18: INTRO TO GENERALIZED LINEAR MODELS

Prof Amanda Luby

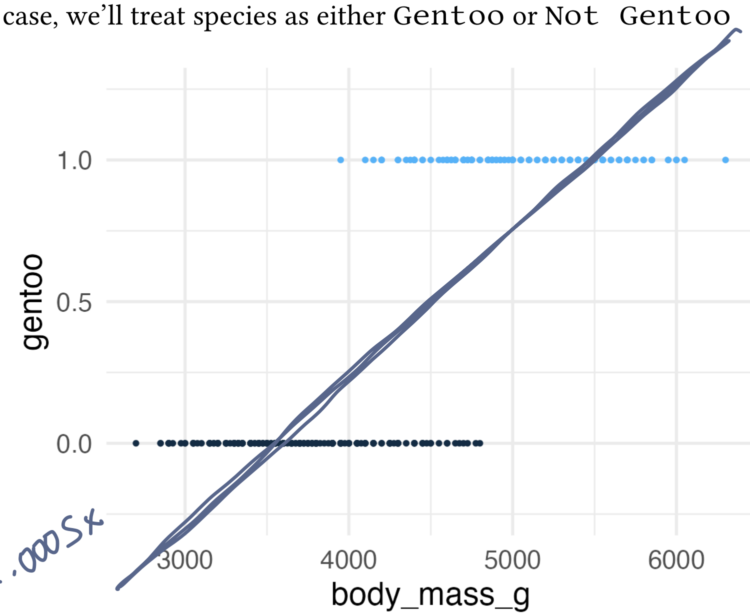
Let's start with our dear old `penguins` friends. The full dataset contains information about three different species of penguins. Rather than understanding the relationship between `body_mass` and `flipper_length`, we might instead be interested in how `body_mass` is related to species. In this case, we'll treat species as either `Gentoo` or `Not Gentoo`

gentoo on average have higher body mass

Data: $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$

x_i: body mass (assumed constant)

Y_i: $\{0, 1\}$ random variable



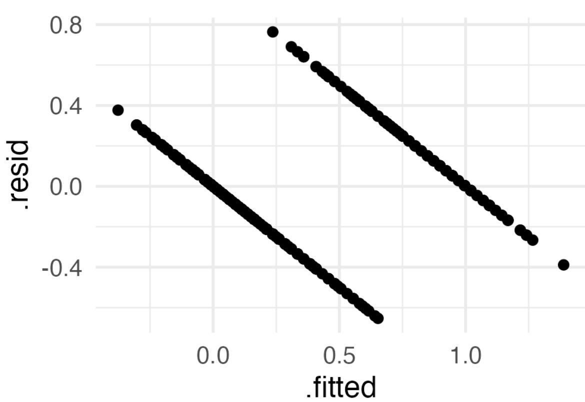
$\hat{y} = -1.7 \times .0005x$

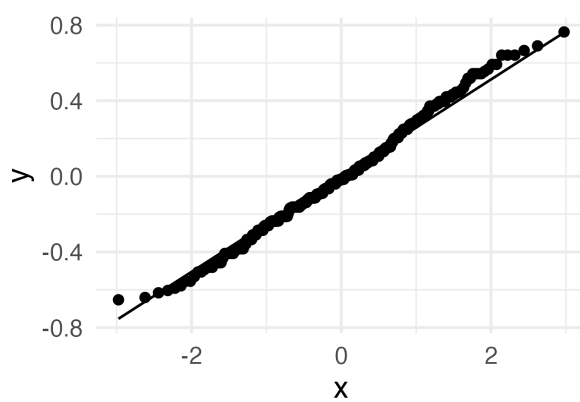
On first glance, it looks like we could go ahead and fit a linear regression model for this problem:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.7006	0.0799	-21.2771	0
body_mass_g	0.0005	0.0000	26.2408	0

residual plots:

QQ plot of $\hat{\epsilon}_i$





Let's list some reasons why this approach is not ideal:

1. Residuals are perfectly correlated w/ predictions $\Rightarrow$ $V(\epsilon_i) \neq \sigma^2 I$

2. How do we assess the equal variance assumption?

3. $\epsilon_i \sim N(0, \sigma^2)$
   $Y_i \sim N(X\beta, \sigma^2)$
   $\uparrow$
   Y_i's are $\{0,1\}$

What distribution does `gentoo` have? A better approach would be to start there.

$Y_i \sim \text{Bernoulli}(p_i)$ $\leftarrow$ rather than modelling Y_i, we model p_i

1. $p_i = \beta_0 + \beta_1 x_i$
   - end up w/ p's < 0 or > 1
   - "diminishing returns" - changes in x matter more if we're close to $1/2$ than if we're far away

2. $\log(p_i) = \beta_0 + \beta_1 x$
   $p_i = e^{\beta_0} e^{\beta_1 x}$
   - only bounded in 1 direction

3. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$
   $p \in [0,1]$
   - not guaranteed to be wrong before we start!

1 Logistic Regression

Logistic Regression Model

$$\underbrace{\log\left(\frac{p_i}{1-p_i}\right)} = \beta_0 + \beta_1 x_i \quad , \quad Y_i \sim \text{Bernoulli}(p_i)$$

$$f(x) = \log\left(\frac{x}{1-x}\right) = \text{"logit"}$$

Solving for p, this gives:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_i$$

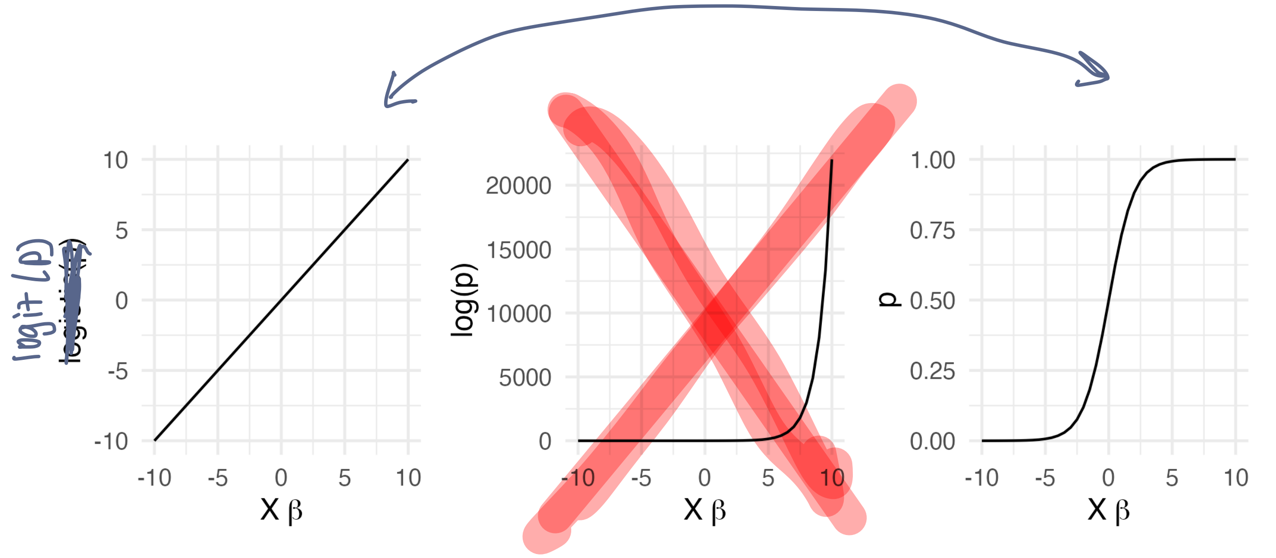
$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_i}$$

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

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

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

$\leftarrow$ logistic function $f(x) = \frac{e^x}{1+e^x}$



MLE's for the linear model relied on the normal PDF for y

1.1 Maximum Likelihood Estimation

Now that we have the structure of the model, we have to think about how to estimate the β's. Recall that the likelihood function for a n Bernoulli random variables is:

$$l(p) = \sum \left[y_i \ln p + (1 - y_i) \ln(1 - p) \right]$$

But, since we now have an X variable, $p = p(x_i)$

$$p = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \qquad \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_i$$

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

$$\begin{aligned} \ell(\beta) &= \sum y_i \ln p + \ln(1-p) - y_i \ln(1-p) \\ &= \sum y_i \ln \frac{p}{1-p} + \ln(1-p) \\ &= \sum y_i (\beta_0 + \beta_1 x_i) + \ln\left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}\right) \\ &= \sum y_i (\beta_0 + \beta_1 x_i) + (-1) \ln(1 + e^{\beta_0 + \beta_1 x_i}) \end{aligned}$$

To find MLE's,

$\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}$, set equal to zero and solve → In general, no closed form solution :(

can be solved numerically w/ Newton-Raphson

Sampling distribution of logistic regression coefficients

$$\hat{\beta}_j \sim N\left(\beta_j, \frac{1}{I_n(\beta_j)}\right)$$ ← since MLE's are approximately normal

"generalized"

```r
gentoo_mod = glm(gentoo ~ body_mass_g,
                 data = penguins,
                 family = "binomial")
summary(gentoo_mod)
```

tells R the distribution we're assuming for Y_i's

```
Call:
glm(formula = gentoo ~ body_mass_g, family = "binomial", data = penguins)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.842e+01  3.609e+00  -7.873 3.46e-15 ***
body_mass_g  6.371e-03  8.131e-04   7.835 4.69e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 446.80  on 341  degrees of freedom
Residual deviance: 117.85  on 340  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 121.85

Number of Fisher Scoring iterations: 7
```

$H_0: \beta_j = 0$

$\Sigma(y_i - \bar{y})^2$

$\Sigma(y_i - \hat{y}_i)^2$

~~1.2 Interpretation of coefficients~~

2 Generalized Linear Models

We've now seen two different settings for regression. If X is a vector of predictors and $Y \in \mathbb{R}$, we have assumed a linear model:

$$Y \sim N(X\beta, \sigma^2)$$

and if $Y \in \{0, 1\}$, we assumed a logistic model:

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = X\beta$$

In both settings, we are assuming that a transformation of the conditional expectation is a linear function of X:

Linear

$$E[Y_i | X_i] = \mu_i = X\beta$$

transformation: identity
$g(x) = x$

Logistic

Binomial RV's: $E[Y] = \mu = p$

$$\log\left(\frac{E(Y_i | X_i)}{1 - E(Y_i | X_i)}\right) = X\beta$$

transformation: logit
$g(x) = \log\left(\frac{x}{1-x}\right)$

Recall that exponential families of distributions can be written as

$$f(x;\theta) = h(x)g(\theta)\exp(T(x)\eta(\theta))$$

$T(x)$: sufficient statistic

$\eta(\theta)$: "natural parameter"

Normal

$T(x) = \Sigma Y_i$

$\eta(\theta) = \mu$ (assuming σ^2 known)

$=$ identity function

Bernoulli

$T(x) = \Sigma Y_i$

$\eta(\theta) = \log\left(\frac{p}{1-p}\right)$

$=$ logit function

Can set up a GLM for any exponential family, where $\eta(\theta)$ tells us what $g(E(Y_i | X_i))$ should be.