14: LEAST SQUARES REGRESSION

Larsen & Marx 11.1-11.3
Prof Amanda Luby

---

Up to this point, we have largely concerned ourselves with **univariate settings**. That is, we observe one sample, $X_1, ..., X_n$, and wish to draw a conclusion about some parameter or estimator. This setting is actually quite restrictive: we rarely are interested in solely one random variable. Most research questions are instead interested in how various components of a complex system are related to one another: how is cancer incidence related to diet, genetics, pollution, or behaviors? How do salaries for new grads vary depending on degree, internship experience, industry, gender, or race?

In order to answer these types of questions, we have to extend our statistical toolbox to include *multivariate* samples. This week, we're going to focus on building up the theory for analyzing the relationship between two variables. Rather than assuming we have a sample of $x_1, x_2, ..., x_n$, we're going to assume we have a bivariate sample $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

1 Method of Least Squares

If we draw a scatterplot of our bivariate sample, we might obtain a graph something like one of the below. What do each of the graphs tell you about the *direction* and *strength* of the relationship? What are you looking for?
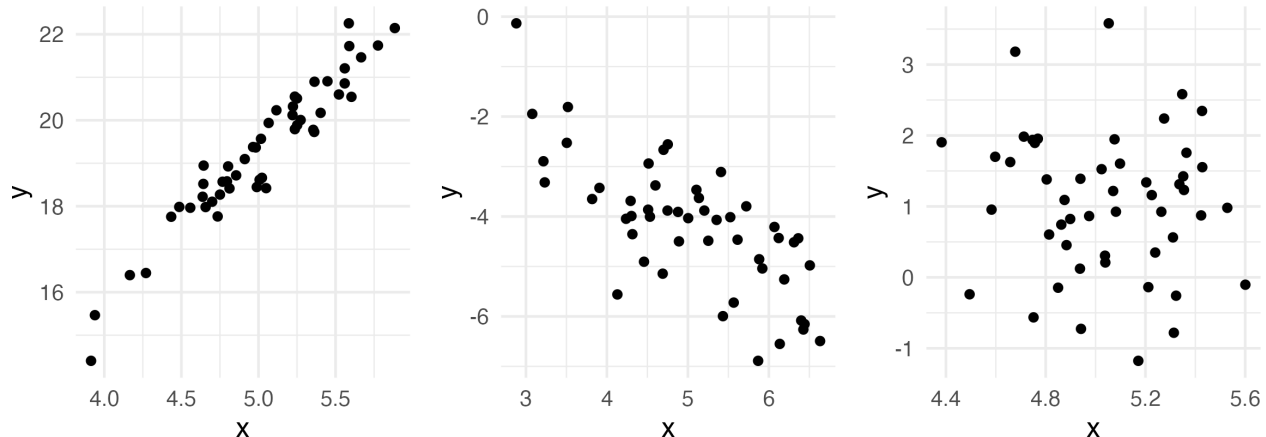
> **Least Squares Line**
>
> Given n points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the straight line $y = a + bx$ minimizing
>
> $$L = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$$
>
> is given by:
>
> $$b =$$
>
> $$a =$$

Proof:

> **Residual**
>
> Let a and b be the least squares coefficients associated with the sample $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. For any x, the quantity $\hat{y} = a + bx$ is the *predicted value* of y. For any i, the difference
>
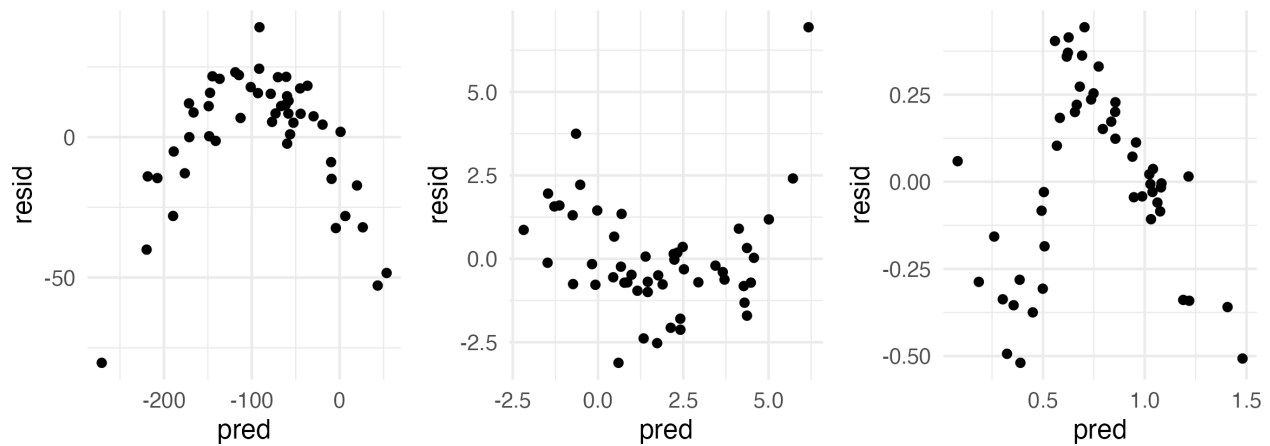> $$y_i - \hat{y}_i = y_i - (a + bx_i)$$
>
> is called the

As statisticians, we often gauge the appropriateness of the least squares line using *residual plots*.

Example: Here are the residual plots after fitting the least squares line to the three plots above. What do you notice?
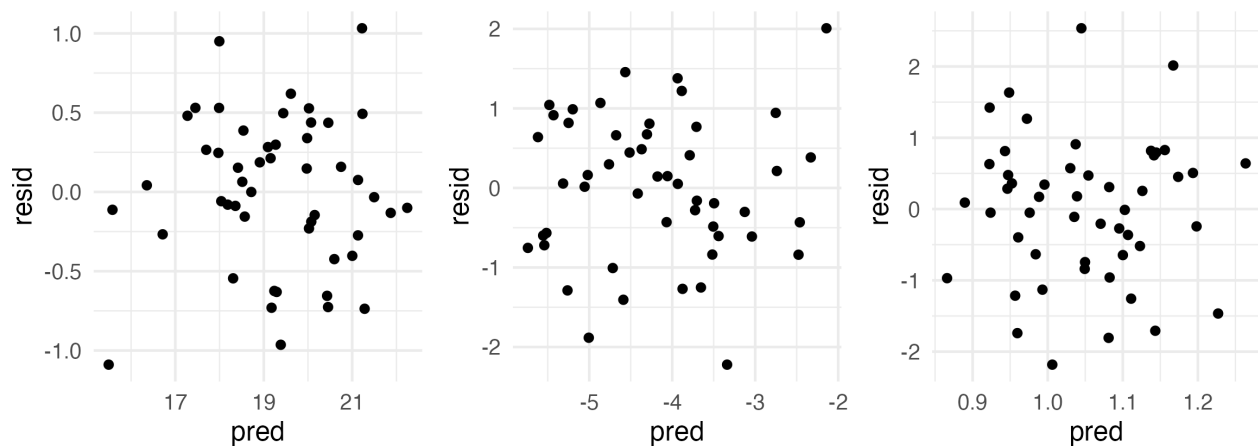
Below are three additional residual plots. What do you suspect about the original X-Y scatterplots?

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.

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
$y = a + bx^2$		
$y = ae^{bx}$		
$y = ax^b$		
$y = \frac{1}{1+\exp(a+bx)}$		
$y = \frac{1}{a+bx}$		
$y = \frac{x}{a+bx}$		
$y = 1 - e^{-x^b/a}$		

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: Gentoo penguins are a species of penguin. The Long Term Ecological Research Network (LTER) has collected data on a group of Gentoo 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. The researchers would like to know how body mass is related to flipper length, and specifically whether they could predict flipper length using body mass alone.

$y = 137 + 0.015 \cdot x$

(plot axes: flipper_length_mm vs body_mass_g)
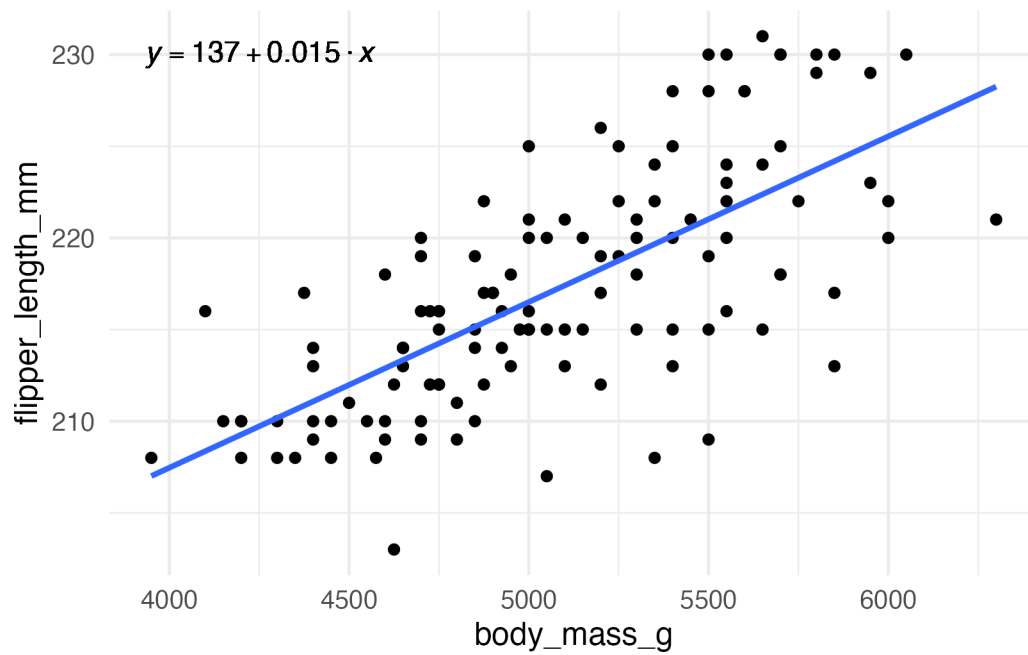
Regression model

Let $f_{Y|X=x}(y)$ denote the PDF of the random variable Y for a given value of $X = x$. Then the function

is called the *regression model* of Y on x.

Simple Linear Regression (SLR) model

Let $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ be a set of points satisfying $E(Y|X = x) = \beta_0 + \beta_1 x$. The MLE's for β_0, β_1, and σ^2 are given by:

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

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

$$\hat{\sigma}^2 = \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i =$

Proof: