· Quiz 3 back today
· Wed: final project proposal + lab 7 due
· OH today 11:30-12:30
wed 2:30-4

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, $X_1, \ldots, X_n$, 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 $x_1, x_2, \ldots, x_n$, we're going to assume we have a bivariate sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

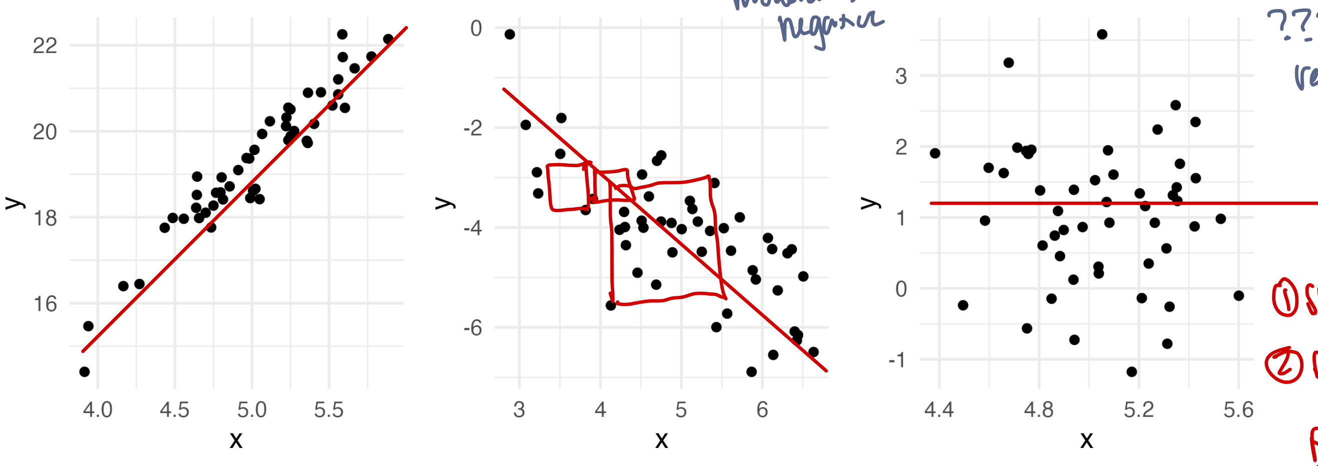
1 Method of Least Squares

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?

strong, positive

moderate, negative

??? no relationship



① slope
② spread of points around the line

Least Squares Line

Given n points $(x_1, y_1), (x_2, y_2), \ldots., (x_n, y_n)$, the straight line $y = a + bx$ minimizing

$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

is given by:

$$b = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x}$$

$$L = \sum [y_i - (a+bx_i)]^2 \qquad b = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x}$$

Proof:

$$\frac{\partial L}{\partial a} = \sum (-2)[y_i - (a+bx_i)]$$
$$= \sum(-2y_i + 2a + 2bx_i)$$
$$0 = -\sum y_i + na + b\sum x_i$$
$$\sum y_i - b\sum x_i = na$$
$$a = \frac{1}{n}\sum y_i - b\cdot\frac{1}{n}\sum x_i$$
$$= \bar{y} - b\bar{x}$$
$$= \frac{\sum y_i - b\sum x_i}{n}$$

$$\frac{\partial L}{\partial b} = \sum (-2)x_i[y_i - (a+bx_i)]$$
$$= \sum(-2x_iy_i + 2ax_i + 2bx_i^2)$$
$$0 = -\sum x_iy_i + a\sum x_i + b\sum x_i^2$$
$$\sum x_iy_i - a\sum x_i = b\sum x_i^2$$
$$b\sum x_i^2 = \sum x_iy_i - \left(\frac{\sum y_i - b\sum x_i}{n}\right)\sum x_i$$
$$nb\sum x_i^2 = n\sum x_iy_i - (\sum y_i)(\sum x_i) + b(\sum x_i)^2$$
$$b(n\sum x_i^2 - (\sum x_i)^2) = n\sum x_iy_i - \sum x_i\sum y_i$$
$$b = \frac{n\sum x_iy_i - \sum x_i\sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Residual

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

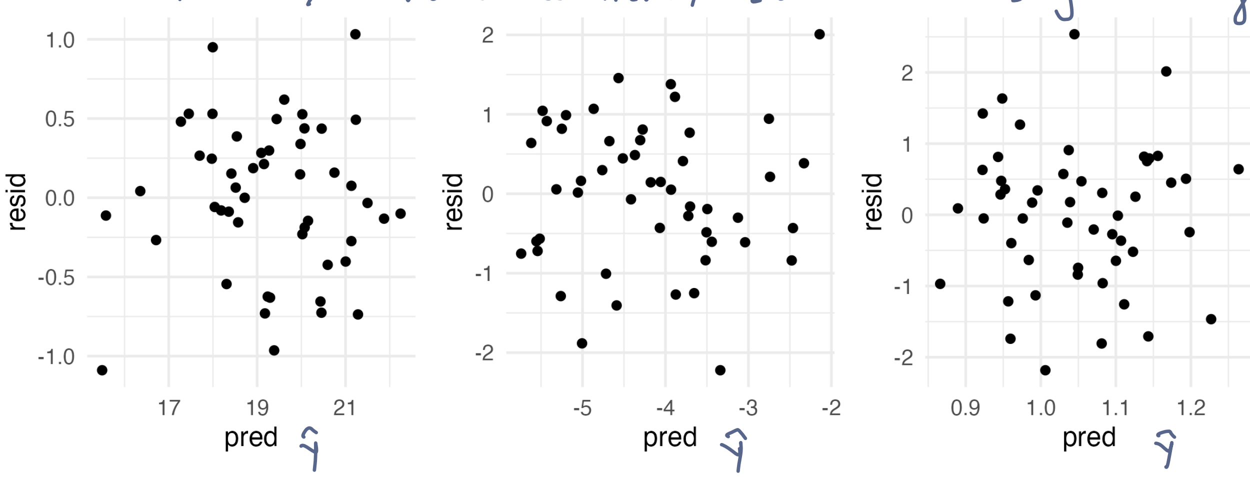
predictions

$$y_i - \hat{y}_i = y_i - (a + bx_i) = \varepsilon_i$$

is called the i^{th} residual

As statisticians, we often gauge the appropriateness of the least squares line using *residual plots*. → $\hat{y}$ on x-axis, ε_i on y-axis

Example: Here are the residual plots after fitting the least squares line to the three plots above. What do you notice? Random, no noticeable trend, "cloud"-like → good thing!

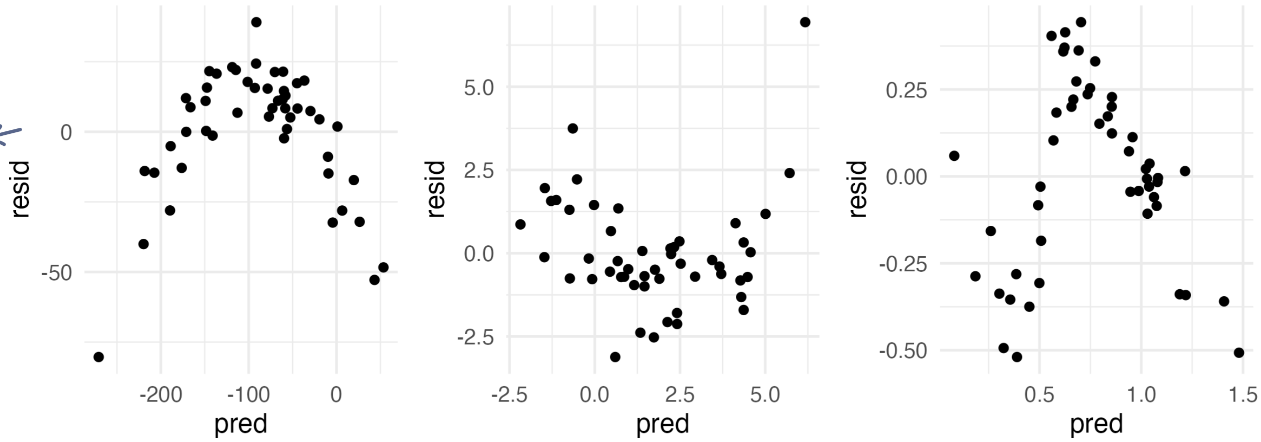


After accounting for the least squares line, there's no additional relationship between X & Y

Below are three additional residual plots. What do you suspect about the original X-Y scatterplots?

Not linear, not cloud like

Regions of data that are systematically above + below the least squares line



"Residual Plots" ⇔ "diagnostic plots" → "diagnose" nonlinearity

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. $y^* = a^* + b^* x^*$

True Relationship	Transformation of Y	Transformation of X
$y = a + bx^2$	y	x^2
$y = ae^{bx}$ → $\ln y = \ln a + bx$	$\ln y$	x
$y = ax^b$	$\ln y$	$\ln x$
$y = \frac{1}{1+\exp(a+bx)}$	$\ln\left(\frac{1-y}{y}\right)$	x
$y = \frac{1}{a+bx}$	$1/y$	x
$y = \frac{x}{a+bx}$	$1/y$	$1/x$
$y = 1 - e^{-x^b/a}$	$\ln \ln\left(\frac{1}{1-y}\right)$	$\ln x$

(4) $\frac{1}{y} = 1 + \exp(a+bx)$

$\frac{1}{y} - 1 = \exp(a+bx)$

$\ln\left(\frac{1-y}{y}\right) = a + bx$

Today:
- Quiz 3
- Notes 14
- Notes 15

- Lab 7
- "Project Proposal" due on Gradescope

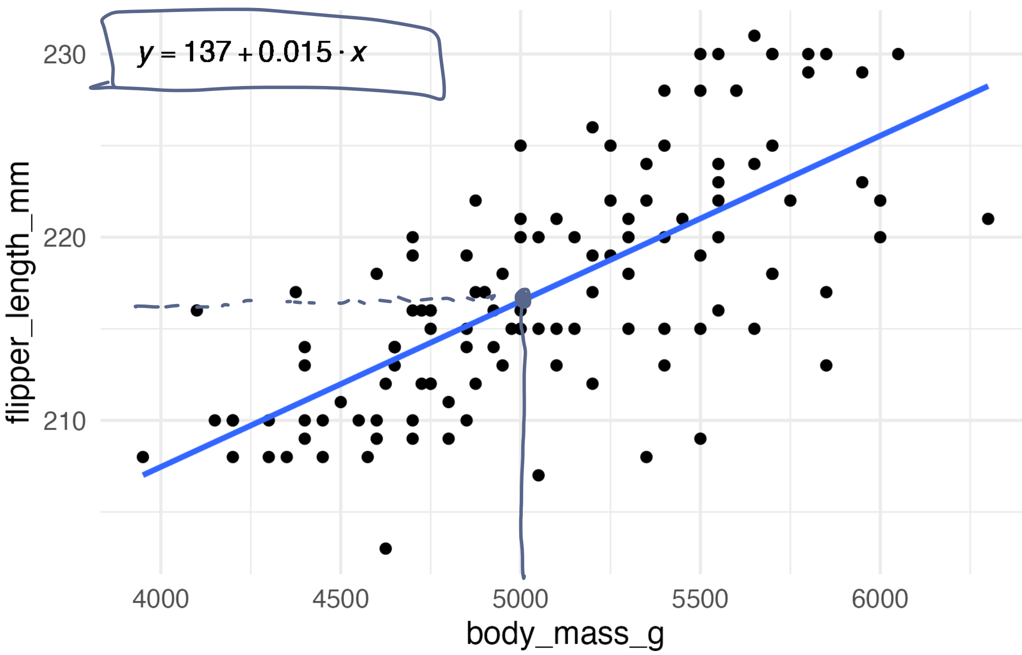
OH today 2:30-4

Post HW tonight/tomorrow

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.



Say we observe a 5,000 g penguin:

$\hat{y} = 137 + .015 \cdot 5{,}000$

≈ 216

would every 5,000 g penguin have flipper length 216?

NO

least squares line captures average relationship, but we also care about variance

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

$$y = E(Y \mid X = x) = E(Y \mid x)$$

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

* for the most part, we assume $X=x$ are constants instead of RV's

Additional Assumptions:

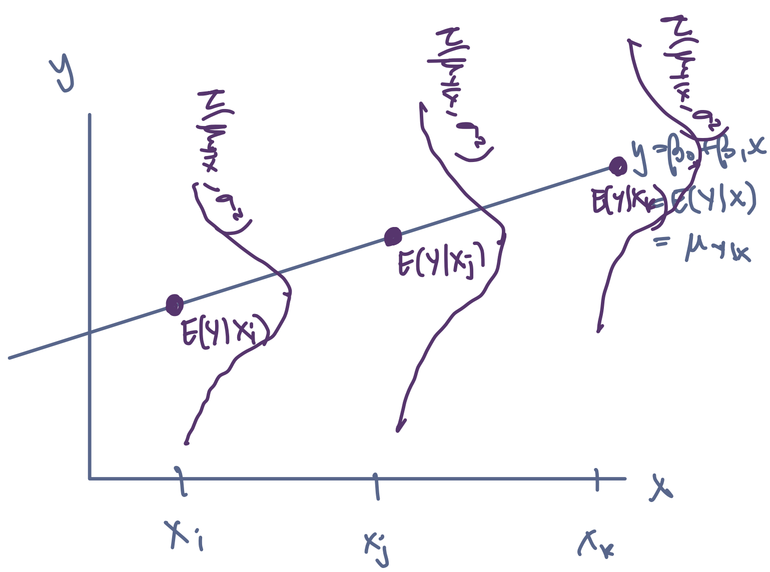
1. $f_{Y|x} \sim N(\mu, \sigma^2) \quad \forall x$

2. σ is constant for all x

3. $\mu = E(Y|x) = \beta_0 + \beta_1 x$

4. All conditional distributions are independent

$E(Y|4000) \perp E(Y|5000)$



Population parameters: $\beta_0, \beta_1, \sigma^2$

Estimators: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$

Simple Linear Regression (SLR) model

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ 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 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Proof: $\boxed{Y_i | X_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)}$

$$L = \prod_{i=1}^{n} f_{y|x} = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-1/2\sigma^2 \sum(y_i - (\beta_0 + \beta_1 x_i))^2}$$

$$\ell = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\frac{\partial \ell}{\partial \beta_0} = -\frac{1}{\sigma^2}\sum(y_i - (\beta_0 + \beta_1 x_i))(-1) = 0$$

$$\sum(y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum(y_i - (\beta_0 + \beta_1 x_i))(-x_i) = 0$$

$$\sum(y_i - (\beta_0 + \beta_1 x_i))(x_i) = 0$$

$\hat{\beta}_0$ + $\hat{\beta}_1$ are the least squared estimates!

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2\pi\sigma^2} \cdot 2\pi \cdot \frac{n}{2} - \frac{1}{2(\sigma^2)^2}\sum(y_i - (\beta_0 + \beta_1 x_i))^2 = 0$$

$$\frac{n}{\sigma^2} - \frac{1}{(\sigma^2)^2}\sum(y_i - (\beta_0 + \beta_1 x_i))^2 = 0$$

let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

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

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