$

$$= N \cdot \ln\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + \frac{N}{\boxed{\cancel{\sum \hat{\epsilon}_i^2}}} \cdot \boxed{\cancel{\sum \epsilon_i^2}}$$

$$= N \cdot \ln\left(2 \cdot \pi \cdot \hat{\sigma}^2_{\epsilon,ML}\right) + N$$

$$\text{deviance} = N\left[\ln\left(2\cdot\pi\cdot\hat{\sigma}^2_{\epsilon,ML}\right)+1\right]$$

$$= N\left[\ln\left(2\cdot\pi\cdot N^{-1}\boxed{\sum\hat{\epsilon}_i^2}\right)+1\right]$$

Sum of squared residuals (SSR)

- ★ SSR is only term in deviance that is not a constant
- ★ For a fixed sample size (*N*), the **SSR is the only influence** on the size of the deviance
- ★ Minimizing the deviance is equivalent to minimizing the SSR
  - ✓ OLS is a special case of ML (but only for LM)

```
> lm.0 <- lm(wage ~ 1, data = wg)     ## Fit intercept-only model
> -2 * logLik(lm.0)[1]                ## Compute deviance
[1] 25.5618

> lm.1 <- lm(wage ~ age, data = wg)  ## Fit slope model
> -2 * logLik(lm.1)[1]                ## Compute deviance
[1] 24.57864
```

- ★ Deviance is **smaller for slope model**, model fits better to data
- ★ Any model with more parameters will fit the data better
- ★ SSR and deviance will **always decrease as more predictors added** to the model
- ★ Worthless predictors will still decrease the SSR and deviance

Information Criteria

★ To guard against adding potentially worthless predictors, the deviance is "penalized"

★ Penalized indexes are known generally as *information criteria* (IC)

$$\boxed{\text{IC} = \textit{deviance} + \textit{penalty}}$$

★ Smaller IC values indicate better fit

✓ Penalty term is always non-negative

✓ Increases as parameters are added to the model

- ★ Two popular IC are AIC and BIC
  - ✓ Akaike information criteria (Akaike, 1973, 1974, 1981)
  - ✓ Schwartz's Bayesian information criteria (Schwartz, 1978)

$$\text{AIC} = \text{deviance} + 2 \cdot K$$

$$\text{BIC} = \text{deviance} + K \ln(N)$$

K is the number of estimated parameters in the model

- ★ Debate about which IC should be used in practice
  - ✓ Each has advantages, depending on goals of analysis
  - ✓ Within this course the use of AIC is emphasized

Likelihood Ratio Test

- ★ Statistical test based on the deviance for **comparing two nested models**
  - ✓ Models are nested when parameters in more complex model, referred to as **full model**, can be set equal to 0 to obtain **reduced model**
  - ✓ Intercept-only model (reduced model) is nested in the slope model (full model)
- ★ Test statistic: difference in deviances between full and reduced models
  - ✓ Distributed as chi-squared, with *df* equal to the difference in the number of parameters

$$\boxed{\chi^2 = \text{deviance}_R - \text{deviance}_F}$$

- ★ Larger values of χ^2 indicate better fit
  - ✓ Parallels to AIC (discussed in future notes)

ML Standard Errors

- ★ Emphasis has been on computing ML point estimates via minimum deviance
- ★ Sampling fluctuation suggests uncertainty about every part of any analysis
  - ✓ Uncertainty in point estimates indexed by the standard error (SE)
  - ✓ Precision refers to extent of **uncertainty indexed by SE**
- ★ Precision is numerically indexed in many ways
  - ✓ Compute **ratio of estimate to its estimated SE** (i.e., *t*-ratio)

$$\boxed{t = \frac{\hat{\beta}_k}{\hat{SE}_{\hat{\beta}_k}}}$$

- ★ Absolute values of $t \approx 0$ indicate relatively low precision
  - ✓ High precision is desirable
- ★ t-ratio is a type of standardized effect
  - ✓ Use of t as a relative measure (without cutoffs or statistical tests) is emphasized
- ★ Precision is can also be numerically indexed using a CI
  - ✓ CI for fixed effect

$$\boxed{\hat{\beta}_k \pm 2 \cdot \hat{SE}_{\hat{\beta}_k}}$$

- ★ Interval offers applied researchers plausible estimates of the parameter values

★ Size of an SE is determined by **curvature of the deviance function**

- ✓ Relatively **flat** deviance functions indicate **low precision** (i.e., greater uncertainty)
- ✓ Relatively **high curvature** indicates **high precision**