Lecture 3:
Inference in Simple Linear Regression

BMTRY 701
Biostatistical Methods II

Interpretation of the SLR model

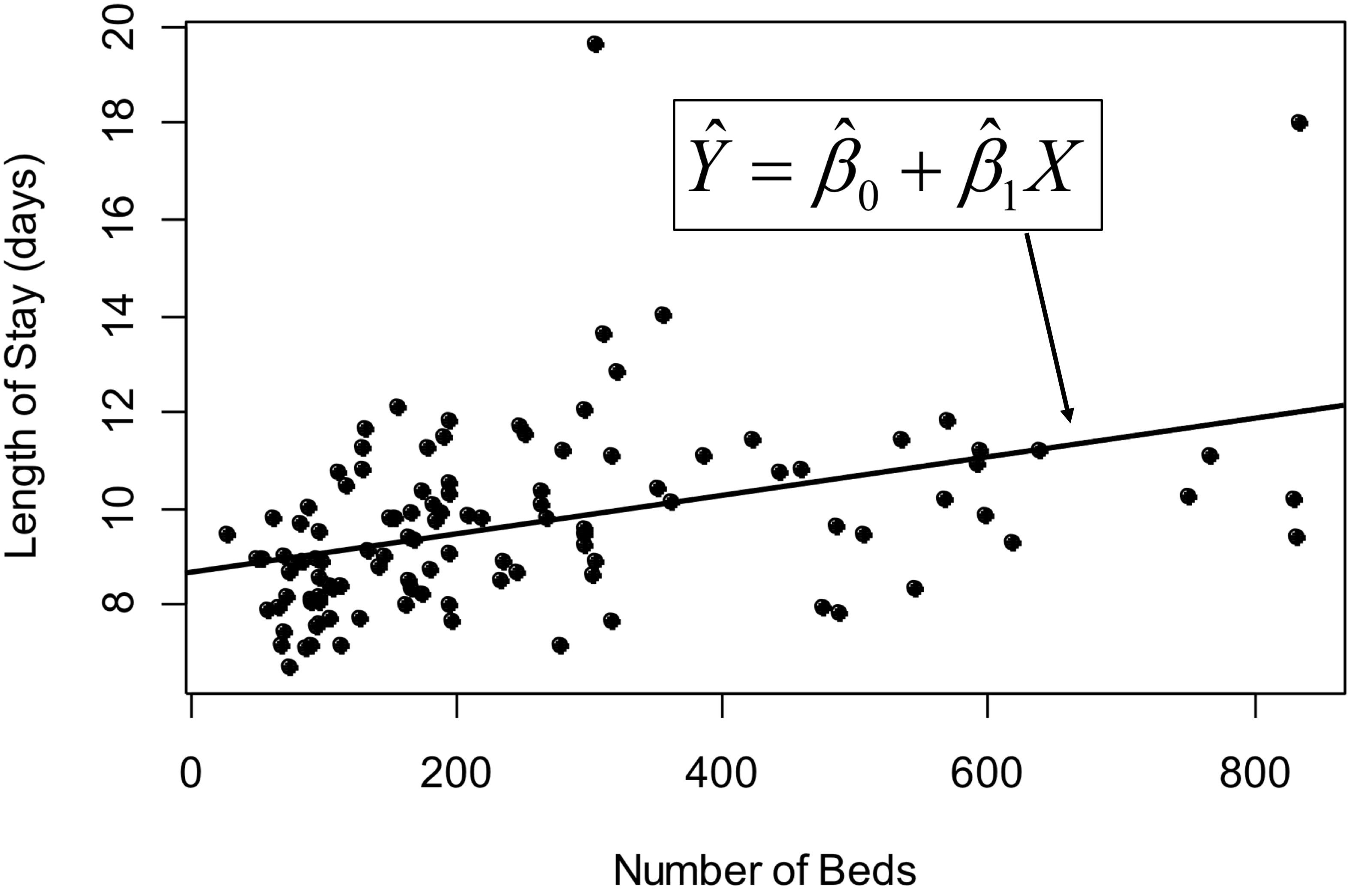
- Assumed model:

$$\boxed{E(Y) = \beta_0 + \beta_1 X}$$

- Estimated regression model: Takes the form of a line.

$$\boxed{\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X}$$

SENIC data


$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Length of Stay (days)
8
10
12
14
16
18
20
0
200
400
600
800
Number of Beds

Predicted Values

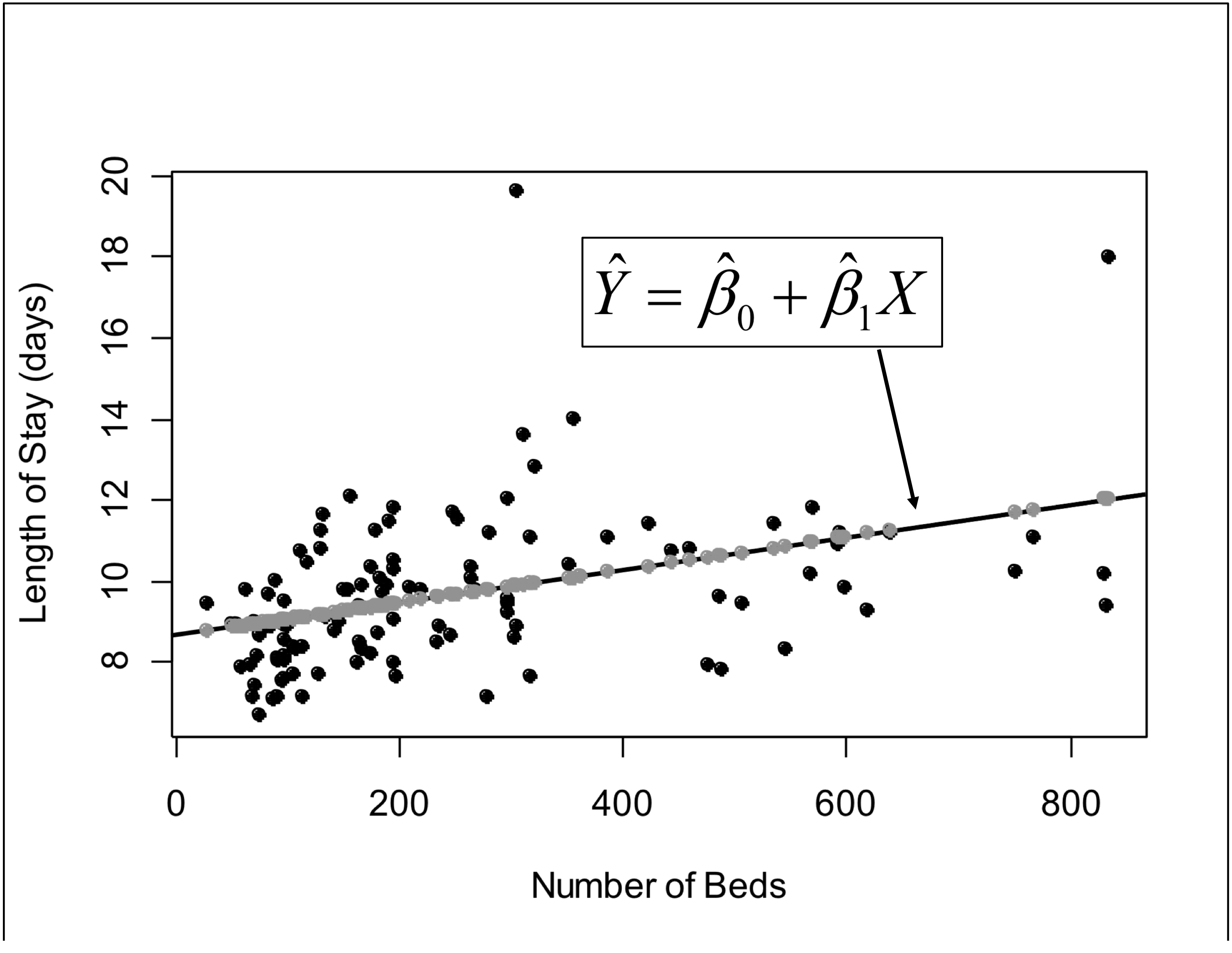
- For a given individual with covariate *X_i*:

$$\boxed{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i}$$

- This is the *fitted* value for the ith individual

- The fitted values fall on the regression line.

SENIC data



$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Length of Stay (days)
8
10
12
14
16
18
20
Number of Beds
0
200
400
600
800

SENIC Data

```
> plot(data$BEDS, data$LOS, xlab="Number of Beds",
       ylab="Length of Stay (days)", pch=16)
> reg <- lm(data$LOS~ data$BEDS)
> abline(reg, lwd=2)
> yhat <- reg$fitted.values
> points(data$BEDS, yhat, pch=16, col=3)
> reg

Call:
lm(formula = data$LOS ~ data$BEDS)

Coefficients:
(Intercept)    data$BEDS
   8.625364     0.004057
```

Estimating Fitted Values

- For a hospital with 200 beds, we can calculate the fitted value as

$$8.625 + 0.00406*200 = 9.44$$

- For a hospital with 750 beds, the estimated fitted value is

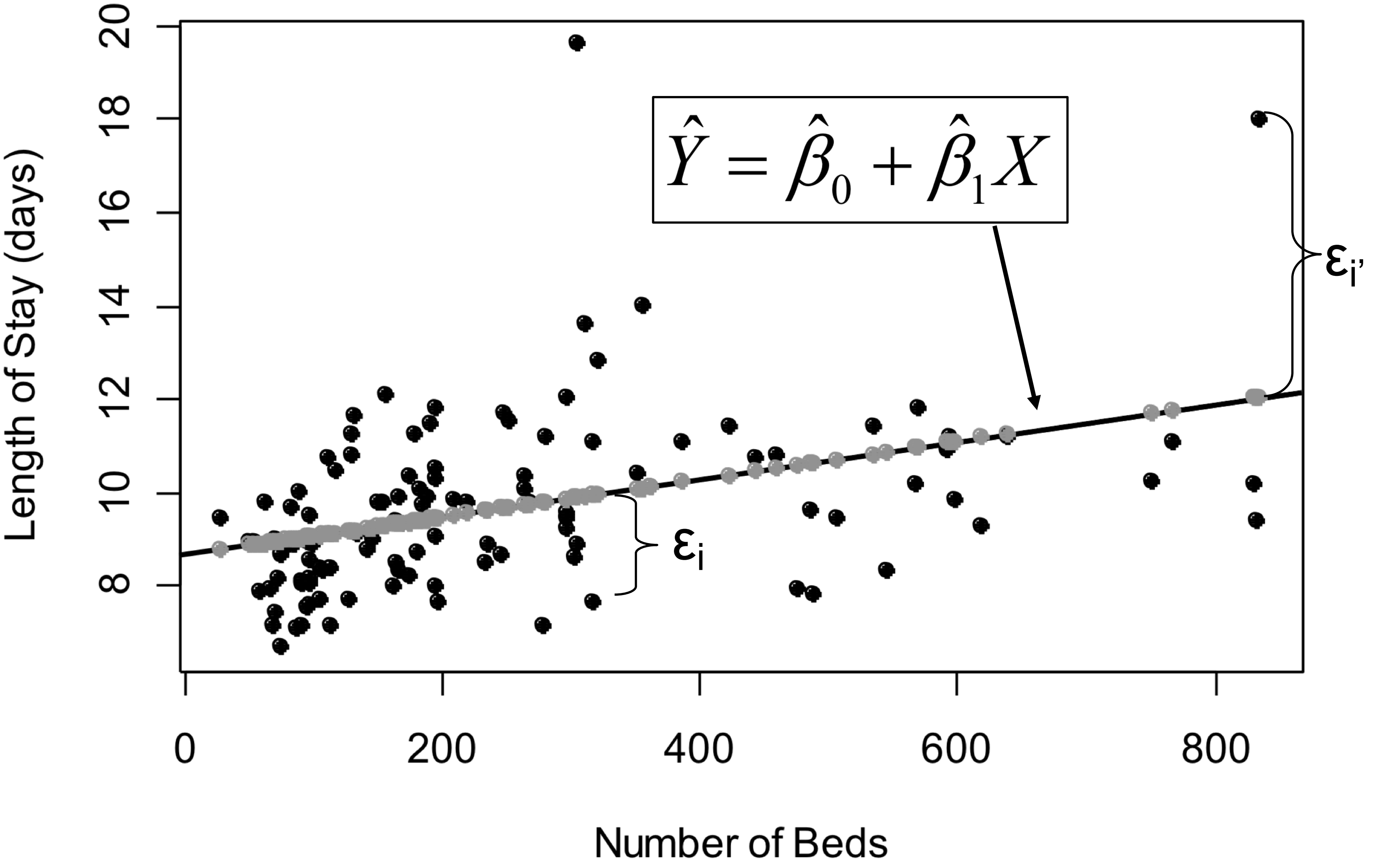
$$8.625 + 0.00406*750 = 11.67$$

Residuals

- The difference between observed and fitted
- Individual-specific
- Recall that $E(\varepsilon_i) = 0$

$$\boxed{\hat{\varepsilon}_i = e_i = Y_i - \hat{Y}_i}$$

SENIC data


$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
$\varepsilon_{i'}$
ε_i
Length of Stay (days)
8
10
12
14
16
18
20
Number of Beds
0
200
400
600
800

R code

- Residuals and fitted values are in the regression object

```
# show what is stored in ‘reg’
attributes(reg)
# show what is stored in ‘summary(reg)’
attributes(summary(reg))
# obtain the regression coefficients
reg$coefficients
# obtain regression coefficients, and other info
# pertaining to regression coefficients
summary(reg)$coefficients
# obtain fitted values
reg$fitted.values
# obtain residuals
reg$residuals
# estimate mean of the residuals
mean(reg$residuals)
```

Making pretty pictures

- You should plot your regression line!
- It will help you ‘diagnose’ your model for potential problems

```
plot(data$BEDS, data$LOS, xlab="Number of Beds",
       ylab="Length of Stay (days)",pch=16)
reg <- lm(data$LOS~ data$BEDS)
abline(reg, lwd=2)
```

A few properties of the regression line to note

- Sum of residuals = 0

- The sum of squared residuals is minimized (recall least squares)

- The sum of fitted values = sum of observed values

- The regression line always goes through the mean of X and the mean of Y

Estimating the variance

- Recall another parameter: σ^2
- It represents the variance of the residuals
- Recall what we know about estimating variances for a variable from a single population:

$$\boxed{s^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

- What would this look like for a corresponding regression model?

Residual variance estimation

- “sum of squares”
- for residual sum of squares:
  - RSS = residual sum of squares
  - SSE = sum of squares of errors (or error sum of squares)

$$\boxed{SSE = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \overline{\varepsilon})^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}$$

Residual variance estimation

- What do we divide by?
- In single population estimation, why do we divide by n-1?

$$\boxed{\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2}}$$

- Why n-2?
- MSE = **mean** square error
- RSE = residual standard error = sqrt(MSE)

Normal Error Regression

- New: assumption about the distribution of the residuals

$$\boxed{\varepsilon_i \sim N(0,\sigma^2)}$$

- Also assumes independence (which we had before).
- Often we say they are ‘iid”: “independent and identically distributed”

How is this different?

- We have now added “probability” to our model
- This allows another estimation approach: Maximum Likelihood
- We estimate the parameters ($\beta_0, \beta_1, \sigma^2$) using this approach instead of least squares
- **Recall least squares: we minimized Q**
- **ML: we maximize the likelihood function**

The likelihood function for SLR

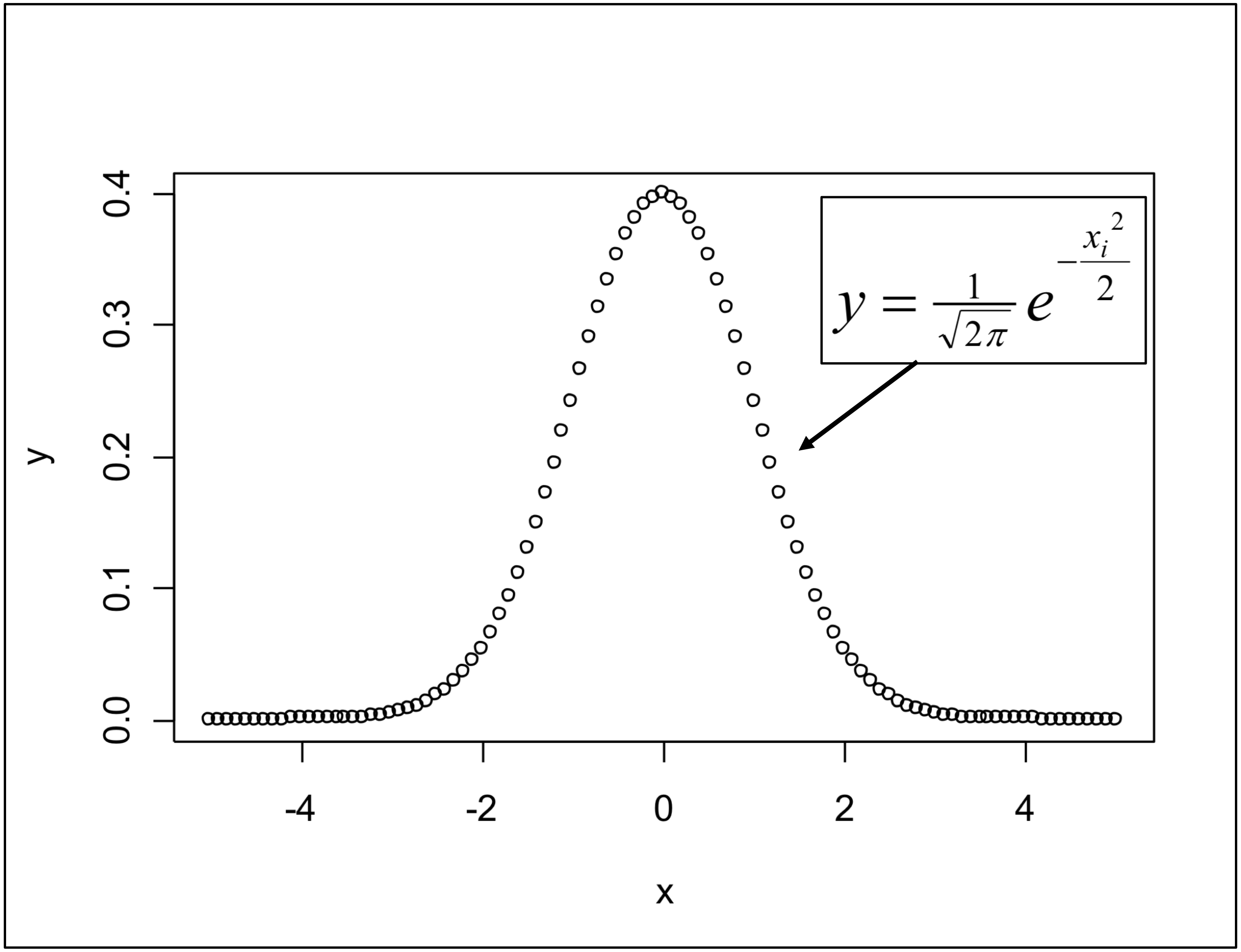
- Taking a step back
- Recall the pdf of the normal distribution

$$\boxed{f(x;\mu,\sigma^2)=\tfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

- This is the probability density function for a random variable X.
- For a ‘standard normal’ with mean 0 and variance 1:

$$\boxed{f(x;\mu=0,\sigma^2=1)=\tfrac{1}{\sqrt{2\pi}}e^{-\frac{x_i{}^2}{2}}}$$

Standard Normal Curve



$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^{\,2}}{2}}$$

The likelihood function for a normal variable

- From the pdf, we can write down the likelihood function

$$\boxed{pdf : f(x;\mu,\sigma^2) = \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right)}$$

- The likelihood is the product over n of the pdfs:

$$\boxed{L(\mu,\sigma^2 \mid x) = \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right)}$$

The likelihood function for a SLR

- What is “normal” for us?
- the residuals
  - what is E(ε) = μ?

$$\boxed{L(\mu,\sigma^2 \mid x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{{\varepsilon_i}^2}{2\sigma^2}\right)}$$

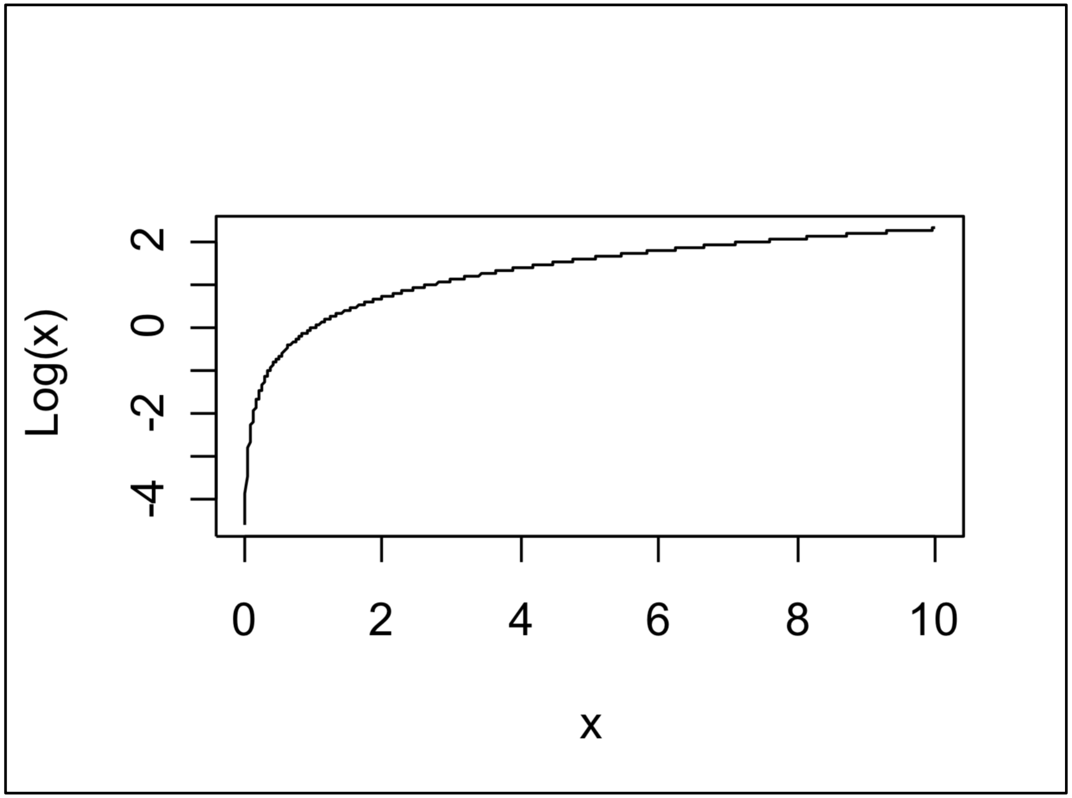
Maximizing it

- We need to maximize ‘with respect to’ the parameters.
- But, our likelihood is not written in terms of our parameters (at least not all of them).

$$\boxed{\begin{aligned} L(\mu,\sigma^2 \mid x) &= \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{{\varepsilon_i}^2}{2\sigma^2} \right) \\ &= \\ &= \\ &= \\ &= \end{aligned}}$$

Maximizing it

- now what do we do with it?
- it is well known that maximizing a function can be achieved by maximizing it’s log. (Why?)



Log-likelihood

$$L(\mu, \sigma^2 \mid x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right)$$

$$l(\mu, \sigma^2 \mid x) = \log\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right) \right]$$

Still maximizing….

- How do we maximize a function, with respect to several parameters?
- Same way that we minimize:
    - we want to find values such that the first derivatives are zero (recall slope=0)
    - take derivatives with respect to each parameter (i.e., partial derivatives)
    - set each partial derivative to 0
    - solve simultaneously for each parameter estimate
- This approach gives you estimates of β_0, β_1, σ^2:

$$\boxed{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2}$$

No more math on this….

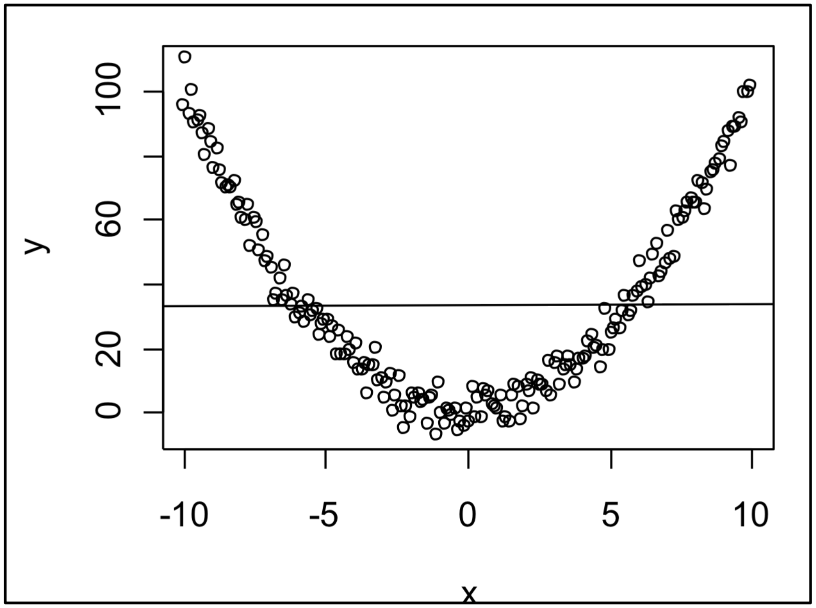
- For details see MPV, Page 47, section 2.10
- We call these estimates “maximum likelihood estimates”
- a.k.a “MLE”
- The results:
  - MLE for β_0 is the same as the estimate via least squares
  - MLE for β_1 is the same as the estimate via least squares
  - MLE for σ^2 is the same as the estimate via least squares

So what is the point?!

- Linear regression is a **special case** of regression
- for linear regression Least Squares and ML approaches give same results
- For later regression models (e.g., logistic, poisson), they differ in their estimates
- Going back to LS estimates
  - what assumption did we make about the distribution of the residuals?
  - LS has fewer assumptions than ML
- **Going forward: We assume normal error regression model**

The main interest: β_1

- The slope is the focus of inferences
- Why? If $\beta_1 = 0$, then there is no linear association between x and y
- But, there is more than that:
  - it also implies no relation of ANY type
  - this is due to assumptions of
    - constant variance
    - equal means if $\beta_1 = 0$
- Extreme example:



Inferences about β_1

- To make inferences about β_1, we need to understand its sampling distribution

$$\boxed{\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1))}$$

- More details:
  - The expected value of the estimate of the slope is the true slope

$$\boxed{E(\hat{\beta}_1) = \beta_1}$$

  - The variance of the sampling distribution for the slope is

$$\boxed{\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}}$$

Inferences about β_1

- More details (continued)
  - **Normality stems from the knowledge that the slope estimate is a linear combination of the Y's**
  - Recall:
    - Yi are independent and normally distributed (because residuals are normally distributed)
    - The sum of normally distributed random variables is normal
    - Also, a linear combination of normally distributed random variabes is normal.
    - (what is a linear combination?)

So much theory! Why?

- We need to be able to make inferences about the slope
- If the sampling distribution is normal, we can standardize to a standard normal:

$$\boxed{\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1))}$$

$$\boxed{\frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} \sim N(0,1)}$$

Implications

- Based on the test statistic on previous slide, we can evaluate the “statistical significance” of our slope.
- To test that the slope is 0:

$$\boxed{\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}}$$

- Test statistic:

$$\boxed{Z = \frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)} \sim N(0,1)}$$

But, there is a problem with that….

$$\boxed{Z = \frac{\hat{\beta}_1}{\sigma(\hat{\beta}_1)} \sim N(0,1)}$$

Do we know what the true variance is?

- Recall $$\boxed{\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}}$$

which depends on the true SD of the residuals

But, we have the tools to deal with this

- What do we do when we have a normally distributed variable but we do not know the true variance?
- Two things:
  - we estimate the variance using the “sample” variance
    - in this case, we use our estimated MSE
    - we plug it into our estimate of the variance of the slope estimate

$$\boxed{\hat{\sigma}^2(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum (X_i - \overline{X})^2}}$$

  - we use a t-test instead of a Z-test.

$$\boxed{t^* = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \sim t_{n-2}}$$

The t-test for the slope

- Why n-2 degrees of freedom?
- **The ratio of the estimate of the slope and its standard error has a t-distribution**

$$\boxed{t^* = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \sim t_{n-2}}$$

- For more details, page 22, section 2.3.

What about the intercept?

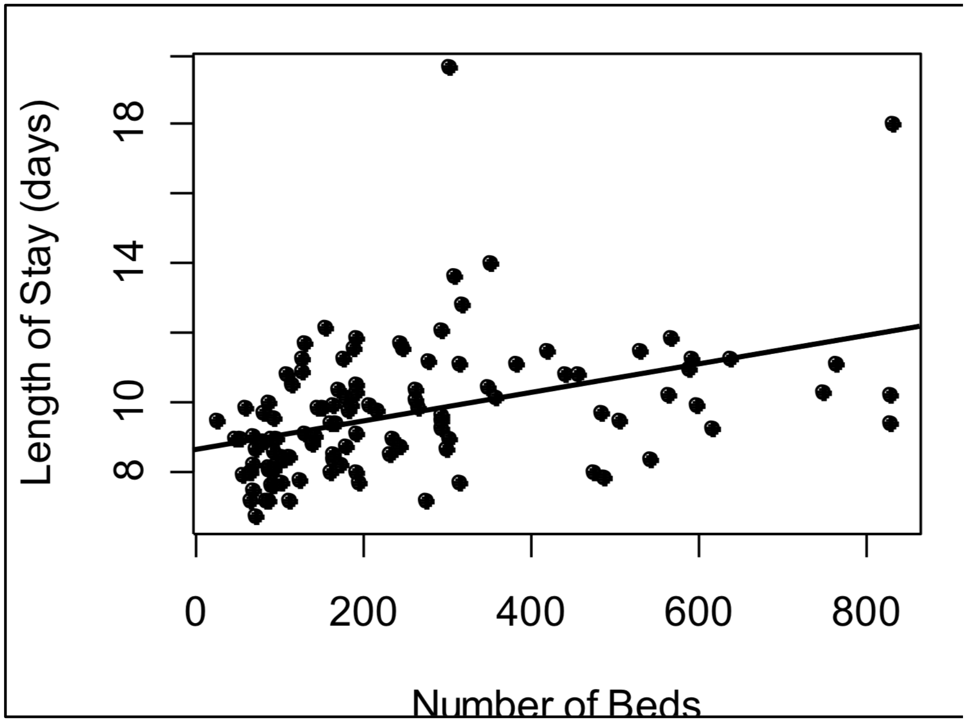
- ditto
- All the above holds

$$\boxed{t^* = \frac{\hat{\beta}_0}{\hat{\sigma}(\hat{\beta}_0)} \sim t_{n-2}}$$

- However, we rarely test the intercept

Time for data (phewf!)

Is Number of Beds associated with Length of Stay?



```
> reg <- lm(data$LOS ~ data$BEDS)
> summary(reg)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.6253643  0.2720589  31.704  < 2e-16 ***
data$BEDS   0.0040566  0.0008584   4.726 6.77e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.752 on 111 degrees of freedom
Multiple R-Squared: 0.1675,     Adjusted R-squared:  0.16
F-statistic: 22.33 on 1 and 111 DF,  p-value: 6.765e-06
```

Important R commands

- `lm`: fits a linear regression model
  - for simple linear regression, syntax is

    ```
    reg <- lm(y ~ x)
    ```

  - more covariates can be added:

    ```
    reg <- lm(y ~ x1+x2+x3)
    ```

- `abline`: adds a regression line to an already existing plot if object is a regression object
  - syntax: `abline(reg)`
- Extracting results from regression objects:
  - residuals: `reg$residuals`
  - fitted values: `reg$fitted.values`