

“This is statistics”

by **Dr. Genevera Allen**


Associate Professor at Rice University

https://www.youtube.com/watch?v=xURkTKDq_M

1

Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM


$$y = a + bx$$


Ch 4: Simple Estimation

Multiple Regression Analysis

Too complicated by hand!

Ouch!



44

2

General linear models (not Generalized linear model)

Linear Model	Common name
$Y = \mu + X$	Simple linear regression
$Y = \mu + A_1$	One-factorial (one-way) ANOVA
$Y = \mu + A_1 + A_2 + A_1 \times A_2$	Two-factorial (two-way) ANOVA
$Y = \mu + A_1 + X (+A_1 \times X)$	Analysis of Covariance (ANCOVA)
$Y = \mu + X_1 + X_2 + X_3$	Multiple regression
$Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
$Y_1 + Y_2 = \mu + A_1 + A_2 + A_1 \times A_2$	Multivariate ANOVA (MANOVA)

Y (response) is a continuous variable
 X (predictor) is a continuous variable
 A represents categorical predictors (factors)
 g represents groups of data (more on this later)
 (+A₁ × X) - step 1 on an ANCOVA, but not in the final analysis
 Multiple factors A₁ + A₂ + etc (and their interactions)

3

Multiple regression – the “model of all models”!

Part I:

Causation, regression model, properties of estimators and sensibility to assumptions

Part II:


Goodness of fit and model simplicity metrics, hypotheses testing, standardized slopes, model selection, examples and diagnostics

4

Multiple regression – the “model of all models”!

The essential idea with regression models is to find driving forces like the train engine and determine the path of the railway track.

The “driving force” in statistics is often called “generating process”



5

Correlation, Causation, & Coincidence

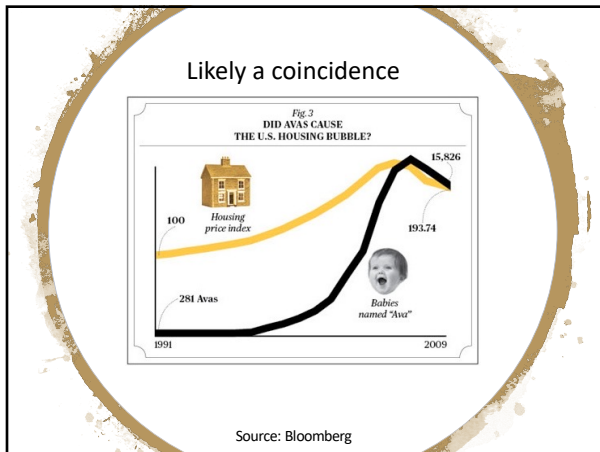
One of the key concepts in regression models, or science in general, is to distinguish between correlation and causation.

source - <http://ucanalytics.com/>

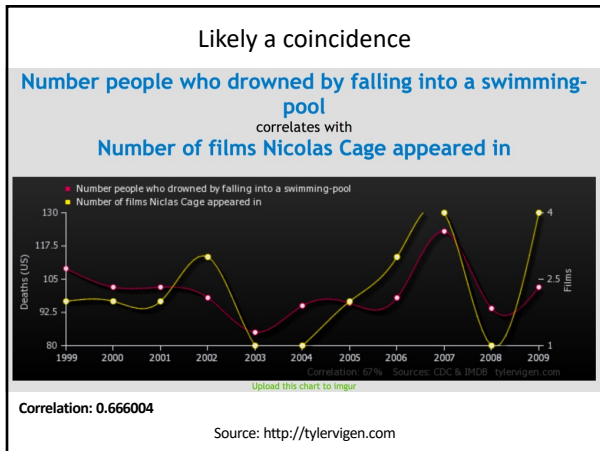
Unless in experimental settings and in some time series (and even then), regression models cannot necessarily distinguish between causation and correlation.

The role of researchers when using regression is to provide strong evidence and a narrative of causation (even though it can't always be confirmed).

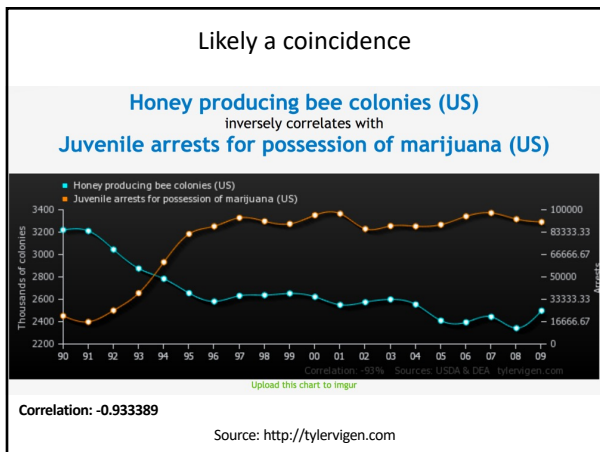
6



7



8



9

Coincidence =
spurious correlations

http://tylervigen.com/discover?type_select=fun

10

Likely a correlation

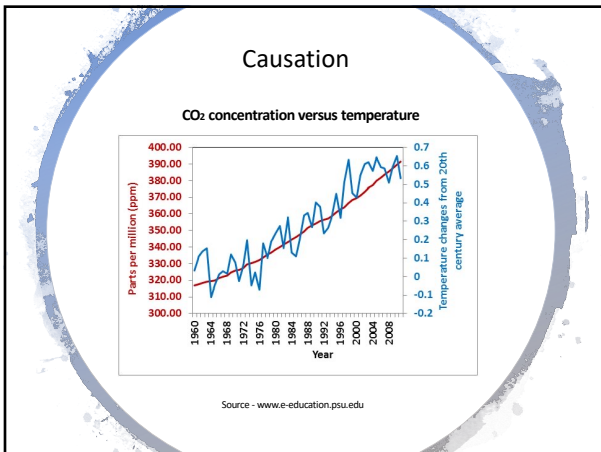
Source: Bloomberg

11

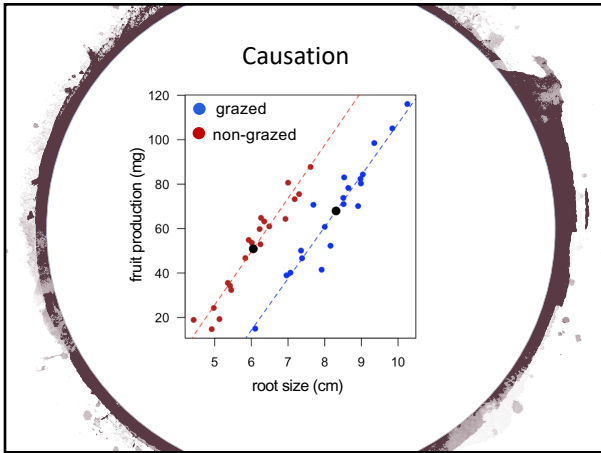
Correlation

Source: Nature

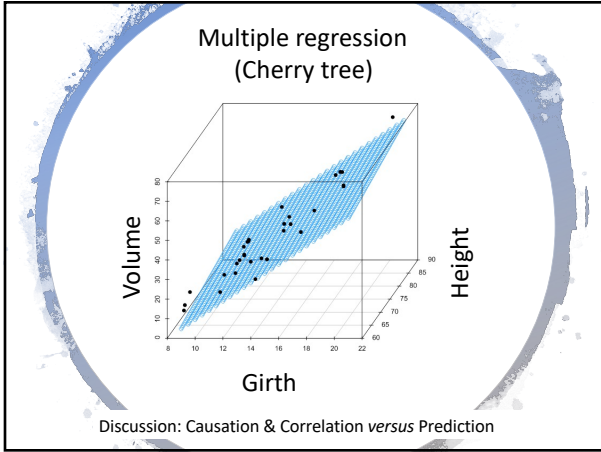
12



13



14



15

Some thoughts on « explanation »

In 1964, during a lecture at Cornell University, the physicist Richard Feynman articulated a profound mystery about the physical world. He told his listeners to imagine two objects, each gravitationally attracted to the other. How, he asked, should we predict their movements? Feynman identified three approaches, each invoking a different belief about the world.

source – The New Yorker

16

Some thoughts on « explanation »

In 1964, during a lecture at Cornell University, the physicist Richard Feynman articulated a profound mystery about the physical world. He told his listeners to imagine two objects, each gravitationally attracted to the other. How, he asked, should we predict their movements? Feynman identified three approaches, each invoking a different belief about the world.

- 1) The first approach used Newton’s law of gravity, according to which the objects exert a pull on each other.
- 2) The second imagined a gravitational field extending through space, which the objects distort.
- 3) The third applied the principle of least action, which holds that each object moves by following the path that takes the least energy in the least time.

source – The New Yorker

17

Some thoughts on « explanation »

In 1964, during a lecture at Cornell University, the physicist Richard Feynman articulated a profound mystery about the physical world. He told his listeners to imagine two objects, each gravitationally attracted to the other. How, he asked, should we predict their movements? Feynman identified three approaches, each invoking a different belief about the world.

- 1) The first approach used Newton’s law of gravity, according to which the objects exert a pull on each other.
- 2) The second imagined a gravitational field extending through space, which the objects distort.
- 3) The third applied the principle of least action, which holds that each object moves by following the path that takes the least energy in the least time.

All three approaches produced the same, correct prediction. They were three equally useful descriptions of how gravity works. “One of the amazing characteristics of nature is this variety of interpretational schemes,” Feynman said.

source – The New Yorker

18



19

Multiple regression – the “models of all models”!

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

β_0 model intercept (or constant)

$\beta_1, \beta_2, \dots, \beta_p$ Partial regression coefficients (or partial slopes)

e model residuals or error

The general purpose of *multiple regression* are:

- 1) Describe, investigate and learn about the relationship between several independent or predictor variables and a dependent variable.
- 2) Make predictions.
- 3) Plan experiments to test causality (in regression, causality is often implied).

20

Multiple regression – the “models of all models”!

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

β_0 model intercept (or constant)

$\beta_1, \beta_2, \dots, \beta_p$ Partial regression coefficients (or partial slopes)

e model residuals or error

Fitting method = Ordinary least square (OLS)

The OLS method minimizes the sum of square differences between the observed and predicted values.

21

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42cm + \beta_1 X_1 + \beta_2 X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)

β_0

- Model intercept (or constant) is the value that is predicted for Y if predictors X₁ and X₂ are zero, i.e., the expected plant height if there is no bacteria in the soil and no sun light.

22

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42cm + \beta_1 X_1 + \beta_2 X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)

β_0

- Model intercept (or constant) is the value that is predicted for Y if predictors X₁ and X₂ are zero, i.e., the expected plant height if there is no bacteria in the soil and no sun light.
- This is only a reasonable interpretation if either X₁ and X₂ can be zero and if the data include values for X₁ and X₂ that are closer to zero). For instance, the intercept could be negative for this model even though a plant can't have negative height.
- The unit of the intercept is the same as the response variable (i.e., cm).

23

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42cm + 2.3X_1 + \beta_2 X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)

β_1

- It represents the difference in predicted value of Y (plant height) for each one unit difference in bacteria amount if sun exposure is kept constant (i.e., as if plants were exposed to the same amount of mean sun light) – called partial effects/slopes
- Plants with 5000/ml bacteria counts would, on average, be 2.3 cm taller (in average) than plants in soils with 4000/ml (which would be 2.3 cm taller in average than plants with 3000/ml).

The slope of any single partial regression line (partial regression slope) represents the rate of change or effect of that specific predictor variable (holding all the other predictor variables constant to their respective mean values) on the response variable.

24

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42\text{cm} + 2.3X_1 + \beta_2X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)

β_1

Represents the difference in predicted value of Y (plant height) for each one unit difference in bacteria amount if sun exposure is kept constant (i.e., as if plants were exposed to the same mean amount of sun light).

25

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42\text{cm} + 2.3X_1 + \beta_2X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)

β_1

- It represents the difference in predicted value of Y (plant height) for each one unit difference in bacteria amount if sun exposure is **kept constant** (i.e., as if plants were exposed to the same amount of sun light).
- Plants with 5000/ml bacteria counts would, on average, be 2.3 cm taller (in average) than plants in soils with 4000/ml (which would be 2.3 cm taller in average than plants with 3000/ml).
- “**Kept constant**” means that that the association between bacterial amount and plant height is independent (controlled for) of amount of sun.

26

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42\text{cm} + 2.3X_1 + \beta_2X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)

β_1

- It represents the difference in predicted value of Y (plant height) for each one unit difference in bacteria amount if sun exposure is **kept constant** (i.e., as if plants were exposed to the same amount of sun light).
- Plants with 5000/ml bacteria counts would, on average, be 2.3 cm taller (in average) than plants in soils with 4000/ml (which would be 2.3 cm taller in average than plants with 3000/ml).
- “**Kept constant**” means that that the association between bacterial amount and plant height is independent (controlled for) of amount of sun.
- **The unit attached to the slope is the unit of the response divided by the unit of the predictor (i.e., cm/ 1000 bacteria per ml)**

27

A small fictional example to facilitate understanding of what regression coefficients mean!

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (1000 bacteria per ml of soil)
 X₂ is amount of plant exposure to sun light (% exposure)


β_1

- It represents the difference in predicted value of Y (plant height) for each one unit difference in bacteria amount if amount of sun is kept constant (i.e., as if plants were exposed to the same amount of sun light).
- Plants with 5000/ml bacteria counts would, on average, be 2.3 cm taller (in average) than plants in soils with 4000/ml (which would be 2.3 cm taller in average than plants with 3000/ml).
- "Kept constant" means that the association between bacterial amount and plant height is independent (controlled for) of amount of sun.

β_2 Reverse interpretation in relation to β_1
 Units attached - cm / % exposure

28

What do model slopes represent?



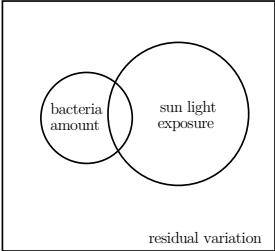
29


Model slopes - represents the difference in predicted value of Y (plant height) for each one unit difference in bacteria amount if amount of sun is kept constant (i.e., as if plants were exposed to the same amount of sun light).

To do that, we use partial slopes – this is important because continuous predictors will rarely be orthogonal and as such we can't assign its effects to one or the other predictor.

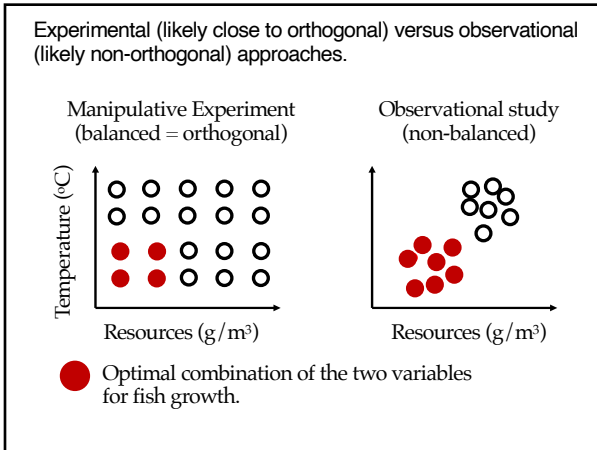
Total variation in plant height

=

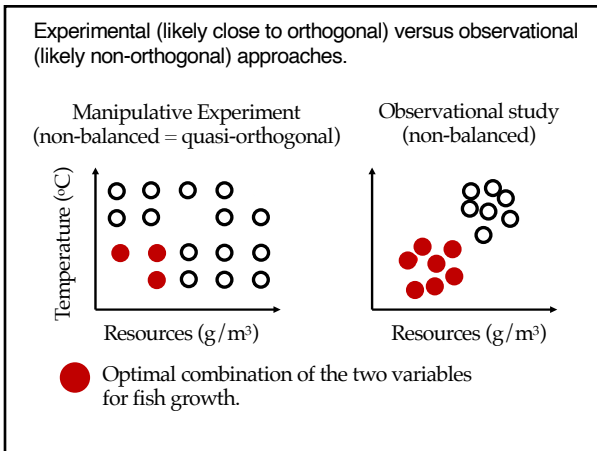


SOUNDS FAMILIAR 

30



31



32

**The properties of a regression model
(let's use a small simulation)**

Regression estimation (based on a sample) of the true population regression involves assumptions.

These assumptions are necessary so that the sample model is an unbiased estimate of the true population model; and that the tests involved have correct behaviour (e.g., Type I error rates = selected alpha).

A word on simulations *versus* math!

33

The properties of a regression model
(let's use a small simulation)

$$Y = 42cm + 2.3X_1 + 11X_2 + e$$

e residual error assumed to be $N(0, \sigma^2)$

Let's start with a really large sample size

```

4
5 n = 1000000
6 constant = 42
7 X1 = rnorm(n,1000,10)
8 X2 = rnorm(n,40,4)
9 error = rnorm(n,0,10)
10
11 Y = constant + 2.3*X1 + 11*X2 + error
12
    
```

34

The properties of a regression model
(let's use a small simulation)

$$Y = 42cm + 2.3X_1 + 11X_2 + e$$

e residual error assumed to be $N(0, \sigma^2)$

```

4
5 n = 1000000
6 constant = 42
7 X1 = rnorm(n,1000,10)
8 X2 = rnorm(n,40,4)
9 error = rnorm(n,0,10)
10
11 Y = constant + 2.3*X1 + 11*X2 + error
12
    
```

*Model results from simulated data
(large sample size, more accuracy)*

```

> lm(Y~X1+X2)
Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)          X1          X2
    42.687         2.299        10.998
    
```

35

The properties of a regression model
(let's use a small simulation)

Let's reduce sample size

36

The properties of a regression model
(let's use a small simulation)

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

e residual error are assumed to be $N(0, \sigma^2)$

```

19 n = 30
20 constant = 42
21 X1 = rnorm(n,1000,10)
22 X2 = rnorm(n,40,4)
23 error = rnorm(n,0,10)
24 Y = constant + 2.3*X1 + 11*X2 + error
25
26
27

```

Model results from simulated data
(smaller sample size, less accuracy;
compare with previous example)

```

> lm(Y~X1+X2)
Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)      X1         X2
  247.123      2.076     11.322

```

37

The properties of a regression model -

Predicted and residual variation

38

Understanding predicted values and residuals

$$Y = 247.12 + 2.08X_1 + 11.32X_2 + e$$

$$\hat{Y} = 247.12 + 2.08X_1 + 11.32X_2$$

$$e = Y - \hat{Y}$$

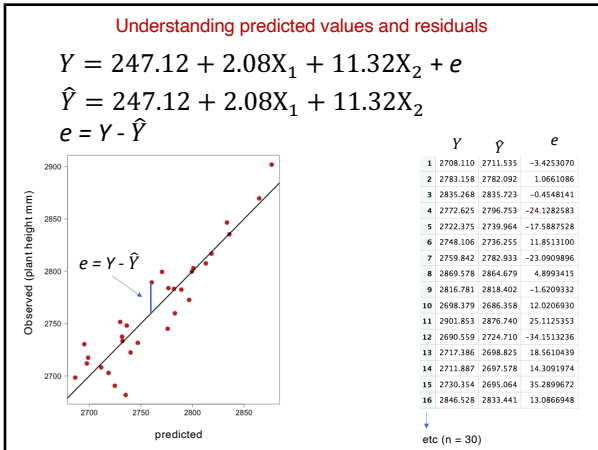
```

> lm(Y~X1+X2)
Call:
lm(formula = Y ~ X1 + X2)

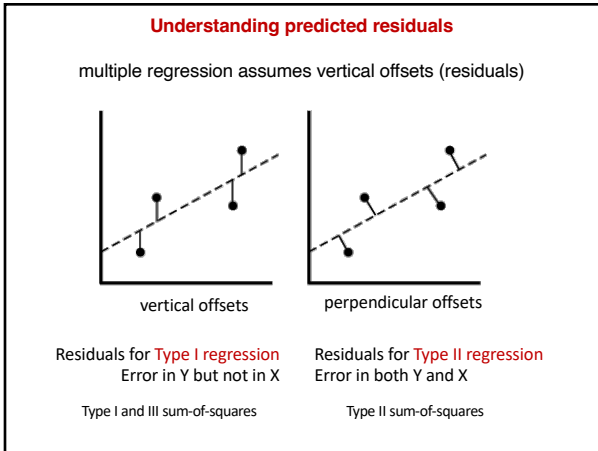
Coefficients:
(Intercept)      X1         X2
  247.123      2.076     11.322

```

39



40



41

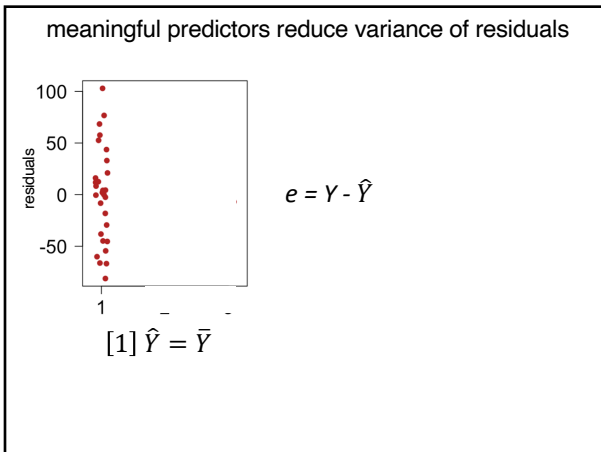
meaningful predictors reduce variance of residuals

A small fictional example to facilitate understanding of what regression coefficients mean!

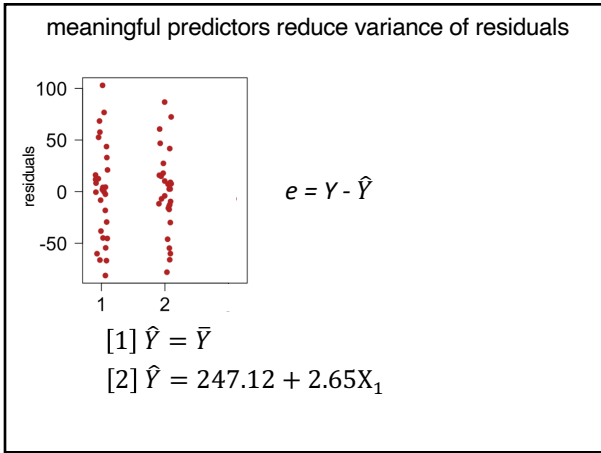
$$Y = 42cm + \beta_1X_1 + \beta_2X_2 + e$$

Y is plant height (cm)
 X₁ is amount of bacteria in the soil (count per ml)
 X₂ is amount of plant exposure to sun light (% exposure)

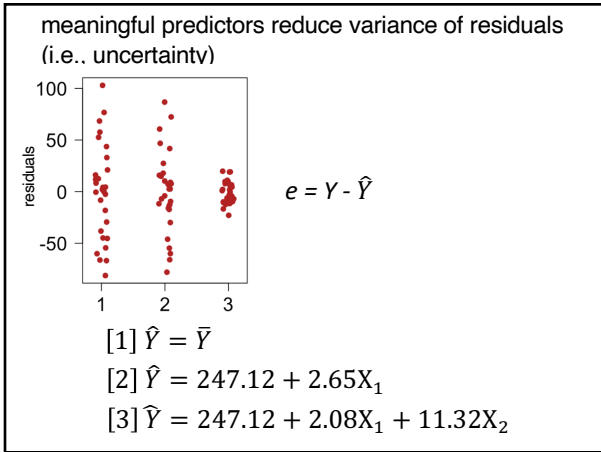
42



43



44



45



46

The properties/assumptions of a regression model

Linearity assumption
(big one)

47

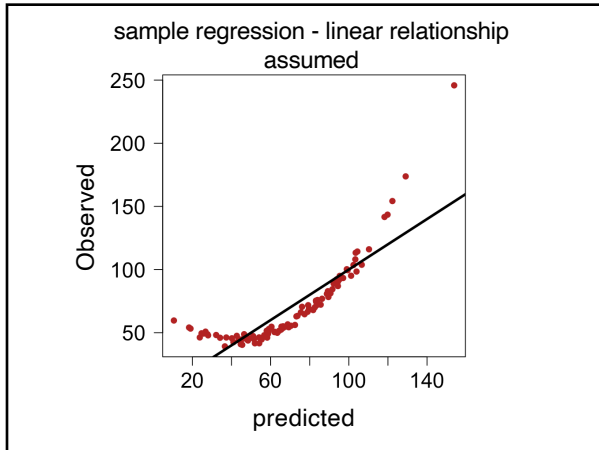
population regression

$$Y = 42 + 2.3X_1 + 11X_2^2 + e$$

```

260
261 n = 100
262 constant = 42
263 X1 = rnorm(n,1,1)
264 X2 = rnorm(n,1,1)
265 error = rnorm(n,0,1)
266 Y = constant + 2.3*X1 + 11*X2^2 + error
267
    
```

48



49

population regression

$$Y = 42 + 2.3X_1 + 11X_2^2 + e$$

sample regression

Call:
`lm(formula = Y ~ X1 + X2)`

Coefficients:

	X1	X2
(Intercept)	-0.7586	25.6188

treated as linear

$$Y = 42 - 0.76X_1 + 25.62X_2$$

50

population regression

$$Y = 42 + 2.3X_1 + 11X_2^2 + e$$

sample regression (non-linear regression)

`> lm(Y~X1+I(X2^2))`

Call:
`lm(formula = Y ~ X1 + I(X2^2))`

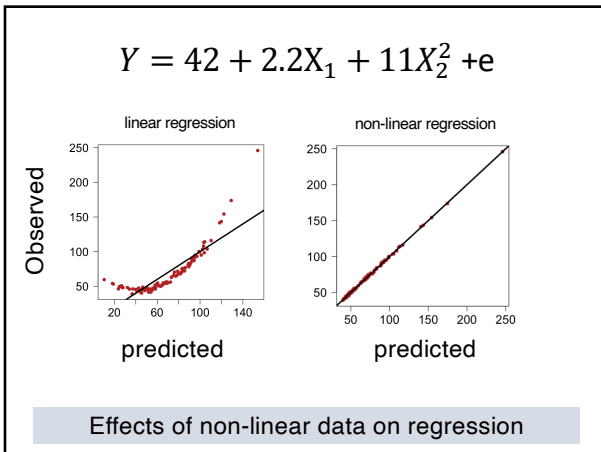
Coefficients:

	X1	I(X2^2)
(Intercept)	2.20	10.98

treated as non-linear

$$Y = 42 + 2.2X_1 + 11X_2^2$$

51



52

More on multiple regressions and assumptions - Lecture 12

53
