Week 8: Regression

Wenhao Jiang

Department of Sociology
New York University

October 28, 2022

Regression

Regression Basics

- The main aim of regression is to find the correlation (or relationship, or association) between two variables
  - We are interested in the relationship between vaccinations per person and deaths per $100k$ across US states.
  - We are interested in the relationship between neighborhood racial segregation and the chance of children's upward mobility

Regression Basics

- The main aim of regression is to find the correlation (or relationship, or association) between two variables
  - We are interested in the relationship between vaccinations per person and deaths per $100k$ across US states.
  - We are interested in the relationship between neighborhood racial segregation and the chance of children's upward mobility
- We call the variable we want to use to explain the other the **explanatory**, or the **independent** variable
- We call the variable being explained the **outcome**, or the **dependent** variable

Regression Basics

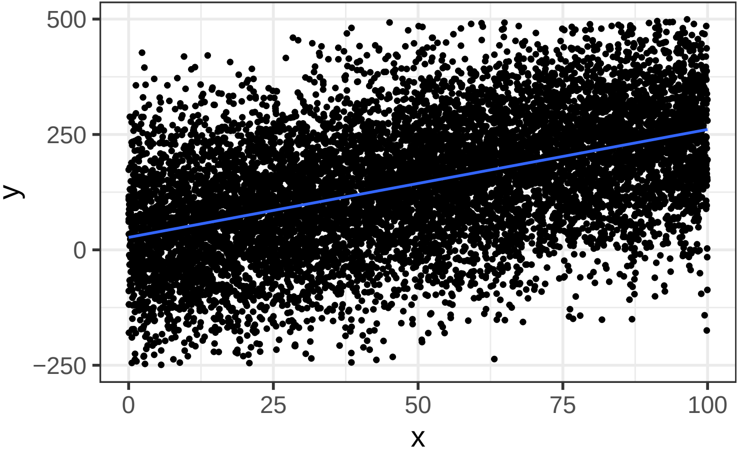
- The main aim of regression is to find the correlation (or relationship, or association) between two variables
- We call the variable we want to use to explain the other the **explanatory**, or the **independent** variable
- We call the variable being explained the **outcome**, or the **dependent** variable
- We use x_i to denote the **explanatory** variable, and y_i to denote the **outcome** variable, with i representing individual observations (i.e., there can be many values for the explanatory and the outcome variable, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$)

Regression Basics

- The main aim of regression is to find the correlation (or relationship, or association) between two variables
- We call the variable we want to use to explain the other the **explanatory**, or the **independent** variable
- We call the variable being explained the **outcome**, or the **dependent** variable
- We use x_i to denote the **explanatory** variable, and y_i to denote the **outcome** variable, with i representing individual observations (i.e., there can be many values for the explanatory and the outcome variable, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$)
- Therefore, y_i is always at the left side of the equation to be "predicted" or "fitted" by x_i, which is always at the right side of the equation.

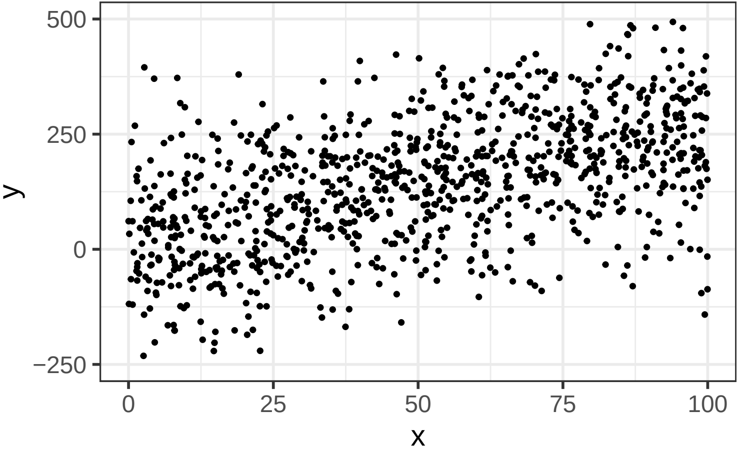
Regression Basics

- At the **population** level, there is a true **data-generating process** subject to
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$



Regression Basics

- We never observe the population-level process. We can only estimate the true **data-generating process** by the **single** sample ($n = 1000$) we draw



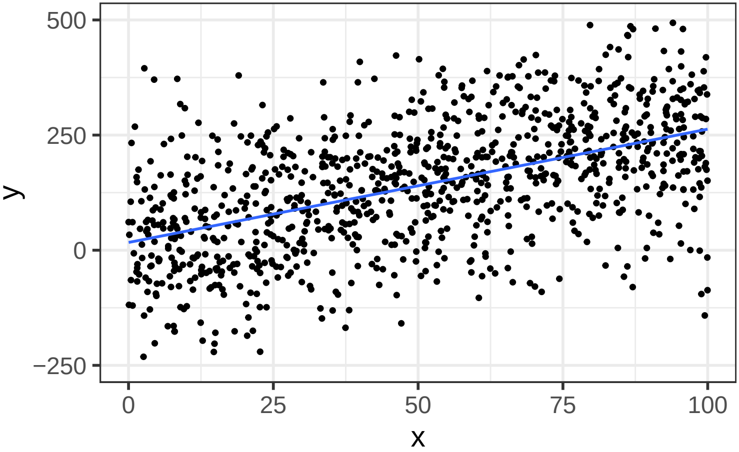
Regression Basics

- We never observe the population-level process. We can only estimate the true **data-generating process** by the **single** sample ($n = 1000$) we draw
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

Regression Basics

- We never observe the population-level process. We can only estimate the true **data-generating process** by the **single** sample ($n = 1000$) we draw
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
- We derive $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the total sum of e_i^2, $S = \sum_{i=1}^{n} e_i^2$

Regression Basics



Regression Basics

- We never observe the population-level process. We can only estimate the true **data-generating process** by the **single** sample ($n = 1000$) we draw
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
- $\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(x_i)}$
- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

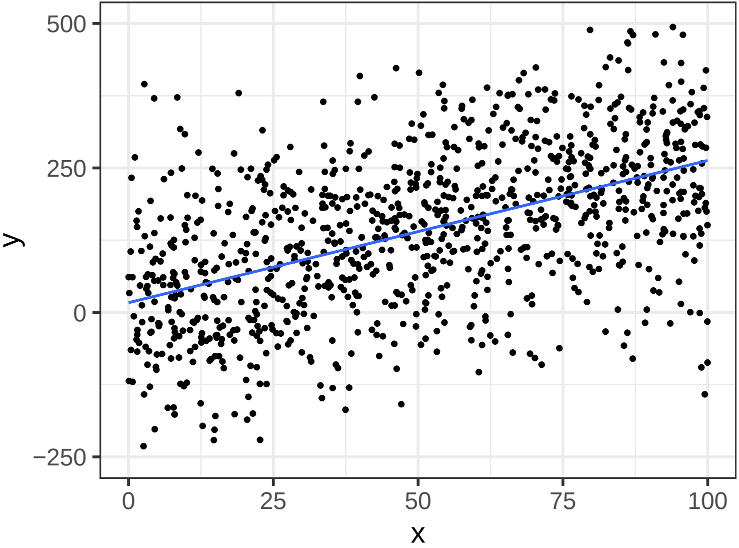
Regression Basics

- Just as other statistics e.g. sample mean, $\hat{\beta}_1$ and $\hat{\beta}_0$ involves uncertainty

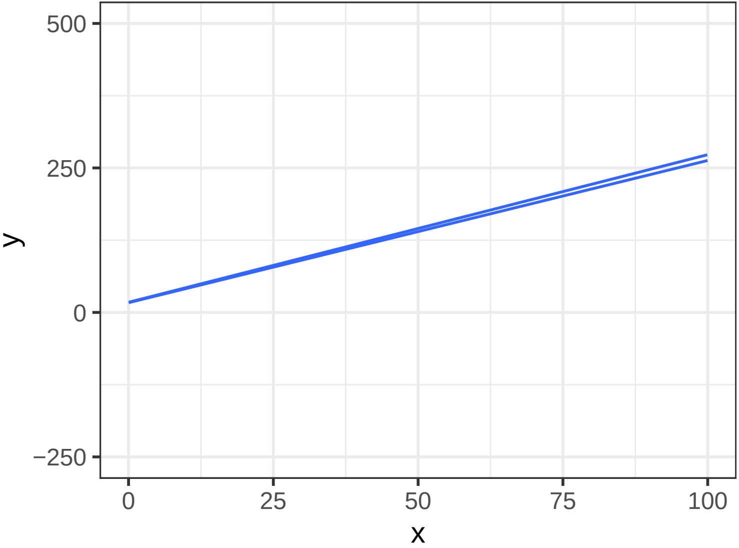
Regression Basics

- Just as other statistics e.g. sample mean, $\hat{\beta}_1$ and $\hat{\beta}_0$ involves uncertainty
- The uncertainty comes from the sampling, i.e., for different samples, we will get a different $\hat{\beta}_1$ and $\hat{\beta}_0$

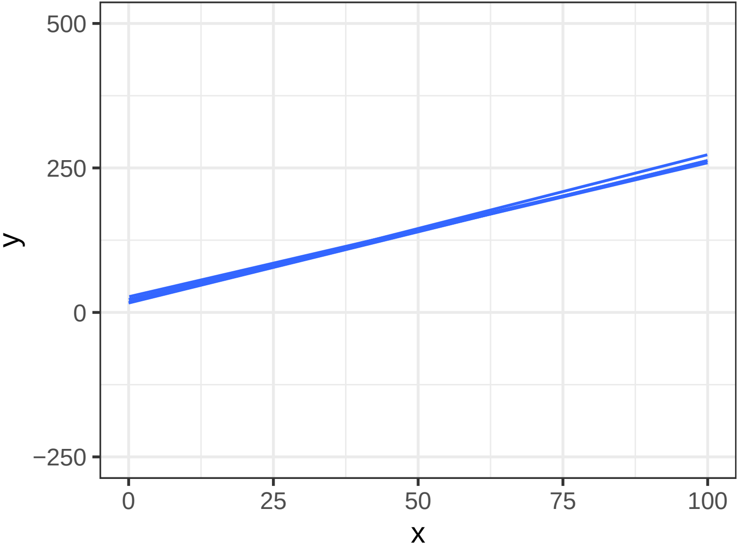
Regression Basics



Regression Basics

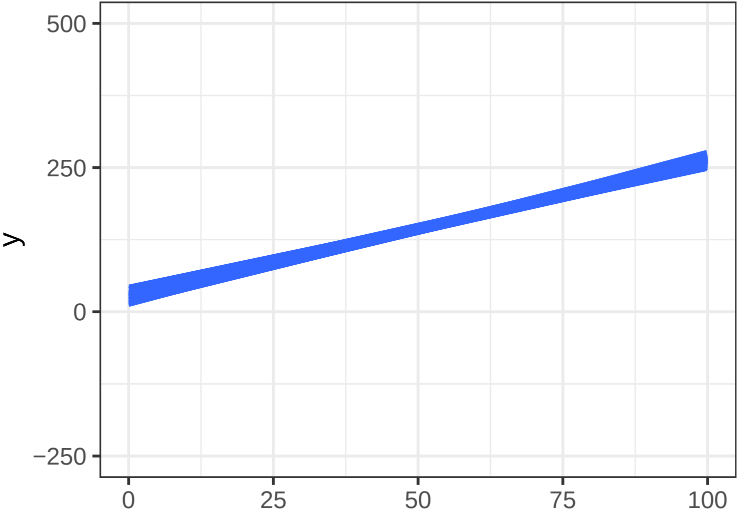


Regression Basics



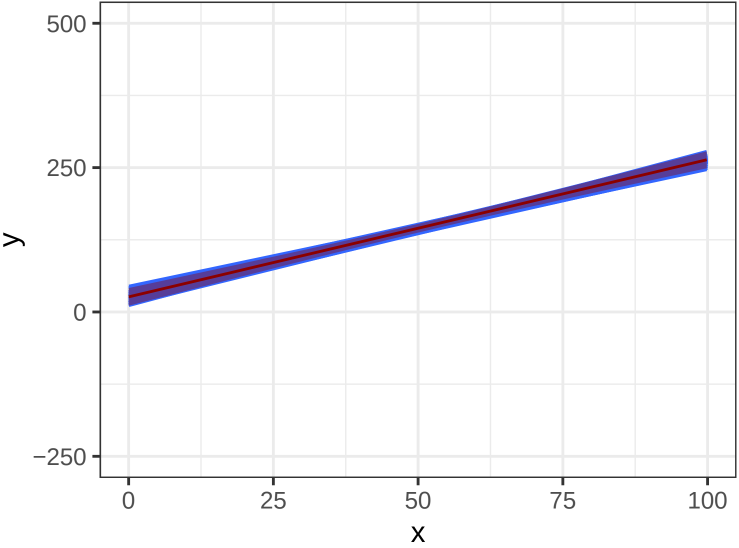
Regression Basics

- We sample from the population for $K = 2000$ times and calculate $\hat{\beta}_1$ and $\hat{\beta}_0$ for each sample.



Regression Basics

- In around 95% of the cases, $\hat{\beta}_1$ will fall into:



Regression Basics

- The standard deviation of these K $\hat{\beta}_1$s, i.e., the standard error of $\hat{\beta}_1$, is
- $\frac{\sigma_{\epsilon_i}/s_{x_i}}{\sqrt{n}}$

Regression Basics

- The standard deviation of these K $\hat{\beta}_1$s, i.e., the standard error of $\hat{\beta}_1$, is
- $\frac{\sigma_{\epsilon_i}/s_{x_i}}{\sqrt{n}}$
- We never observe σ_{ϵ_i}, instead, we use the sample error standard deviation s_{e_i} to estimate σ_{ϵ_i}
- This leads to $\frac{s_{e_i}/s_{x_i}}{\sqrt{n-2}}$ as an unbiased estimate (no need to memorize any of these)

Regression Basics

- The standard deviation of these K $\hat{\beta}_1$s, i.e., the standard error of $\hat{\beta}_1$, is
- $\frac{\sigma_{\epsilon_i}/s_{x_i}}{\sqrt{n}}$
- We never observe σ_{ϵ_i}, instead, we use the sample error standard deviation s_{e_i} to estimate σ_{ϵ_i}
- This leads to $\frac{s_{e_i}/s_{x_i}}{\sqrt{n-2}}$ as an unbiased estimate (no need to memorize any of these)
- This is also why in a regression table, only $t-$score rather than $z-$score is reported. We almost always underestimate σ_{ϵ_i} by using s_{e_i}

Exercise

- To estimate the Data-Generating Process in the population: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, a sample of $n = 1000$ is drawn. A linear regression line is fitted to the observed data: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, where e_i describes the random error that cannot be explained by the linear regression line. How are $\hat{\beta}_0$ and $\hat{\beta}_1$ determined?
  - A. By crossing the mean point $(\bar{x}_i, \bar{y}_i)$
  - B. By minimizing $\sum_i^n e_i$
  - C. By minimizing $\sum_i^n |e_i|$
  - D. By maximizing $\sum_i^n e_i^2$
  - E. None of the above

Exercise

- To minimize $\sum_i^n e_i^2$, which of the following(s) have to be satisfied?
  - A. $\hat{\beta}_1 = \frac{Cov(x_i, y_i)}{Var(y_i)}$
  - B. The linear regression line must cross the mean point $(\bar{x}_i, \bar{y}_i)$
  - C. The sum of the error term $\sum_{i=1}^n e_i = 0$
  - D. The sum of the absolute values of the error term $\sum_{i=1}^n |e_i|$ must be minimized

$$
\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right] \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= n\bar{y} - n\hat{\beta}_0 + n\hat{\beta}_1 \bar{x} \\
&= n(\bar{y} - \hat{\beta}_0 + \hat{\beta}_1 \bar{x}) \\
&= 0
\end{aligned}
$$

Intepretation of Regression

- Suppose we are interested in the returns to education in the United States. We sampled from the population $n = 1000$ individuals, surveyed their years of education (x_i) and annual income at the age of 30 (y_i). We use a linear regression model to fit the data and find that $y_i = 5000 + 4000x_i + e_i$
- How do interpret the main result?

Intepretation of Regression

- Suppose we are interested in the returns to education in the United States. We sampled from the population $n = 1000$ individuals, surveyed their years of education (x_i) and annual income at the age of 30 (y_i). We use a linear regression model to fit the data and find that $y_i = 5000 + 4000x_i + e_i$
- How do interpret the main result?
- On average, one additional year of education is associated with 4000 more annual income at the age of 30.

Intepretation of Regression

- Suppose we are interested in the returns to education in the United States. We sampled from the population $n = 1000$ individuals, surveyed their years of education (x_i) and annual income at the age of 30 (y_i). We use a linear regression model to fit the data and find that $y_i = 5000 + 4000x_i + e_i$
- How do interpret the main result?
- On average, one additional year of education is associated with 4000 more annual income at the age of 30.
- Do we consider this association as a causal relationship? That is, if a person who had never been to college after finishing high school hypothetically completed 4-year college education, would he/she have earned $16K$ more annually?

Intepretation of Regression

- Do we consider this association as a causal relationship?
- Very likely no. Other factors like family conditions may also affect the probability of gaining additional year of education and, at the same time, the earning potential
- This is one critical reason why we want to have multivariate regression
- We will review multivariate regression in more detail next week

Regression in R

Read Data

```r
## set your working directory - you should set your own unique one!
setwd("~/Dropbox/Teaching/SOCUA-302/Week 8")

## read csv data - this is 2021 GSS data
gss <- read.csv("GSS_SOCUA_W8.csv")
```

Bivariate Regression

- R reports the estimates of intercept, slope (coefficient), their standard errors, their corresponding $t-$value, and $p-$values

```
## specify regression model
model <- lm(rincome ~ educ, gss)

## check the results
summary(model)
```

```
##
## Call:
## lm(formula = rincome ~ educ, data = gss)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20376  -2324   5035   5318   7439
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17560.98    1001.81  17.529   <2e-16 ***
## educ          141.41      65.22   2.168   0.0302 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8713 on 2512 degrees of freedom
##   (1518 observations deleted due to missingness)
## Multiple R-squared:  0.001868,   Adjusted R-squared:  0.00147
## F-statistic: 4.701 on 1 and 2512 DF,  p-value: 0.03025
```