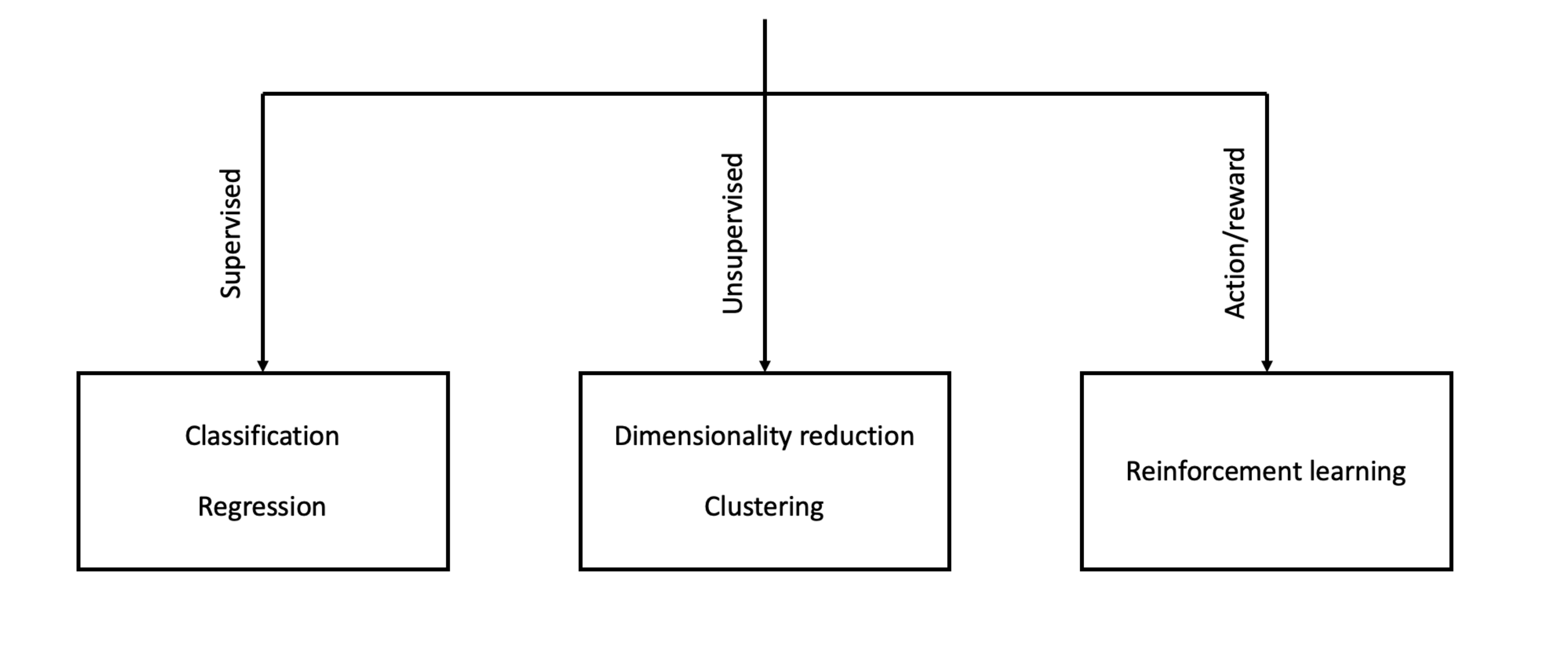
Machine Learning Foundations with Linear and Logistic Regression

Presenter: Micheleen Harris

ML Algorithm Groups



Taking a Closer Look at Regression and Classification with Examples

Linear Regression vs. Logistic Regression

- Linear regression is for quantitative variables
  - A linear, quantitative response is modeled directly
  - Continuous variable(s)
- Logistic regression is for qualitative or categorical variables
  - Classification method
  - Response (Y) is not modeled directly, but rather the probability that Y belongs to a specific category
  - Decision boundaries are linear

Linear regression

Important linear regression assumptions

- There is a relationship between the predictor variable(s) (x_1, x_2, …, x_i) and quantitative response variable (Y)
- For all predictor variables, the relationship to the response variable is linear (however, in some cases non-linear extensions may be used)
- Constant variance (homoscedasticity) in the residual terms ("noise")
- Error terms are normally distributed
- Independent feature variables are not correlated with each other (no collinearity)

Linear regression functions

- Simple linear regression with one variable

$$Y = \beta_0 + \beta_1 x$$

- Multiple linear regression

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

- Multiple linear regression with polynomial regression (extension)

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2$$

A look at the Palmer penguin data

The top five entries:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007
Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN	2007
Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007

- 3 species - 'Chinstrap', 'Gentoo', 'Adelie'
- 3 islands - 'Biscoe', 'Torgersen', 'Dream'

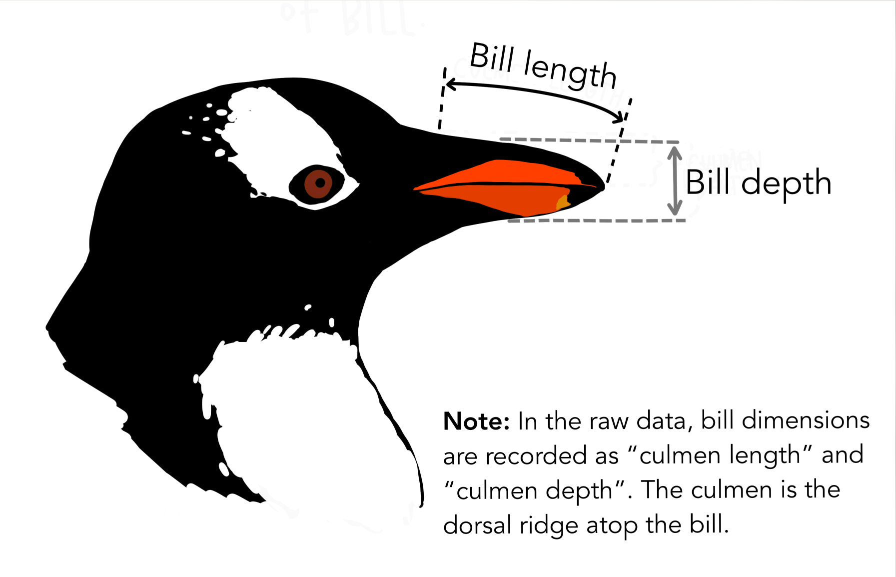
Data source: Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

Linear regression example – estimating the coefficients

- First, we need data. Let's use penguin bill length and bill depth. We want to predict bill depth (Y) from bill length (X).

$$estimate\ of\ bill\_depth = \hat{\beta}_0 + \hat{\beta}_1 * bill\_length$$

- We want to find the "closest" line that fits our data points by getting the best $\hat{\beta}_0$ (the intercept estimate) and $\hat{\beta}_1$ (the slope estimate).
- How is this done?



Artwork by @allison_horst

Aha! This probably reminds you of $y = mx + b$?

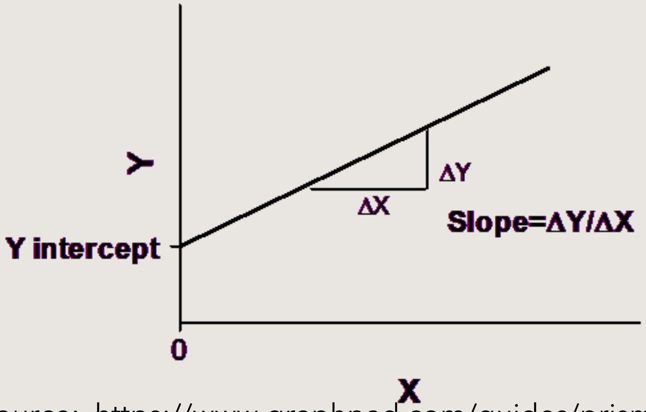


Image source: https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_the_goal_of_linear_regression.htm

Linear regression – estimating the coefficients

estimate of bill_depth = $\hat{\beta}_0 + \hat{\beta}_1 *$ ***bill_length***

One of the most common methods for measuring "closeness" to our data (y's and x's) is the *least-squares* criterion.

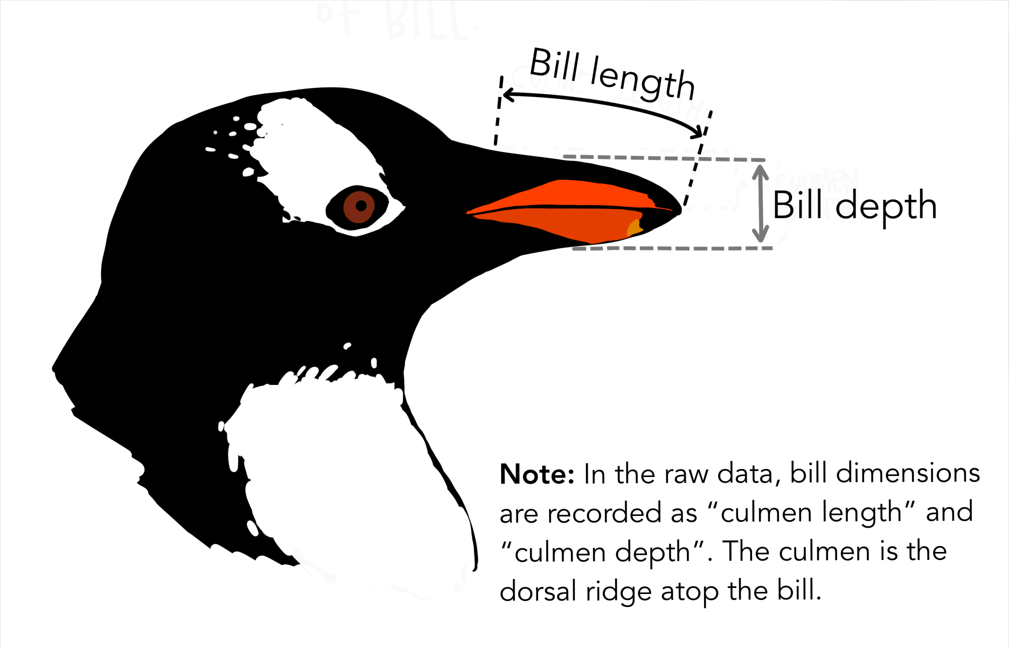
In stats terms, we are actually *minimizing the residual sum of squares*. A residual is $y_i - \hat{y}_i$, or the actual response minus the predicted response. The residual sum of squares is as follows where n is the number of samples.

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 \mathrm{x}_1)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 \mathrm{x}_n)^2$$

We want equations to create an estimate for the slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) that minimize the *residual sum of squares*. They end up looking like:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Artwork by @allison_horst

Where the following are simply the sample means:

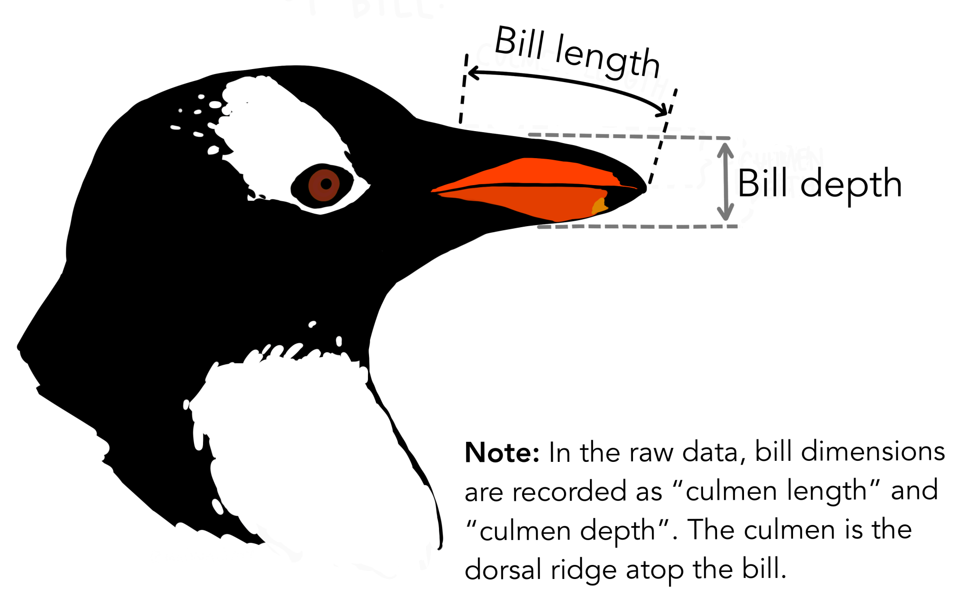
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

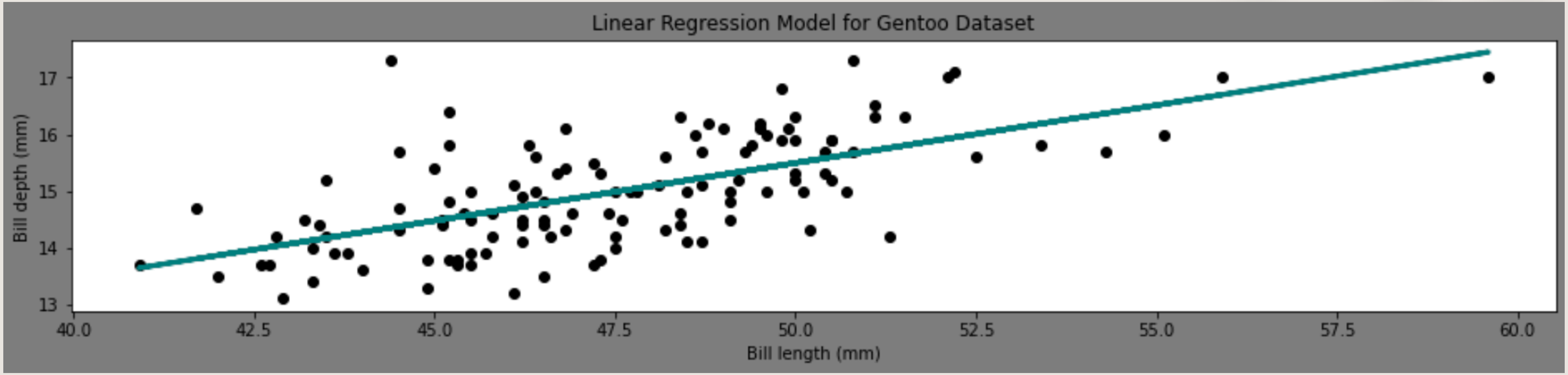
Linear regression example: Gentoo penguin bills

estimate of bill_depth = $\hat{\beta}_0 + \hat{\beta}_1 * $ ***bill_length***

$\hat{\beta}_0 = 5.31$ mm
$\hat{\beta}_1 = 0.204$
$R^2 = 0.41$



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.



Assessing the fit of the model

- Assess the accuracy of the coefficient estimates
  - Compute standard errors
  - Examine the confidence intervals
- Assess the accuracy/fit of the model with RSE
  - Residual standard error – measures *lack of fit* (in units for Y)

$$RSE = \sqrt{\frac{1}{n-2}RSS}$$

Recall:

$$Residual\ sum\ of\ squares(RSS) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Assess the fit of the model with the R^2 statistic – it's proportion-based, so 0-1 in value (1 is a perfect fit!) and uses the RSS and *total sum of squares* (TSS)

$$R^2 = 1 - \frac{RSS}{TSS}$$

where:

$$TSS = \sum(y_i - \bar{y})^2$$

Other topics to look at for linear regression

- null hypothesis vs. alternative hypothesis
- t-statistic
- p-value

Logistic regression

Time to look at the Palmer penguin data again

The top five entries:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007
Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN	2007
Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007

Let's let Y = *sex*; then, X_1 = *bill_depth_mm*

We want the probability of male or female given *bill depth* or written in ML language:

$$\Pr(Y = female \mid X)$$
$$\Pr(Y = male \mid X)$$

Logistic regression assumptions

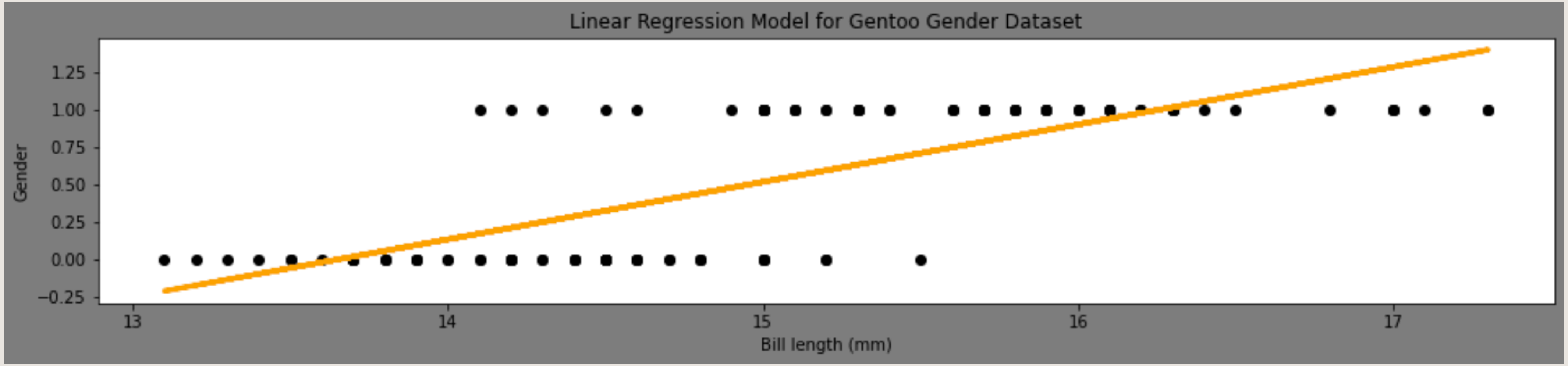
- The response variable is categorical (like the binary male/female) and does not have an ordering
- Logistic regression models binary and multi-class classification problems where decision boundaries are linear
- Data is independent and identically distributed – generated by independent sampling repeatedly from the same distribution.

Use linear regression on a qualitative response?

- Let's take the case for a binary response:

$$Y = \begin{cases} 0 \; if \; female \; penguin \\ 1 \; if \; male \; penguin \end{cases}$$

We could try linear regression and let our model predict female if $\hat{Y} < 0.5$ and male if $\hat{Y} \geq \; 0.5$ (but uh oh, we have negative numbers!). The orange line is our model for the response.



The logistic function – our model

Recall a response with *linear regression* can be written:

$$Y = \beta_0 + \beta_1 x_1$$

The logistic function is as follows where we are modeling probabilities instead!

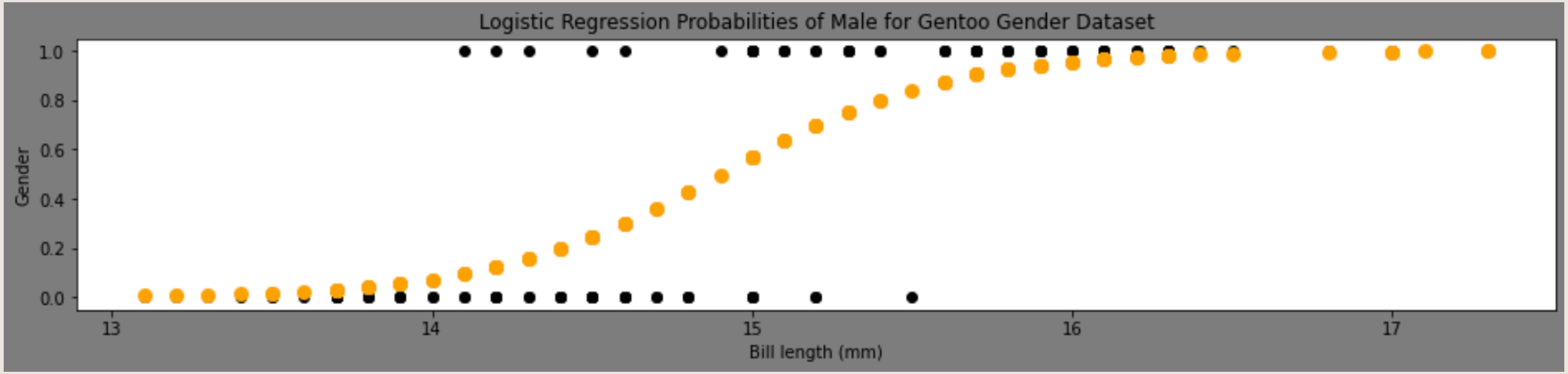
$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Where the probability, $p(X)$, is based on our penguin bill depth data, X.

In logistic regression we care about probabilities (not just a quantitative response like in linear regression) and probabilities should be between 0-1.

Using the logistic function

- Gets us probabilities in the range [0-1]
- For our penguin dataset we can see when we do so and plot these new probabilities (orange) in [0-1] range we see the following (*where 0 is female and 1 is male*):



Finding the coefficient estimates with the maximum likelihood function

- When fitting logistic regression models we use *maximum likelihood* - we try to maximize a function called the maximum likelihood function.
- We use maximum likelihood estimation to find the likelihood estimates for our parameters or coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Now, we want want to be able to plug in the values for our coefficients of our model,

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

such that we get a value close to *0 for female* and close to *1 for male* penguins.

- And the likelihood function, rather out of scope, is as follows:

$$\ell(\beta_0, \beta_1) = \prod_{i:yi=1} p(x_i) \prod_{i\prime:yi_{\prime}=0} (1 - p(x_{i'}))$$

Beware of the curse of dimensionality

- Most classical statistical and machine learning approaches are meant for low-dimensional data where the number of features is much lower than the number of samples.
    - E.g. we have 10,000 bank transactions and 2 features (balance, default status) from a bank statement dataset vs. an image dataset where each image has 1000 features (pixels) and we may only have 100 images.
- Dimensionality reduction algorithms, like principal component analysis (PCA), may be used to transform high dimensional data down to a more manageable number of features.

Tiny bit of interview guidance from experience

- Interviews often look for a solid understanding of the **basics** in stats, classical ML and sometimes deep learning
- Find your best way of learning and go with that (books, videos, tutors, etc.; the math first or the intuition first…)
- Balance becoming a good Python or R programmer with the basics and underlying theory in ML
  - Programming guidance
    - Ask questions and pause to make sure you have **listened** well and understood the problem
    - Have pen and paper ready to brainstorm
    - Write docstrings and comments while coding up a problem
    - Use coding practice platforms to become a sharp programmer (1-2 questions/day)
  - ML guidance
    - Understand the basic theory and intuition
    - Be ready to talk about a school or open-source project that involved some form of DS or ML
- Every interview is practice for the next…

References

- An Introduction to Statistical Learning: with Applications in R by G. James, D. Witten, T. Hastie and R. Tibshirani (https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370/ref=sr_1_1)
- http://www.gatsby.ucl.ac.uk/~porbanz/teaching/W4400S14/W4400S14_01May14.pdf
- https://towardsdatascience.com/assumptions-in-linear-regression-528bb7b0495d