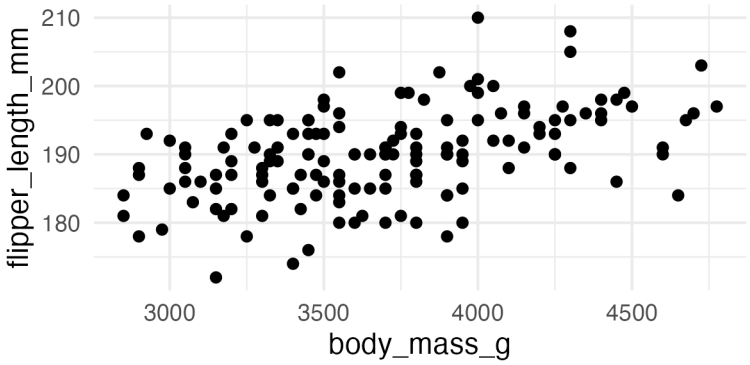
25: LEAST SQUARES

Stat250 S25

Prof Amanda Luby

Example: Let's return to the Adelie penguins that we've seen before. The Long Term Ecological Research Network (LTER) has collected data on a group of Adelie penguins, including their body mass, flipper length, bill length, and bill depth. It's relatively easy to measure their body mass, but harder to get accurate measurements of their flipper length. We generally expect heavier penguins will have longer flippers, but the researcher would like to know (a) how strong this relationship is and (b) the magnitude of this relationship.



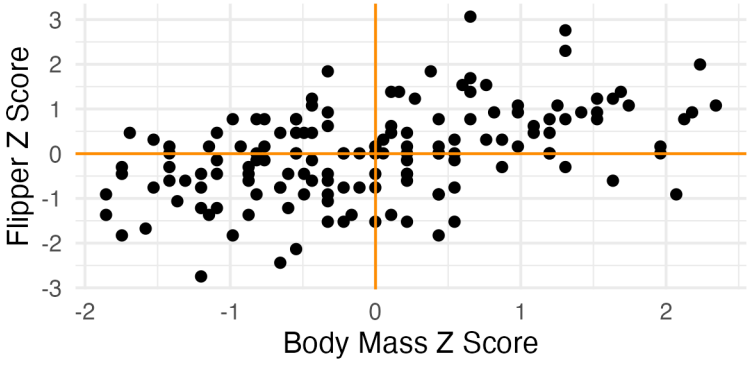
Notation:

Covariance

$$Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

Correlation

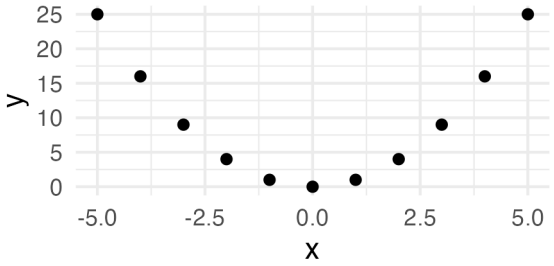
$$\rho(X,Y) =$$



Watch out!

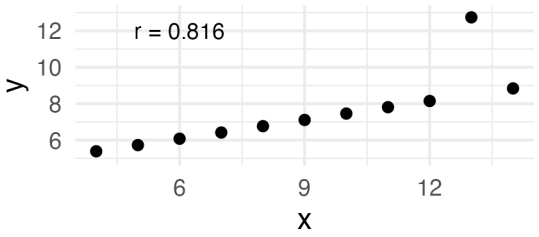
Correlation only measures the strength of **linear** functions

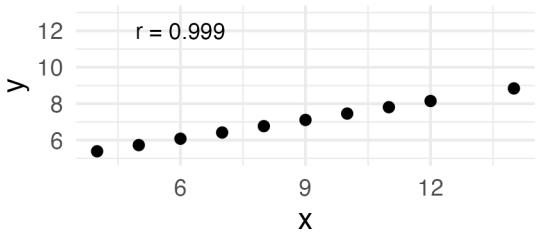
```
x <- -5:5
y <- x^2
cor(x, y)
#> [1] 0
gf_point(y ~ x)
```



Watch out!

Correlation is **sensitive to outliers**





1 Least Squares Line

- Goal: predict flipper length based on body mass using a linear function:

$$\hat{y} = \hat{a} + \hat{b}x_i$$

- Approach: minimize the **residual sum of squares**:

In math:

Least Squares Estimates

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

In R:

```
adelie_fit <- lm(flipper_length_mm ~ body_mass_g, data = adelie)
adelie_fit
#>
#> Call:
#> lm(formula = flipper_length_mm ~ body_mass_g, data = adelie)
#>
#> Coefficients:
#> (Intercept)  body_mass_g
#>   1.652e+02    6.677e-03
```

Interpretations:

$\hat{a}$:

$\hat{b}$:

2 "Nonlinear" Least Squares

Obviously, not every relationship can be adequately described by a straight line. BUT linear models are very "nice" with "easy" solutions (as we saw above). Luckily, we can "linearize" many nonlinear relationship by transforming the X or Y variable.

Example: We suspect that the true relationship between variables X and Y is described by $y = ae^{bx}$. What are the transformations of X and Y that "linearize" this relationship?

Exercise: Fill in the following table to show that all of these nonlinear relationships can be expressed as linear functions of transformations of the original variables.

True Relationship	Transformation of Y	Transformation of X
$y = a + bx^2$		

True Relationship	Transformation of Y	Transformation of X
$y = ae^{bx}$		
$y = ax^b$		
$y = \frac{1}{a+bx}$		
$y = \frac{x}{a+bx}$		
$y = \frac{1}{1+\exp(a+bx)}$		
$y = 1 - e^{-x^b/a}$		

3 Simple Linear Regression Model

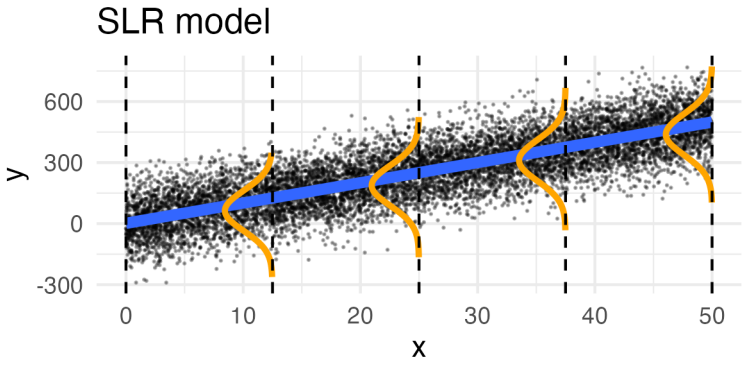
Everything we've talked about up until this point has not used any statistical properties at all: there have been no probability distributions, expectations, independence assumptions, etc. We've gone about "fitting curves" as a purely geometric exercise.

Example: Suppose we observe an Adelie penguin that weighs 4,000g.

- What is the prediction for that penguin's flipper length?
- Would every 4,000g penguin have that flipper length?

Simple Linear Regression Model

- Least Squares only assumes that there is a linear relationship between x and y
- The SLR model adds assumptions that can be written in a few forms:
  - $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$
  - $Y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$



LINE Mnemonic:

1. **Linearity**: $E(Y_i|X = x_i) = \mu_i = \beta_0 + \beta_1 x_i$
2. **Independence**: $\varepsilon_1, \dots, \varepsilon_n$ are independent
3. **Normal error terms**: $\varepsilon_i \sim N(0, \sigma^2)$
4. **Equal error variance**: $Var(\varepsilon_1) = \cdots = Var(\varepsilon_n) = \sigma^2$

Note: This model also assumes that the x variables are **fixed**, which is why they get little x's instead of big X's

How many parameters must be estimated?

MLE's:

Maximum Likelihood Estimators for SLR

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

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$$

Note: