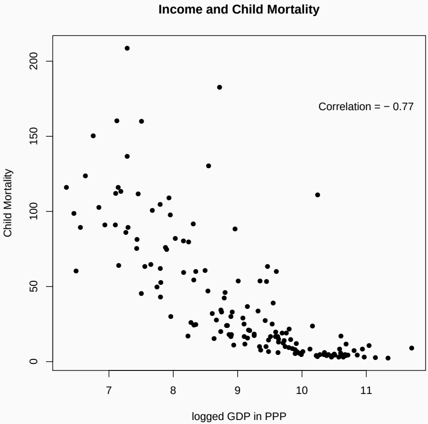
Political Science 209 - Fall 2018

Linear Regression

Florian Hollenbach

11th October 2018

Recall Correlation & Scatterplot



What is the correlation?

Recall the definition of correlation

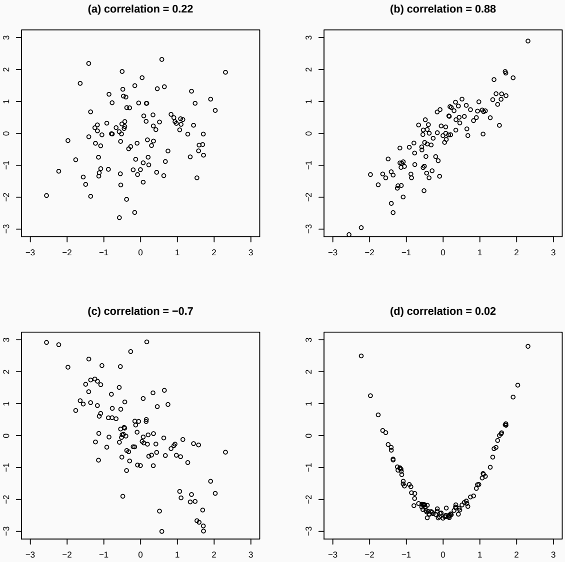
Correlation (x,y) $= \frac{1}{N}\sum_{i=1}^{N}$ z-score of $x_i \times$ z-score of y_i

Correlation (x,y) $= \frac{1}{N}\sum_{i=1}^{N} \frac{x_i - \bar{x}}{sd_x} \times \frac{y_i - \bar{y}}{sd_y}$

Correlations & Scatterplots/Data points

1. positive correlation $\rightsquigarrow$ upward slope
2. negative correlation $\rightsquigarrow$ downward slope
3. high correlation $\rightsquigarrow$ tighter, close to a line
4. correlation cannot capture nonlinear relationship

Correlations & Scatterplots/Data points



Moving from Correlation to Linear Regression

Preview:

- linear regression allows us to create predictions
- linear regression specifies direction of relationship
- linear regression allows us to examine more than two variables at the same time (*statistical control*)

Linear Regression

- regression has one dependent (y) and *for now* one independent (x) variable
- regression is a statistical method to estimate the linear relationship between variables

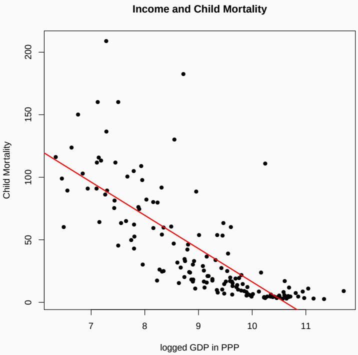
- goal of regression is to approximate the (linear) relationship between X and Y as best as possible

Linear Regression

- goal of regression is to approximate the (linear) relationship between X and Y as best as possible

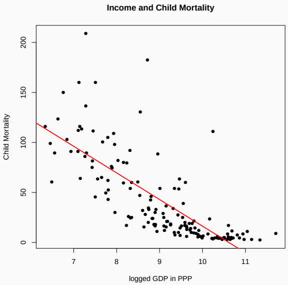
- regression is the mathematical model to draw best fitting line through cloud of points

Linear Regression



Linear regression is the mathematical model to draw best fitting line through cloud of points

Linear Regression



- regression line is an estimate of the (for now bivariate) relationship between x and y
- for each x we have a prediction of y: what would we expect y to be given the value of x?

What is the equation of a line?

Equation of a line?

What is the equation of a line?

Equation of a line? $y = mx + b$

$\rightarrow$ b? m?

What is the equation of a line?

Equation of a line?

$y = mx + b$

b $\rightarrow$ y-intercept

m $\rightarrow$ slope

What is the equation of a line?

Equation of a line?

$y = mx + b$

b $\rightarrow$ y-intercept

m $\rightarrow$ slope

regression equation:

$Y = \alpha + \beta X + \epsilon$

$\rightarrow \alpha$? β? ϵ?

What is the equation of a line?

Equation of a line?

$y = mx + b$

b $\rightarrow$ y-intercept

m $\rightarrow$ slope

regression equation:

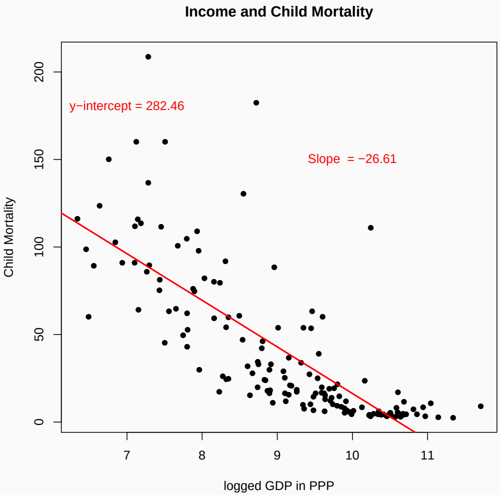
$Y = alpha + \beta X + \epsilon$

$\alpha \rightarrow$ y-intercept

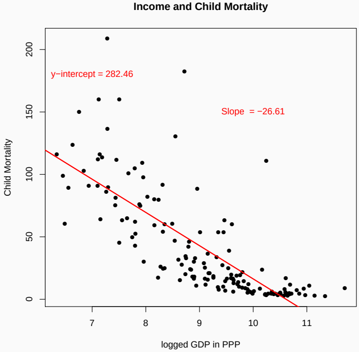
$\beta \rightarrow$ slope

$\epsilon \rightarrow$ error

Regression equation



Regression equation



$$Y = 282.46 + -26.61X + \epsilon$$

Regression equation

Model:

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

- Y: dependent/outcome/response variable
- X: independent/explanatory variable, predictor
- (α, β): coefficients (parameters of the model)
- ϵ: unobserved error/disturbance term (mean zero)

Regression: Interpretation of the Parameters:

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

- $\alpha + \beta X$: average of Y at the given the value of X
- α: the value of Y when X is zero
- β: increase in Y associated with one unit increase in X

Regression equation

- but, we don't know the equation that generates the data
- our regression line is an estimate, based on the collected data

Regression equation

- but, we don't know the equation that generates the data
- our regression line is an estimate, based on the collected data

- estimates are denoted with little hats: $\hat{\beta}$, $\hat{\alpha}$
- $(\hat{\alpha}, \hat{\beta})$: estimated coefficients

Regression equation

- but, we don't know the equation that generates the data
- our regression line is an estimate, based on the collected data

- estimates are denoted with little hats: $\hat{\beta}$, $\hat{\alpha}$
- $(\hat{\alpha}, \hat{\beta})$: estimated coefficients

- we can use $(\hat{\alpha}, \hat{\beta}, X)$ to create *predicted values* of y
- $\widehat{Y} = \hat{\alpha} + \hat{\beta}x$: predicted/fitted value

Regression equation

How far off is our line? How do we know?

Regression equation

How far off is our line? How do we know?

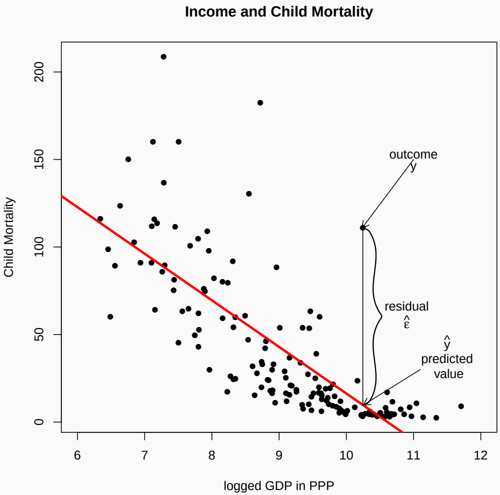
Regression equation

How far off is our line? How do we know?

$\hat{\epsilon} =$ true $Y - \widehat{Y}$: residuals/error

$\hat{\epsilon}$'s are an estimate of how good/bad our line approximates the relationship

Regression



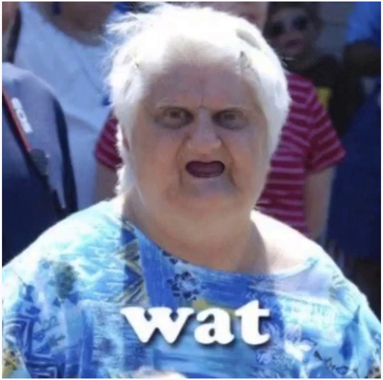
Income and Child Mortality
Child Mortality
logged GDP in PPP
200
150
100
50
0
6
7
8
9
10
11
12
outcome
y
residual
ε̂
ŷ
predicted
value

Regression

- (α, β) are estimated from the data
- How do we find α, β?

Regression: How do we find α, β?

We minimize the sum of the squared residuals



Regression: How do we find α, β?

We minimize the *sum of the squared residuals (SSR)*

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

Regression: How do we find α, β?

We minimize the *sum of the squared residuals (SSR)*

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

This also minimizes the root mean squared error: $\text{RMSE} = \sqrt{\frac{1}{n}\text{SSR}}$

$$
\begin{aligned}
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\
\hat{\beta} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}
\end{aligned}
$$

OR:

Regression by Hand

$$\begin{aligned}\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\end{aligned}$$

OR:

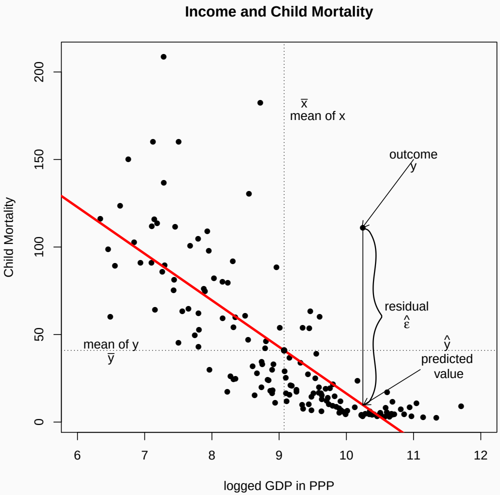
$$\hat{\beta} = \text{correlation of } X \text{ and } Y \times \frac{\text{standard deviation of } Y}{\text{standard deviation of } X}$$

Regression by Hand

Regression line always goes through the point of averages $(\hat{X}, \hat{Y})$

$$\widehat{Y} \;=\; (\overline{Y} - \hat{\beta}\overline{X}) + \hat{\beta}\overline{X} \;=\; \overline{Y}$$

Regression always goes through point of averages



Regression NOT by Hand

Enough math!

Fitting/estimating a regression in *R*:

```
lm(dependent ~ independent, data = data_object)
```

Regression NOT by Hand

Fitting/estimating a regression in *R*:

```
data <- read.csv("bivariate_data.csv")
data <- subset(data, Year ==2010)
result <- lm(Child.Mortality ~ log(GDP) , data = data)
summary(result)
```

Regression NOT by Hand

```
result <- lm(Child.Mortality ~ log(GDP) , data = data)
coef(result) ### coefficients
```

```
(Intercept)    log(GDP)
  282.45870   -26.61347
```

R-output:

(Intercept): α

log(GDP): β

Model Fit

How well does our regression line fit the data?

How well does the model predict the outcome?

Model Fit

How well does our regression line fit the data?

How well does the model predict the outcome?

R^2 or *coefficient of determination*:

$$R^2 \;=\; 1 - \frac{\text{SSR}}{\text{Total sum of squares (TSS)}} \;=\; 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Model Fit

$$R^2 \;=\; 1 - \frac{\text{SSR}}{\text{Total sum of squares (TSS)}} \;=\; 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

R^2 is also defined as the *explained variance* in Y

How much of the deviation of Y from the average is explained by X?

Model Fit

```
result <- lm(Child.Mortality ~ log(GDP) , data = data)
summary(result)
```

```
Call:
lm(formula = Child.Mortality ~ log(GDP), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.455 -15.418  -4.161  10.847 132.136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  282.459     16.569   17.05   <2e-16 ***
log(GDP)     -26.613      1.809  -14.71   <2e-16 ***
---
codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.57 on 150 degrees of freedom
Multiple R-squared:  0.5906,Adjusted R-squared:  0.5878
F-statistic: 216.4 on 1 and 150 DF,  p-value: < 2.2e-16
```