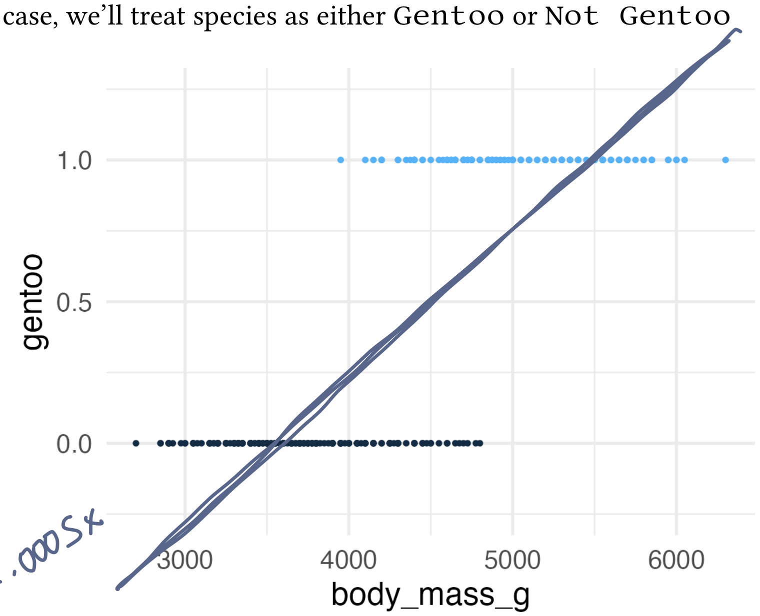
18: INTRO TO GENERALIZED LINEAR MODELS

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)

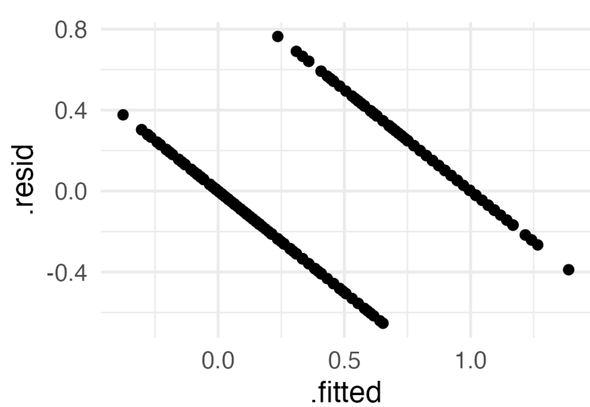
Y_i: $\{0, 1\}$ random variable

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

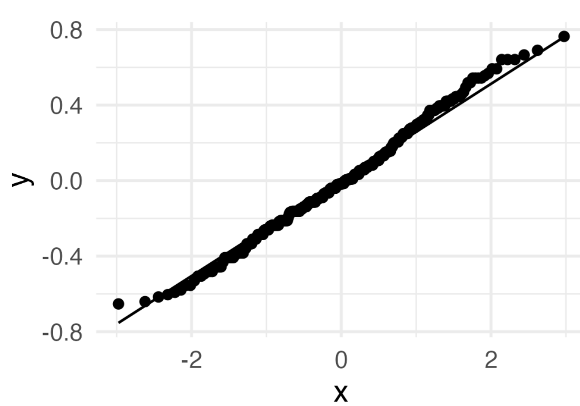
On first glance, it looks like we could go ahead and fit a linear regression model for this problem:

```
             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

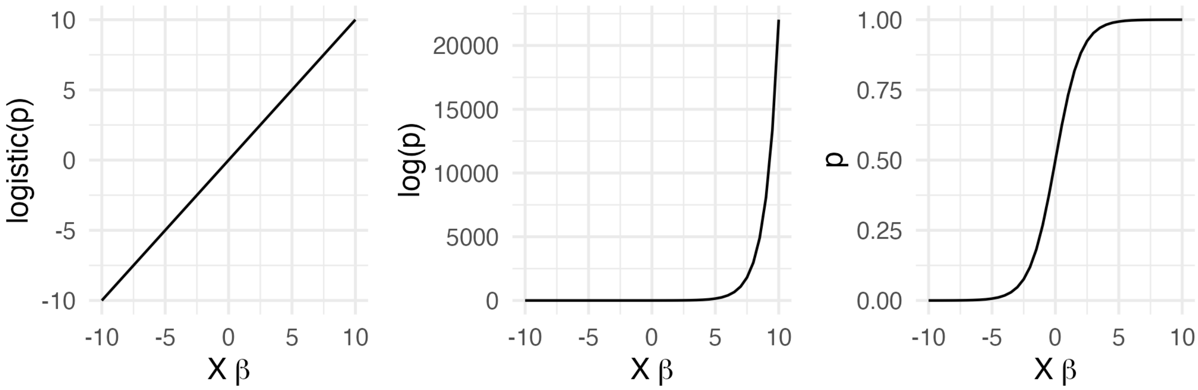
3. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ - not guaranteed to be wrong before we start!

   $p \in [0,1]$

1 Logistic Regression

Logistic Regression Model

Solving for p, this gives:



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:

$$l(p) = \sum y_i \ln p + (1 - y_i) \ln(1 - p)$$

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

Sampling distribution of logistic regression coefficients

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

```
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
```

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:

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

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