LECTURE 8: CORRELATIONS, LINEAR REGRESSION

AP STATS

Contents

As a quick reminder, recall some of the notation:

Mean (average) of X $$\mu_X = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Standard Deviation of X $$\sigma_X = \sqrt{\frac{(x_1 - \mu_X)^2 + (x_2 - \mu_X)^2 + \cdots + (x_n - \mu_X)^2}{n}}$$

Variance of X $$\mathrm{Var}(X) = \sigma_X^2 = \frac{(x_1 - \mu_X)^2 + (x_2 - \mu_X)^2 + \cdots + (x_n - \mu_X)^2}{n}$$

Covariance of X and Y $$\mathrm{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

Note that $\mathrm{cov}(X, X) = \sigma_X^2$, the variance of X.

1. Describing Correlations

Recall that the **correlation** r is given by the formula

$$r(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

for $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$. This correlation r is the standardized score of the covariance of X and Y

When you are asked to describe a scatterplot (or any relationship between two variables), you need to talk about

(1) If r is positive, then the data or negative
(2) The strength of the correlation:

$r-$ value	$0 \leq \lvert r\rvert \leq 0.25$	$0.25 \leq \lvert r\rvert \leq 0.5$	$0.5 \leq \lvert r\rvert \leq 0.8$	$0.8 \leq \lvert r\rvert \leq 1$
correlation strength	very weak	weak	moderate	strong

Use $DUFSC$ whenever you are asked to describe any scatterplots:

D*irection*	positive or negative direction
U*nusual features*	outliers or clusters
F*orm*	linear or nonlinear
S*trength*	weak/moderate/strong
C*ontext*	write answer in a sentence

A few things about r:

- it's a number without units
- $r(X, Y) = r(Y, X)$
- $-1 \leq r \leq 1$

Example 1. "The scatterplot of (units of Y) verses the (units of X) for (the problem setting) shows a moderately strong negative linear association."

2. Linear Regressions

Broadly speaking, **regression** is a method for studying relationships between two quantitative variables.

(1) X-the set which we call the "explanatory variable(s)"[1]
(2) Y- the set which we call the "response variable"

We want to use a *regression function*, whose technical definition is

$$r(x) = \int y\ f\left(y|x\right) \mathrm{d}y$$

We're going to study the simplest case, where the regression is a line.

Recall that we have lines as

$$y = mx + b$$

for some fixed constants m and b. Once you know those constants, you know everything about the line.

[1] also called covariate

Similarly, a **linear regression model** is a line given by the equation

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

for some fixed constants α, β.

- the α is playing the role of y-intercept
- the β is playing the role of slope
- the ε_i are called **residuals**
- You compute the residuals by

$$\varepsilon_i = \textit{actual} - \textit{predicted} = y_i - \widehat{y_i}$$

- We measure the accuracy of our linear regression model by using an **SSR** (sum of squares of residuals)

$$SSR = \sum_{i=1}^{n} \varepsilon_i^2$$

Example 1. Given the data set $\{(2, 11), (3, 17), (4, 29)\}$, suppose we take a linear regression model $y = -8 + 9x$.

Let's compute the residuals for each point. You do this by

(1) plugging in the x value into the model function...this gives $\widehat{y}$
(2) subtract result from corresponding y-coordinate from data set.

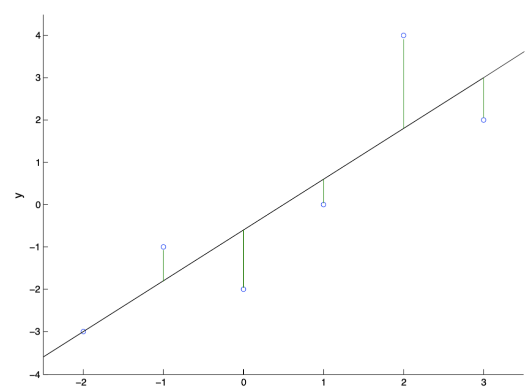
Since

$$\varepsilon_i = y_i - \widehat{y_i}, \quad \widehat{y_i} = -8 + 9x_i$$

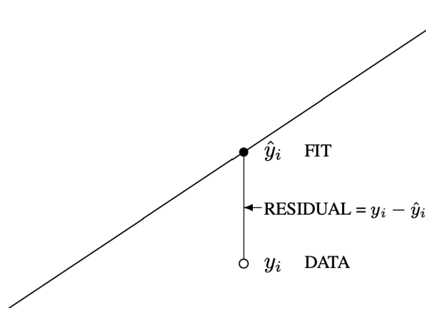
You should get:

$$\begin{aligned}
\varepsilon_{x=2}: &\quad \widehat{y}(x=2) = -8 + 9(2) = 10 \Longrightarrow \varepsilon_{x=2} = 11 - 10 = 1 \\
\varepsilon_{x=3}: &\quad \widehat{y}(x=3) = -8 + 9(3) = 19 \Longrightarrow \varepsilon_{x=3} = 17 - 19 = -2 \\
\varepsilon_{x=4}: &\quad \widehat{y}(x=4) = -8 + 9(4) = 28 \Longrightarrow \varepsilon_{x=4} = 29 - 28 = 1
\end{aligned}$$

Our SSR is $(1)^2 + (-2)^2 + (1)^2 = 6$.



In the picture, the vertical lines are the residuals, and if you zoom in, you see:



This is only *one* example of a linear regression model...there are infinitely many models we can create! However, not all models are going to be as helpful and there is a "best" regression line that will triump all the other lines in accuracy! This line is called the **least squares regression line** or simply ***the* regression line**.

The ***least squares regression line*** is characterized as *the line that minimizes the SSR.*

Theorem 1. *The least squares regression line for a scatterplot exists.*

Moreover, it is of the form

$$y = \alpha + \beta x$$

where $\beta = \dfrac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}$ *and* $\alpha = \mu_Y - \beta\mu_X$

Proof. Calculus magic! This will be an extra credit homework problem. □

Because of the characterizing property, we get the following results:

Proposition 2. Given a scatterplot and *the* regression line (the least squares regression line),

(a) the sum of the residuals is 0.
(b) we can rewrite β as $\beta = r\dfrac{\sigma_Y}{\sigma_X}$. So the slope of the regression line is $slope = r\dfrac{\sigma_Y}{\sigma_X}$
(c) Writing β_X as the slope for the regression of Y on X, and β_Y as the slope for the regression of X on Y, you can see that

$$\beta_X\beta_Y = r^2.$$

Proof. (a) HW (part of Extra credit)
(b) HW on Problem Set 5
(c) HW on Problem Set 5

□