Statistical Thinking for Forensic Practitioners

Part 7: Regression and Analysis of Variance (Video 3)

Joe Zemmels, MSc.

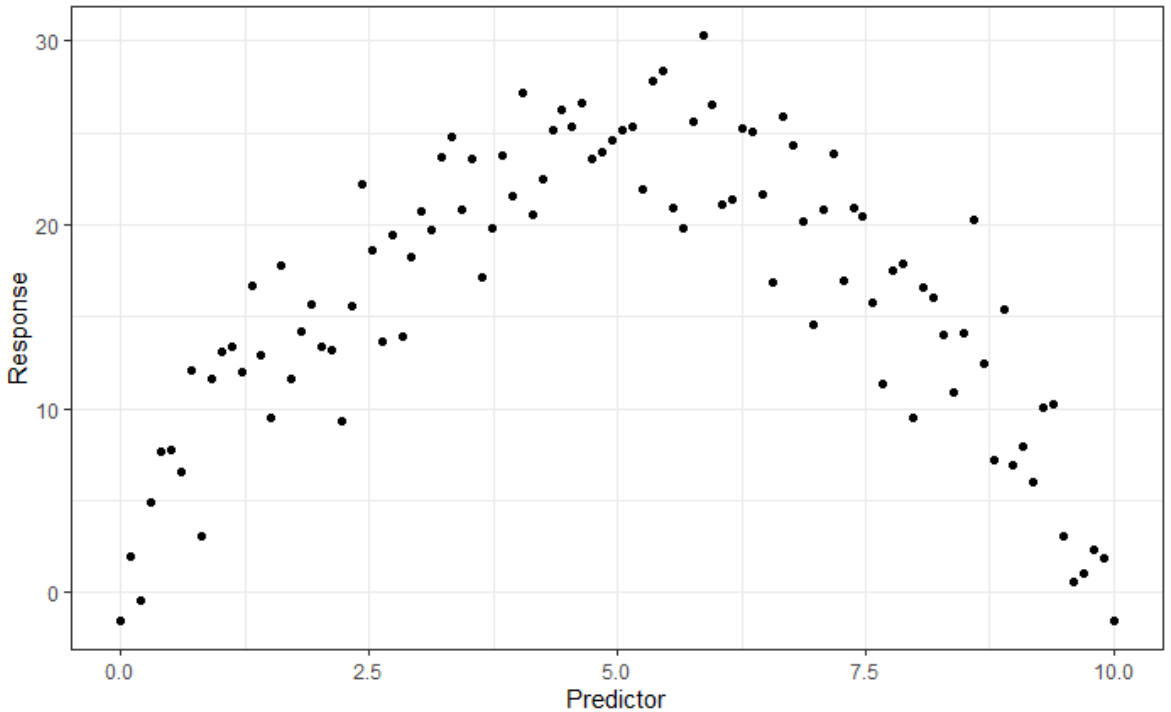
csafe
Center for Statistics and Applications in Forensic Evidence

Outline

- Quadratic regression
- Logistic regression & binary classification
- What we haven't covered.

Quadratic Regression Motivation

- Simple linear regression describes the *linear* relationship between a response and predictor.
- What if the relationship is non-linear?



- A *quadratic* relationship between y and x is $y = a + bx + cx^2$

Quadratic Regression

- Adding new terms to our model makes it more "flexible."
- Let y_i represent the response and x_i the predictor of subject i, $i = 1, ..., n$.
- A quadratic regression model is:

$$y_i = E(y_i|x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where ϵ_i are assumed independent $N(0, \sigma^2)$, $i = 1, ..., n$.

- Our interest is in estimating $\beta_0, \beta_1, \beta_2$.
- Using b_0, b_1, b_2 estimates, an unbiased estimator for σ^2 is:

$$MSE = \hat{\sigma}^2 = \frac{1}{n-3} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i - b_2 x_i^2)^2.$$

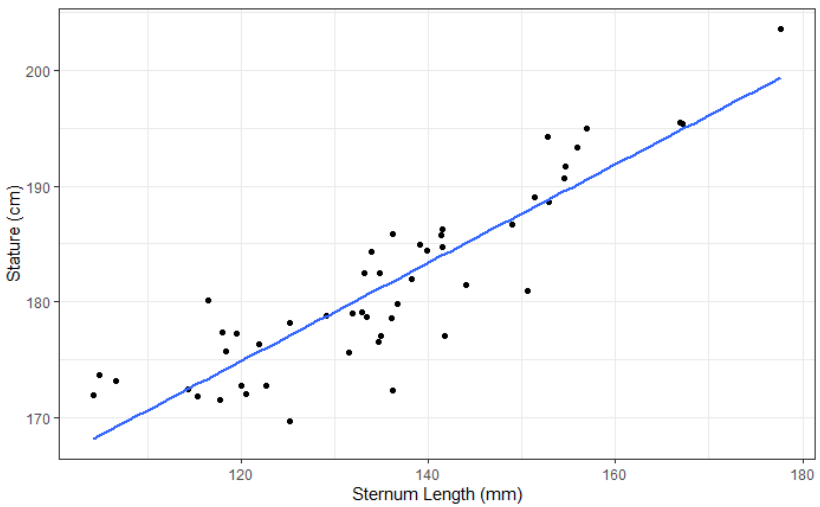
where $n - 3$ are the d.f. after estimating $\beta_0, \beta_1, \beta_2$.

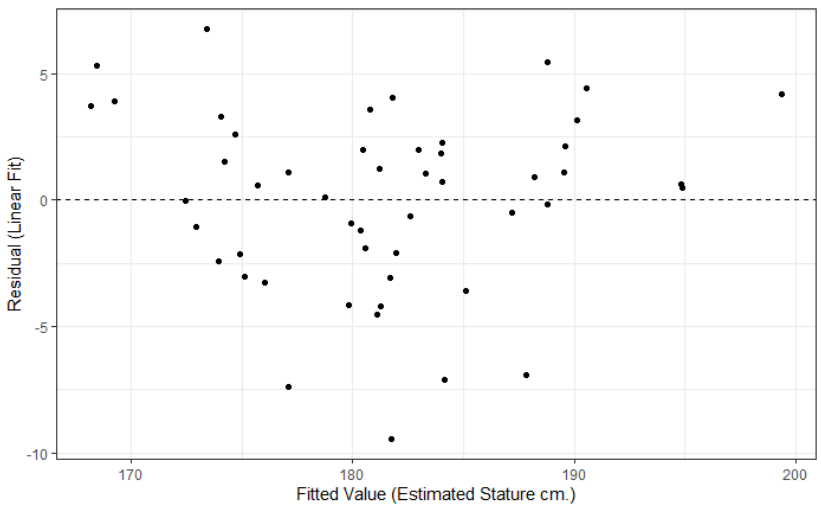
Quadratic Regression Estimation

- We estimate the regression coefficients $\beta_0, \beta_1, \beta_2$ using the method of least squares.
- Unfortunately, the estimators b_0, b_1, b_2 do not have a nice closed form as the estimators in simple linear regression.
- Finding b_0, b_1, b_2 involves linear algebra that is outside the scope of the course (you won't be expected to know it).
- Most statistical packages contain built-in function to calculate b_0, b_1, b_2.
  - Available via the Analysis ToolPak add-in in Excel.

Quadratic Regression Example

- Consider again the problem of predicting stature of a victim given their sternum length.
- Consider a SLR & its residuals for the male data:





- May be worthwhile to explore a quadratic regression.

Quadratic Regression Example

Like in SLR, we need to fit this problem within the quadratic regression "mold" by introducing mathematical notation.

- Let y_i be the stature of the ith subject, $i = 1, ..., 50$.
- Let x_i be the sternum length of the ith subject, $i = 1, ..., 50$.
- Assume

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where ϵ_i are independent $N(0, \sigma^2)$ and $\beta_0, \beta_1, \beta_2, \sigma^2$ are unknown.

Quadratic Regression Example

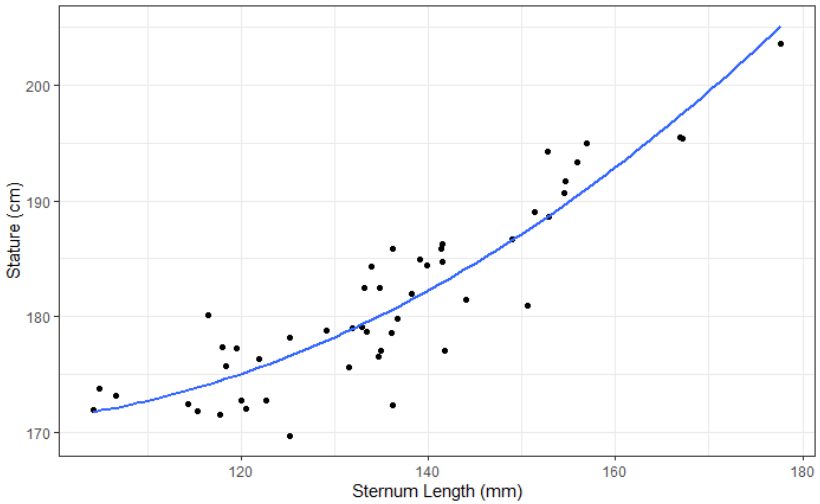
- The estimates are $b_0 = 203.249$, $b_1 = -.746$, $b_2 = .004$.
- Under the model assumptions, this means

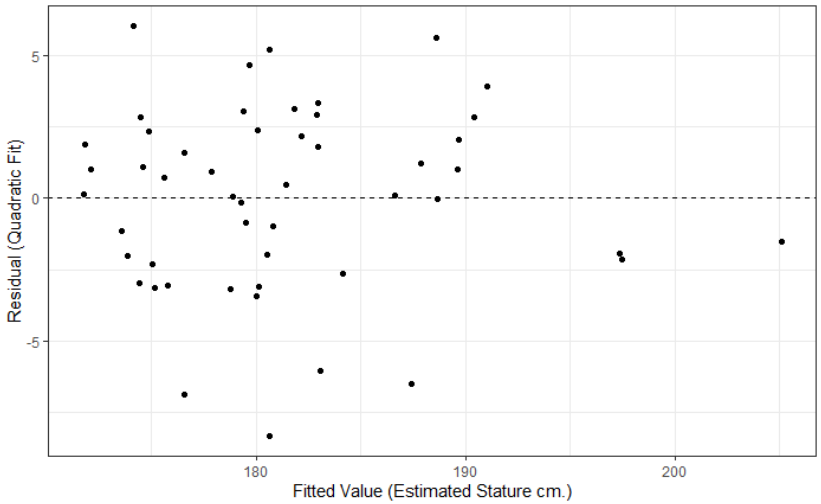
$$\hat{y}_i = 203.249 - .746x_i + .004x_i^2.$$

- Why is this useful?
  - Suppose we observe sternum #51 where $x_{51} = 150$ mm.
  - Our model estimates that this person's height is

$$\hat{y}_{51} = 203.249 - .746(150) + .004(150)^2 = 181.349 \text{ cm}.$$

Quadratic Regression Example





- Residuals look more "well-behaved" than the SLR model's.
- Formal procedures exist to test whether the quadratic model is actually "better" than the SLR model.
  - Analysis of Variance (ANOVA)
- Can use hypothesis tests/confidence intervals as in SLR, although calculations rely on linear algebra beyond the course's scope.

Quadratic Regression in Excel

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.910520065					
5	R Square	0.829046788					
6	Adjusted R Square	0.821772183					
7	Standard Error	3.297484311					
8	Observations	50					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	2478.364104	1239.182052	113.9645129	9.38892E-19	
13	Residual	47	511.0499307	10.87340278			
14	Total	49	2989.414035				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	203.2489206	25.35403505	8.016432894	2.4074E-10	152.2431811	254.25466
18	sternumLength	-0.745960423	0.370662265	-2.012507057	0.049916267	-1.491636718	-0.000284129
19	sternumLengthSquared	0.004256986	0.00134368	3.168153856	0.002696198	0.00155385	0.006960123
20							
21							
22							
23	RESIDUAL OUTPUT						
24							
25	*Observation*	*Predicted stature*	*Residuals*	*Standard Residuals*			
26	1	174.1281952	6.024579608	1.865490839			
27	2	205.0933046	-1.537619375	-0.476118675			
28	3	186.6186936	0.088320433	0.027348125			
29	4	173.8472443	-2.032249184	-0.629279133			
30	5	182.8866641	2.915779041	0.902861186			
31	6	174.5147243	2.842686777	0.880228412			
32	7	175.1916944	-3.132968973	-0.970113319			
33	8	178.9071063	0.06257902	0.019377383			
34	9	188.6150708	5.595608164	1.732661272			
35	10	184.1400452	-2.656692084	-0.82263578			
36	11	182.9388345	3.348072285	1.036719337			
37	12	187.4328982	-6.506370302	-2.014675709			
38	13	175.8062296	-3.069812483	-0.950557124			

cadaverData | **Quadratic Regression**

Categorical Response

- We may be interested in learning about a categorical variable.
- Can we predict the value of a categorical variable given other information?
- Examples:
  - Whether two toolmarks match given the number of CMS
  - Brand of a shoe given the number of triangles on the sole
  - Occurrence of a heart attack given vital signs
- The technical term for predicting a categorical variable is *classification* (assigning observations into "classes")

Binary Classification

- Variable of interest takes on one of two categories (match/non-match, male/female, heart attack/no heart attack, etc.)
- Common to estimate the probability that an observation will take one of these 2 values.
- Choose a *decision rule* to decide how an observation is classified based on its estimated probability.
  - Example: if $\widehat{\Pr}(\text{Male}) > .5$ for a particular observation, then classify the observation as male. Otherwise, classify as female.

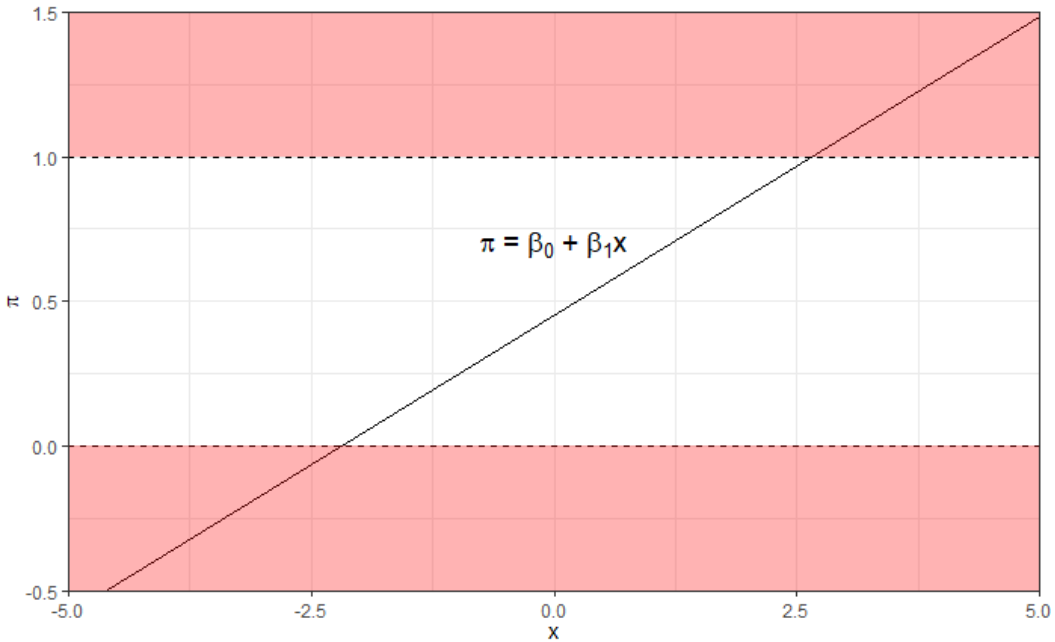
Logistic Regression

- Statistical framework for estimating the probability a response, y_i, equals 1 given a predictor x_i.
- Model assumes responses, y_i, are independently distributed Binomial$(n = 1, \pi_i)$, $i = 1, ..., n$.[1]
  - $n = 1$ implies each y_i takes on values 0 or 1.
  - π_i is the probability that ith observation takes on the value 1 given the value of predictor x_i.
  - We will arbitrarily assign the 2 categories to 0 and 1.
- Example: suppose we choose male $= 1$ and female $= 0$. Then $\pi_i = \Pr(y_i = 1|x_i) = \Pr(i\text{th observation is male } |x_i)$.
- Is a "regression" because we estimate the continuous parameter $\pi_i \in [0, 1]$ using predictors $x_i \in \mathbb{R}$.

[1] You may hear Binomial$(1,\pi_i)$ referred to as the Bernoulli(π_i) distribution.

Logistic Regression

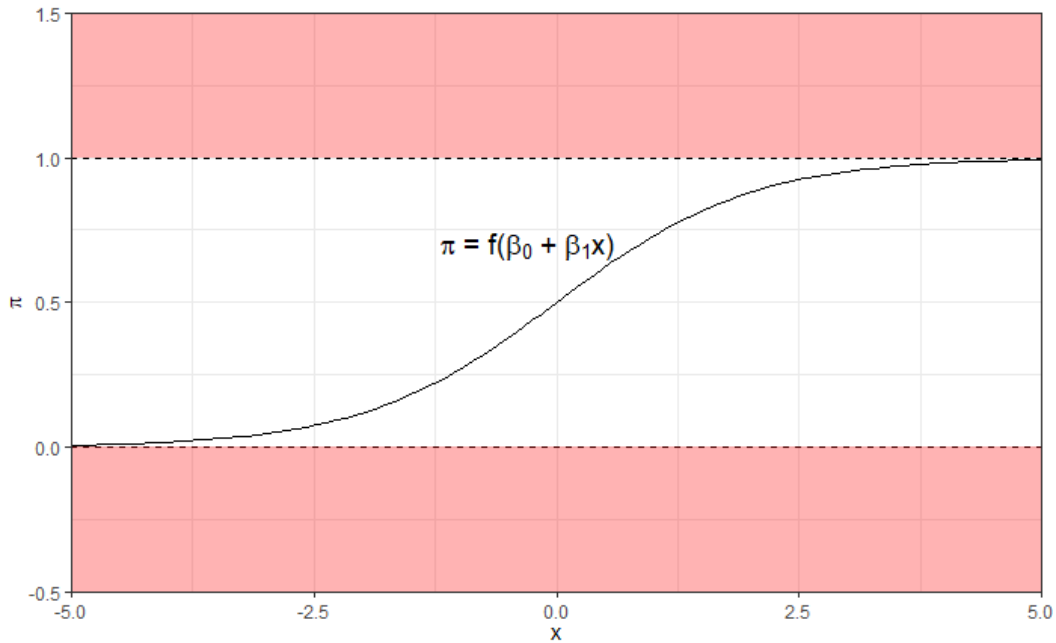
- Want to model the probability π_i, bounded between 0 and 1, with predictors x_i that might take on any real value.
- A linear model like $\pi_i = \beta_0 + \beta_1 x_i$ wouldn't be effective because the estimated $\hat{\pi}_i$ could potentially lie outside $[0, 1]$.



The Logistic Function

- The *logistic function* has the form $f(x) = \frac{e^x}{1+e^x}$.
  - Accepts any $x \in \mathbb{R}$ but returns $f(x) \in (0, 1)$.
- Logistic regression with predictor x_i assumes

$$\pi_i = f(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}.$$



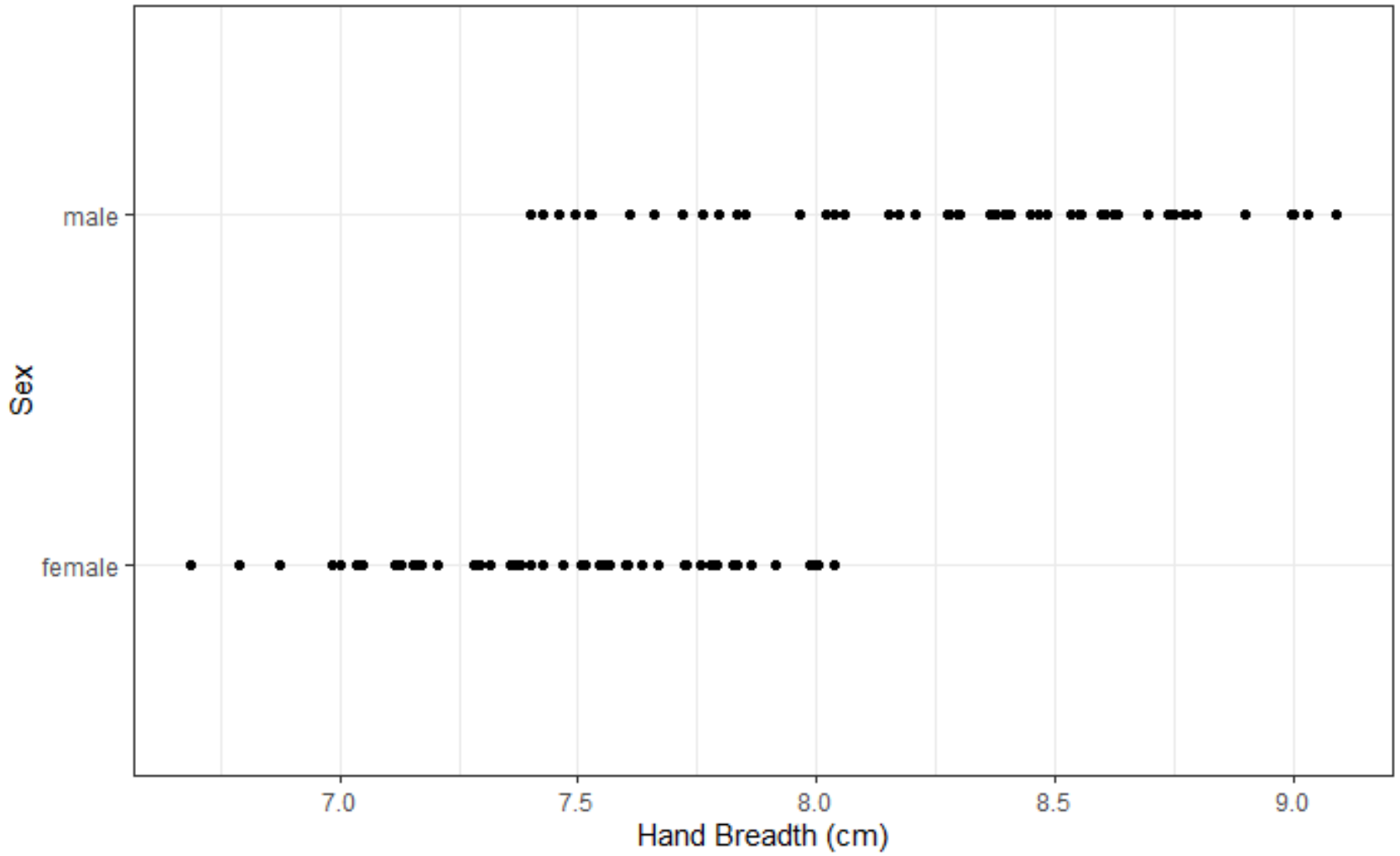
- Like in simple linear regression, our task is to estimate β_0 and β_1.
- Once we obtain b_0, b_1 we can calculate $\hat{\pi}_i = f(b_0 + b_1 x_i)$.
- Estimation of β_0, β_1 involves complicated optimization algorithms that are outside the scope of the course.
  - Most statistical packages have built-in estimation functions (excluding Excel, unfortunately).

Logistic Regression Example

- Determination of sex is useful for establishing a biological profile of the deceased.
- We may only find skeletal remains or body parts.
- Can we predict someone's sex based on anthropometric (physical bodily) measurements?
- Hand breadths measured for 50 males & 50 females.[2]

[2] Source

Logistic Regression Example



Logistic Regression Example

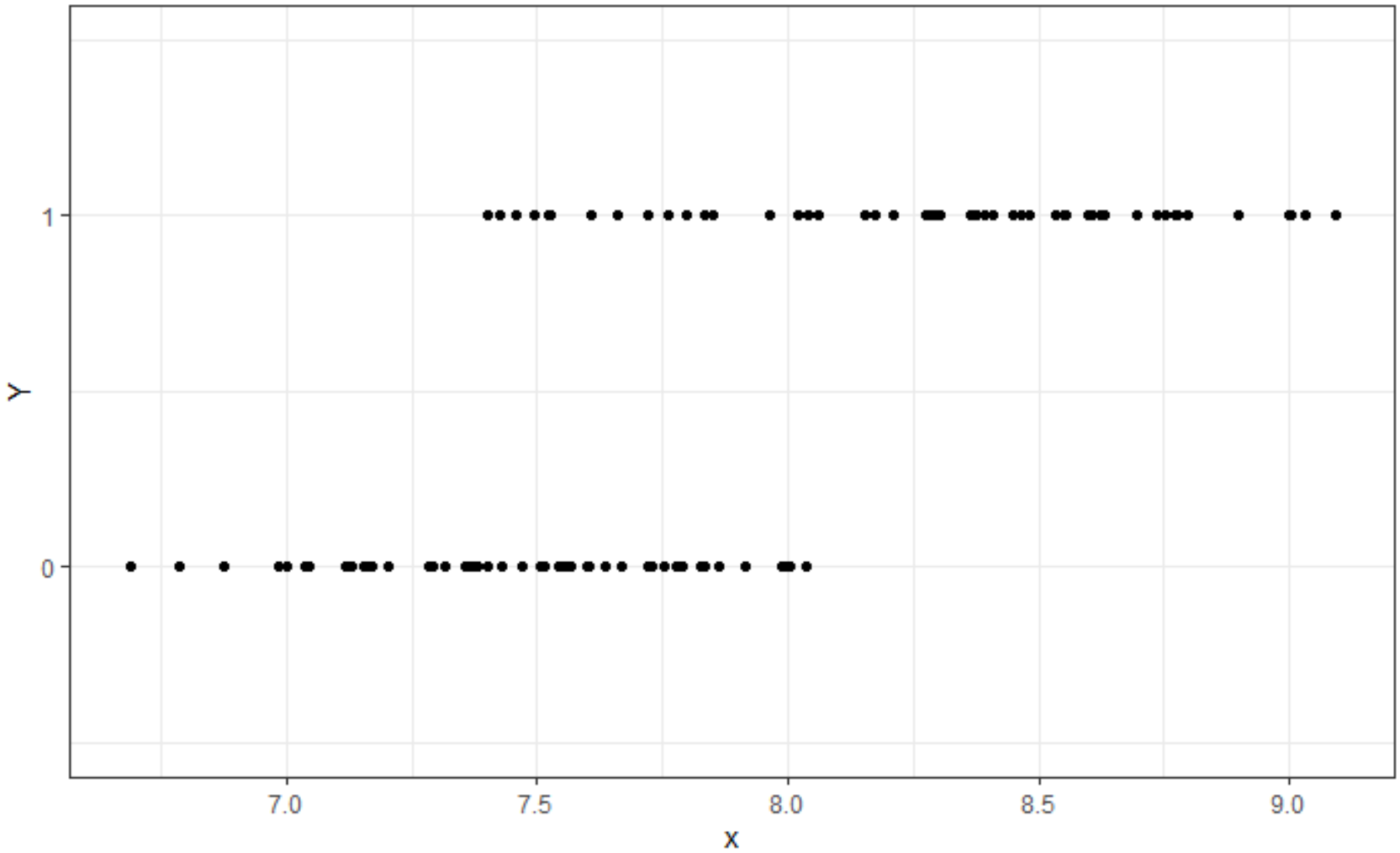
We now introduce notation to fit this problem within the logistic regression "mold."

- Assign male $= 1$ and female $= 0$.
- Let y_i be the sex of the ith subject, $i = 1, ..., 100$.
- Assume y_i are independently distributed as Binomial$(1, \pi_i)$ where π_i is the probability that subject i is male.
- Let x_i be the hand breadth of the ith subject, $i = 1, ..., 100$.
- Assume

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

where β_0, β_1 are unknown parameters to be estimated.

Logistic Regression Example



Logistic Regression Example

- The estimated coefficients are $b_0 = -36.245$ and $b_1 = 4.634$.
- Under the model assumptions, this means

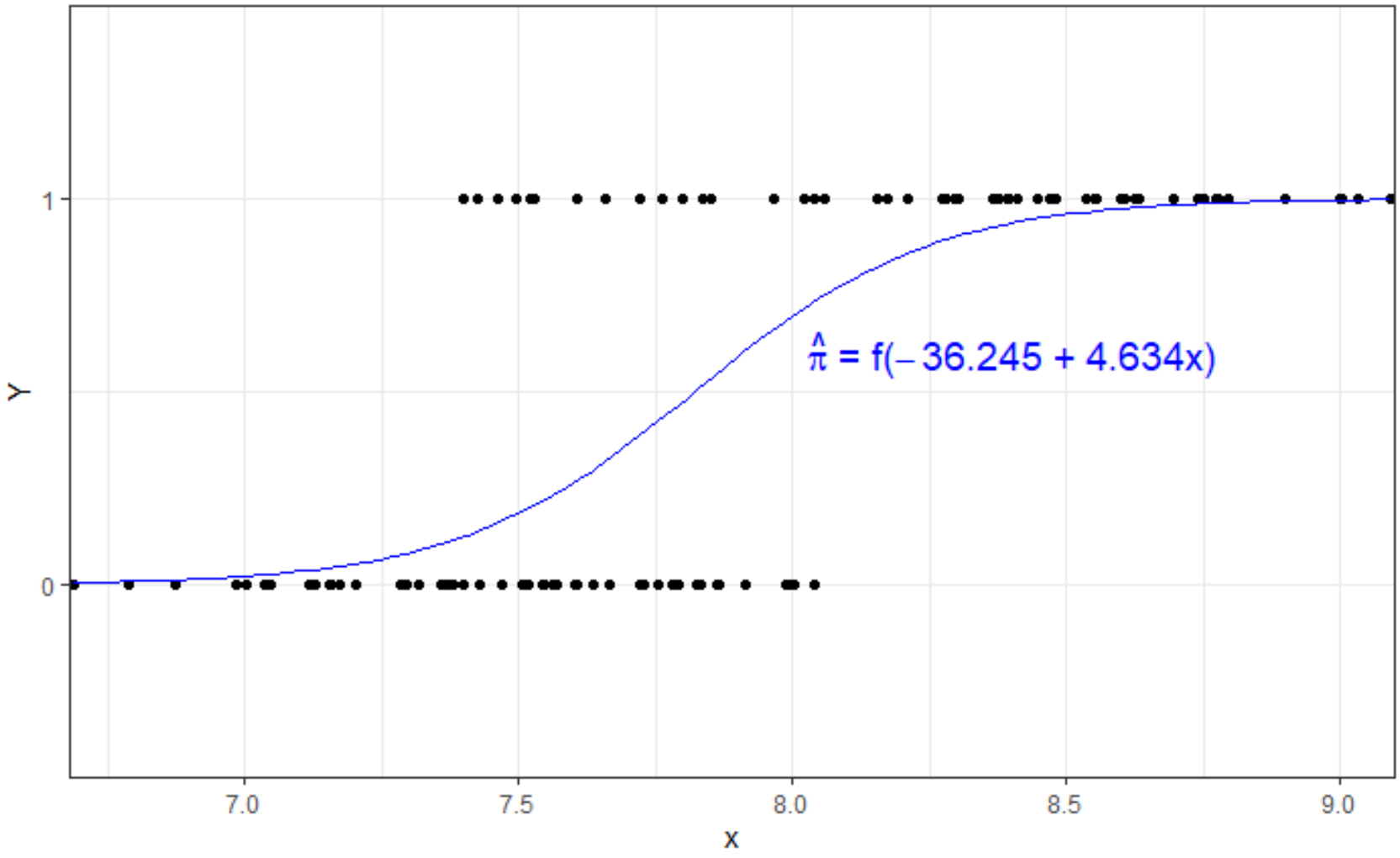
$$\hat{\pi}_i = f(b_0 + b_1 x_i) = \frac{e^{-36.245+4.634x_i}}{1 + e^{-36.245+4.634x_i}}.$$

- Why is this useful?
  - Suppose we observe a new subject #101 where $x_{101} = 7.25$ cm., but we don't know their sex.
  - Our model estimates that the probability the subject is male is

$$\hat{\pi}_{101} = f(b_0 + b_1(7.25)) = \frac{e^{-36.245+4.634(7.25)}}{1 + e^{-36.245+4.634(7.25)}} = .066$$

  or 6.6% chance. We can be confident that they are female!

Logistic Regression Example



Logistic Regression Example

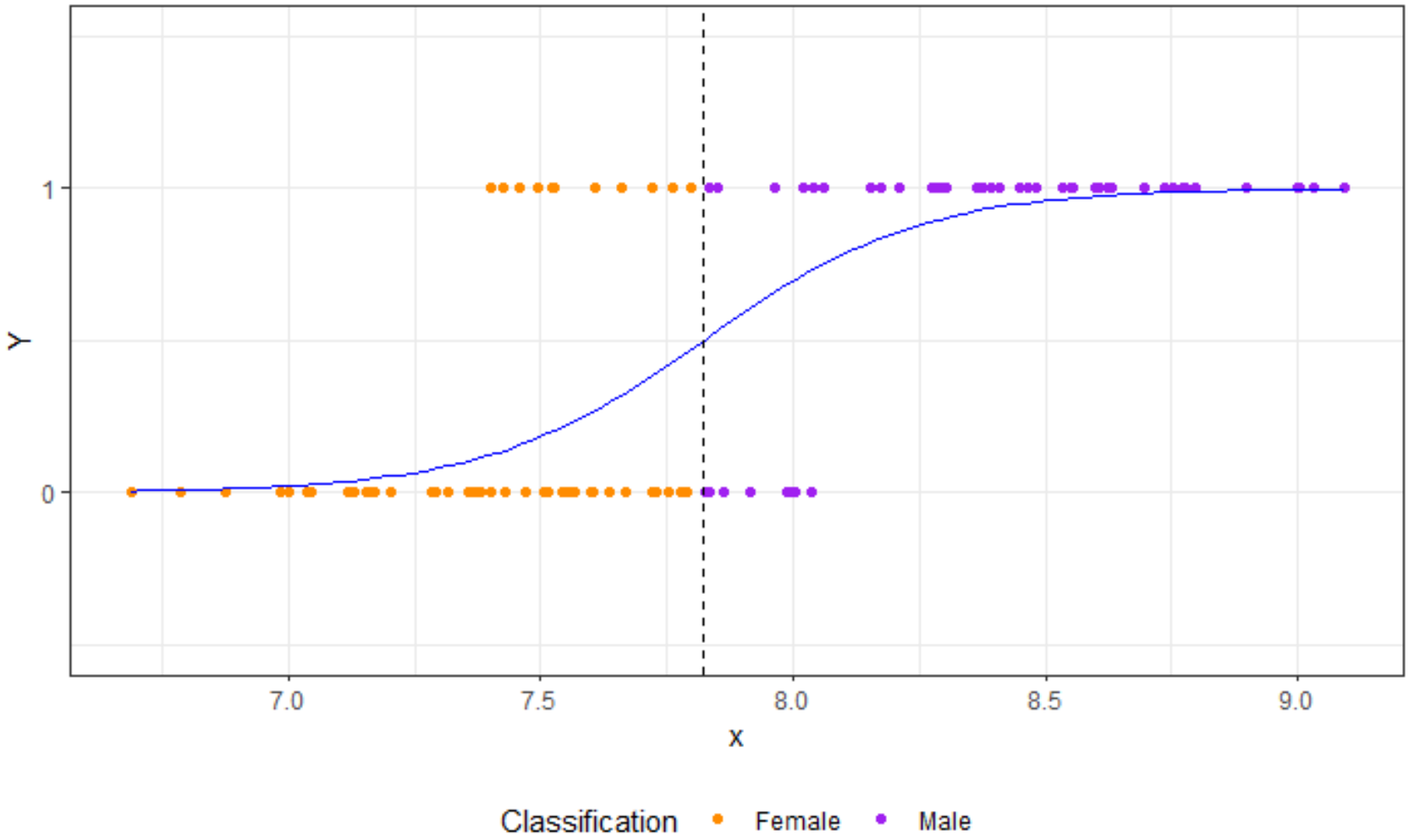
- Given a $\hat{\pi}_i$, we desire a rule that tells us to classify subject i as male or female.
- It seems reasonable that $\hat{\pi}_i > .5$ means we should classify subject i as male.
- Define a *decision rule* $d(\hat{\pi}_i)$ where

$$d(\hat{\pi}_i) = \begin{cases} 1 & \text{if } \hat{\pi}_i > .5 \\ 0 & \text{if } \hat{\pi}_i \leq .5 \end{cases}.$$

- Translation: if the ith subject is more likely a male, then classify them as male. Otherwise, classify them as female.
- Using algebra we can show that $\hat{\pi}_i > .5$ if $x_i > 7.82$ cm.[3]
  - $x_i = 7.82$ cm. is the "decision boundary" for a new observation

[3] see final slide

Logistic Regression Example



Logistic Regression Example

- For the price of having a simple decision rule we have made some misclassifications.
- A *confusion matrix* summarizes our performance

		Truth	
		Male	Female
Classification	Male	39	9
	Female	11	41

- To affect performance:
  - Obtain more data/different predictors (hand length, foot breadth, etc.)
  - Explore other models (can of worms - *all models are wrong, but some are useful*)
- Takeaway: algorithm classifications should never be taken at "face value" (e.g., will recidivate vs. will not recidivate). We need to consider underlying data, assumptions, etc.

- There are numerous ways to calculate confidence intervals for logistic regression probabilities.
- One popular method called the *endpoint method* relies on the fact that the logistic function is a *monotone transformation*.
- If $[L, U]$ are the lower and upper endpoints for an approximate $(1-\alpha)\%$ CI[4] for $\beta_0 + \beta_1 x_i$, an approximate $(1-\alpha)\%$ CI for π_i is

$$\left[\frac{e^L}{1+e^L}, \frac{e^U}{1+e^U}\right].$$

- Hypothesis tests for a hypothesized probability value, $H_0 : \pi_i = \pi_0$ say, can be performed by determining if the associated $(1-\alpha)\%$ CI covers π_0.

[4] we say "approximate" for technical reasons - the confidence interval becomes "exact" as the sample size $n \to \infty$.

What We Haven't Covered

- Model diagnostics for logistic regression[5]
- Multiple linear regression (how to handle multiple predictors)[6]
- Generalized linear models (what if my data are distributed Poisson, Log-Normal, etc.?)[7]
- Locally Estimated Scatterplot Smoothing (what if my data don't follow a known distribution?)[8]
- Time series models (modeling a variable observed repeatedly over time)[9]

... and much, MUCH more.

---

[5] Logistic Regression Lesson
[6] MLR Lesson
[7] GLM Lesson
[8] LOESS Article
[9] Time Series Online Course

Logistic Regression: Inverting a Probability

Assume $\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. Rearranging this for x_i is done by

$$\begin{aligned}
\pi_i \left(1 + e^{\beta_0+\beta_1 x_i}\right) &= e^{\beta_0+\beta_1 x_i} \\
\implies \pi_i &= e^{\beta_0+\beta_1 x_i} - \pi_i e^{\beta_0+\beta_1 x_i} \\
\implies \pi_i &= (1-\pi_i) e^{\beta_0+\beta_1 x_i} \\
\implies \frac{\pi_i}{1-\pi_i} &= e^{\beta_0+\beta_1 x_i} \\
\implies \beta_0 + \beta_1 x_i &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\
\implies x_i &= \frac{1}{\beta_1}\left[\log\left(\frac{\pi_i}{1-\pi_i}\right) - \beta_0\right].
\end{aligned}$$

Now if $\pi_i > .5$, $\beta_0 = -36.245$, and $\beta_1 = 4.634$, then $x_i > 7.82$.