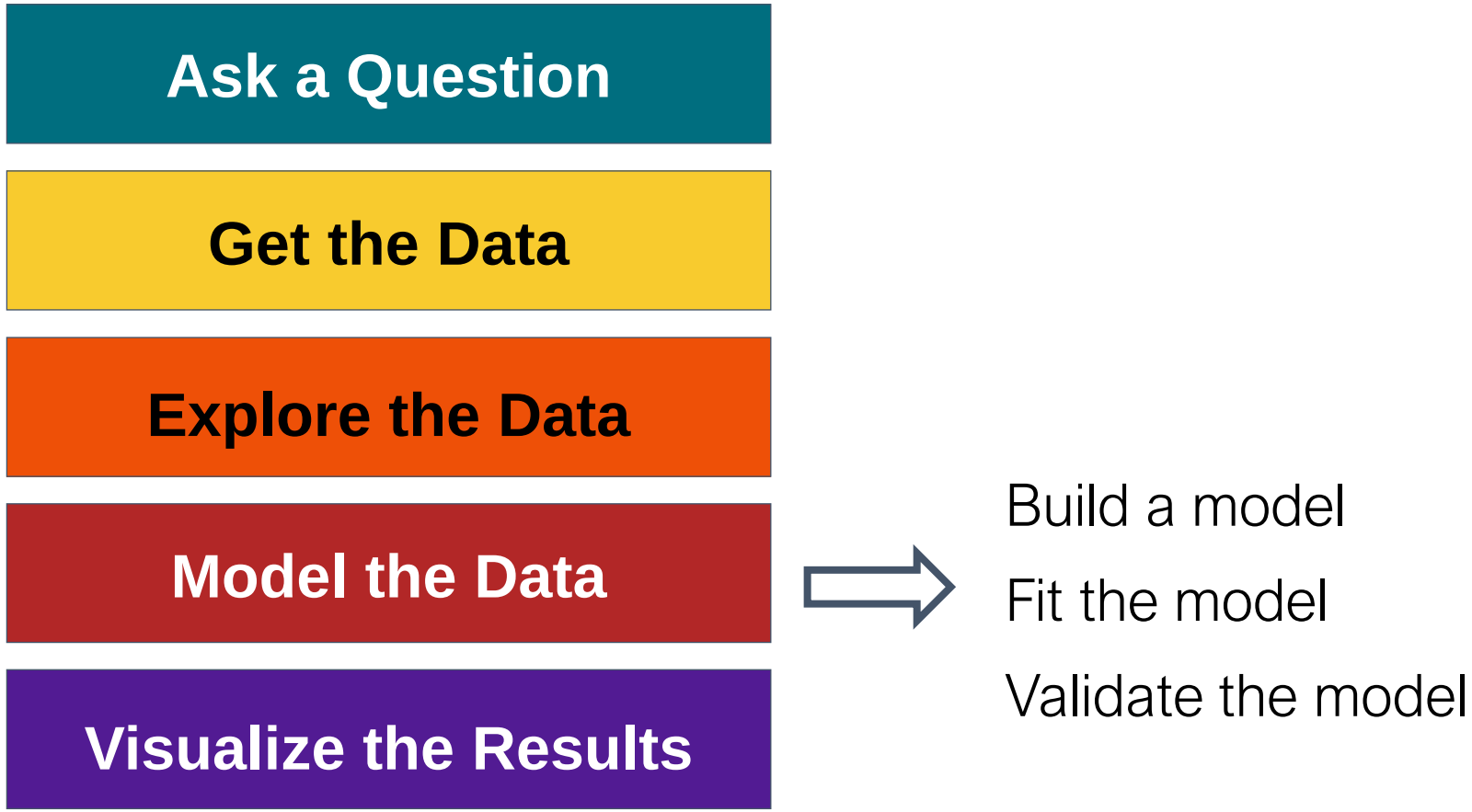
Linear Regression

Esra Suel

CASA0006: Data Science for Spatial Systems

Slides from Ollie Ballinger

The Data Science Process



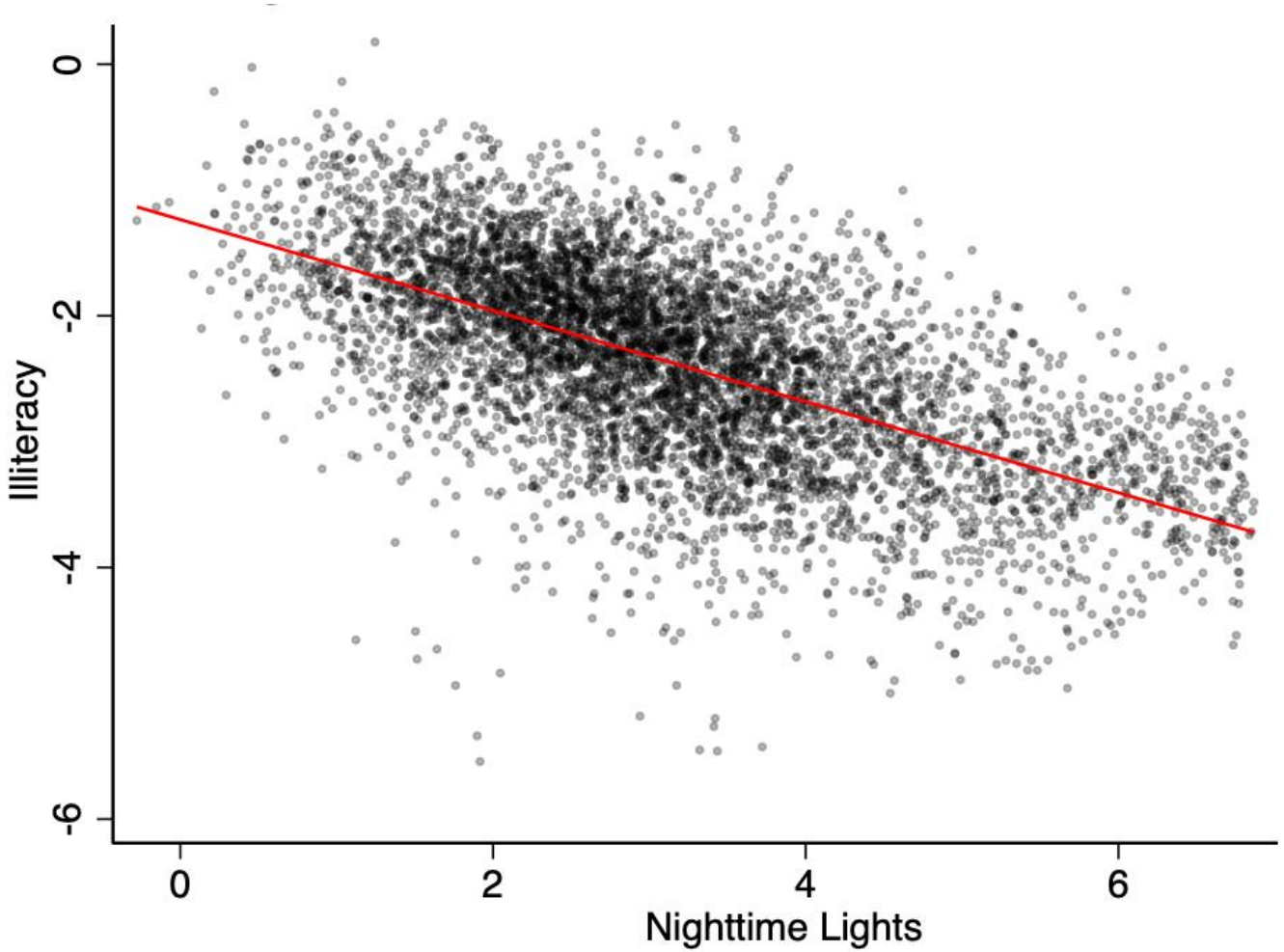
Outline

1. Visualization
2. Ordinary Least Squares (OLS)
    1. Fitting a Model
    2. Estimating Coefficients
    3. Standard Errors
    4. Confidence Interval on Predicted Values
3. Assumptions

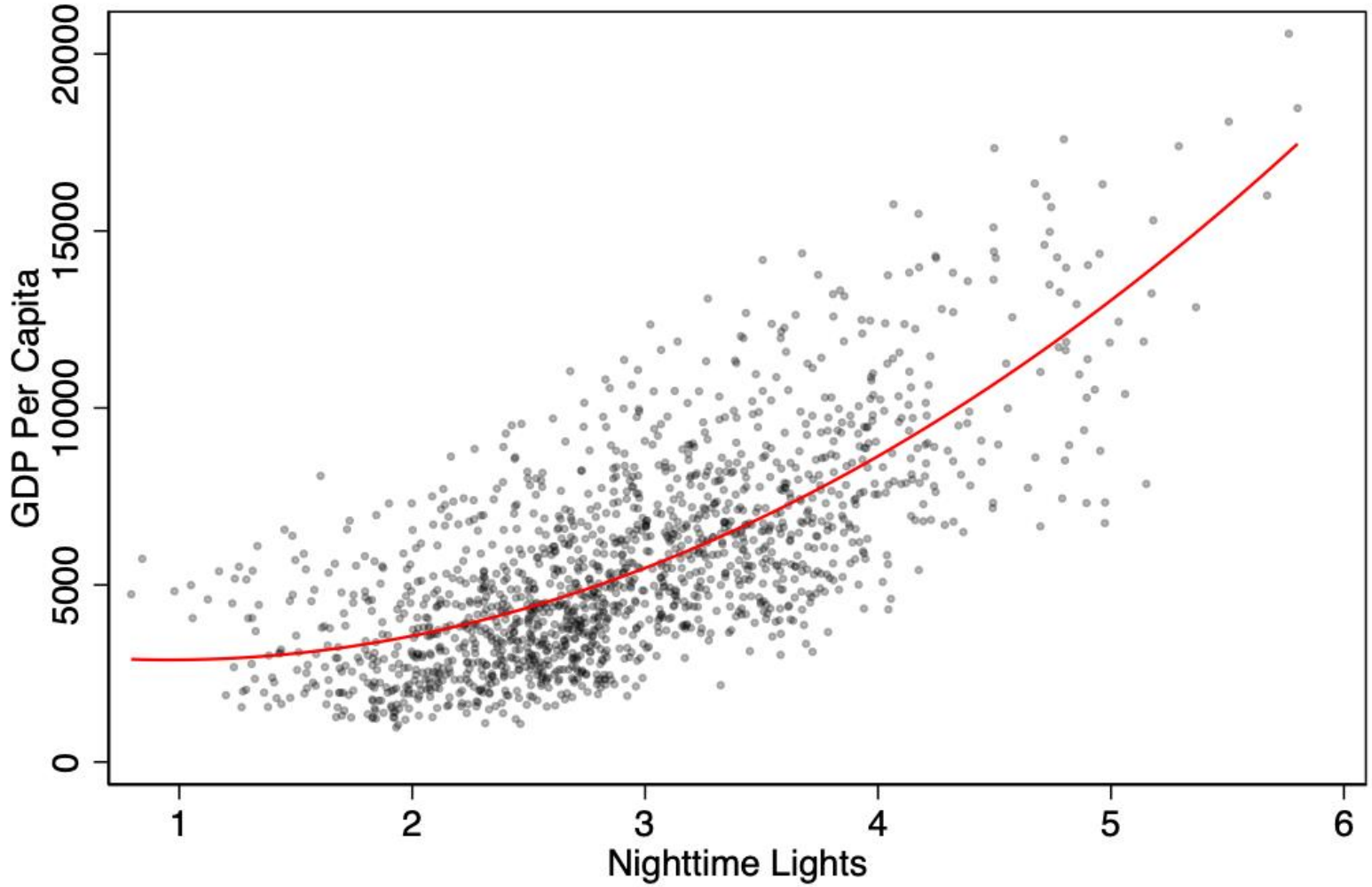
1. Visualisation

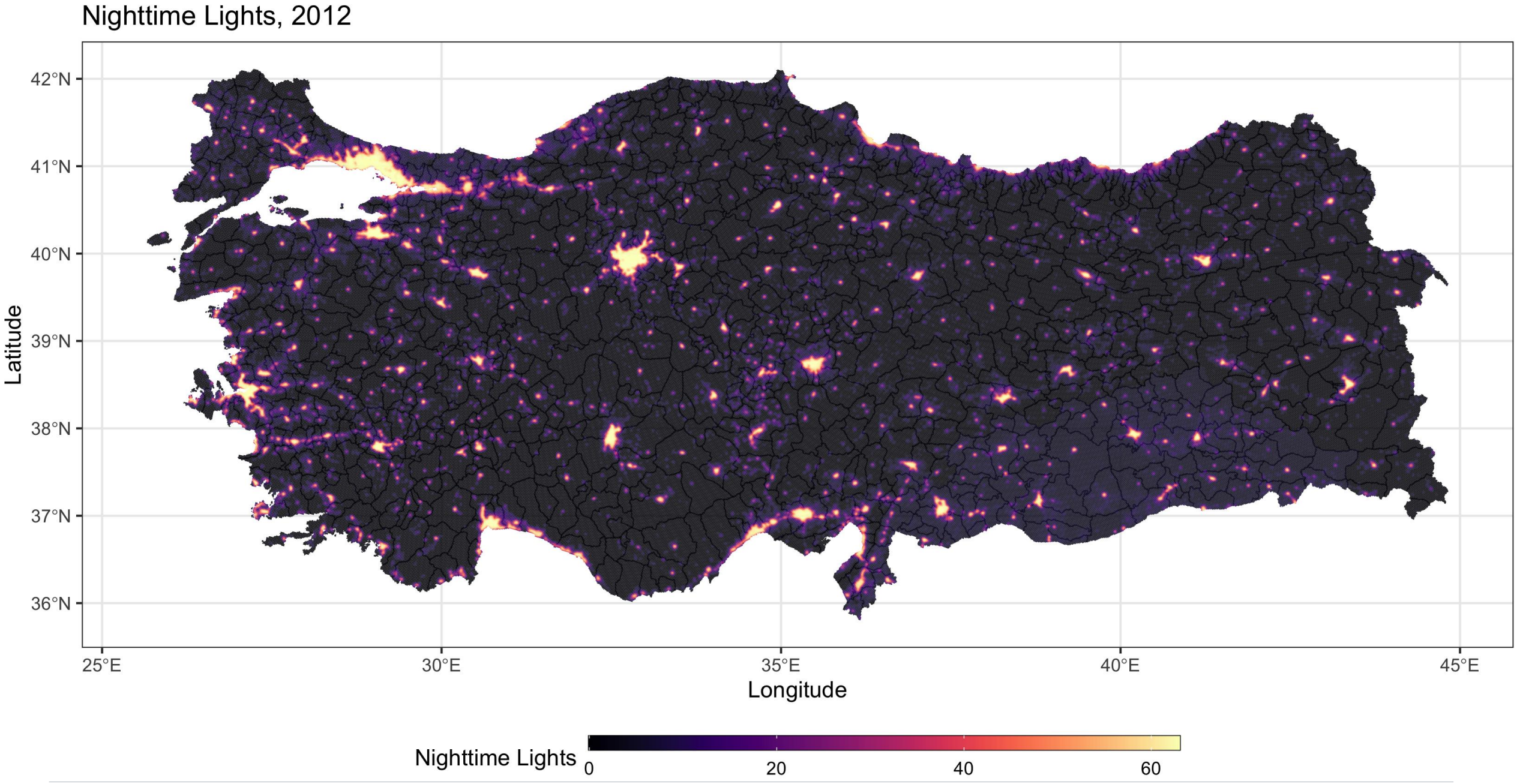

EROTIK SHOP
34 OJ 9639

Nighttime Lights and Literacy



Nighttime Lights and GDP Per Capita




Nighttime Lights, 2012
42°N
41°N
40°N
39°N
38°N
37°N
36°N
Latitude
25°E
30°E
35°E
40°E
45°E
Longitude
Nighttime Lights
0
20
40
60

Two continuous variables

What if we want to measure the relationship between two continuous variables, like income and years of schooling?

2. Ordinary Least Squares

The Regression Equation

Here's the equation for a regular line: $y = b + mx$

Here's the equation for a *regression line:*

$$y = \beta_o + \beta_1 x + \varepsilon$$

y: Dependent Variable

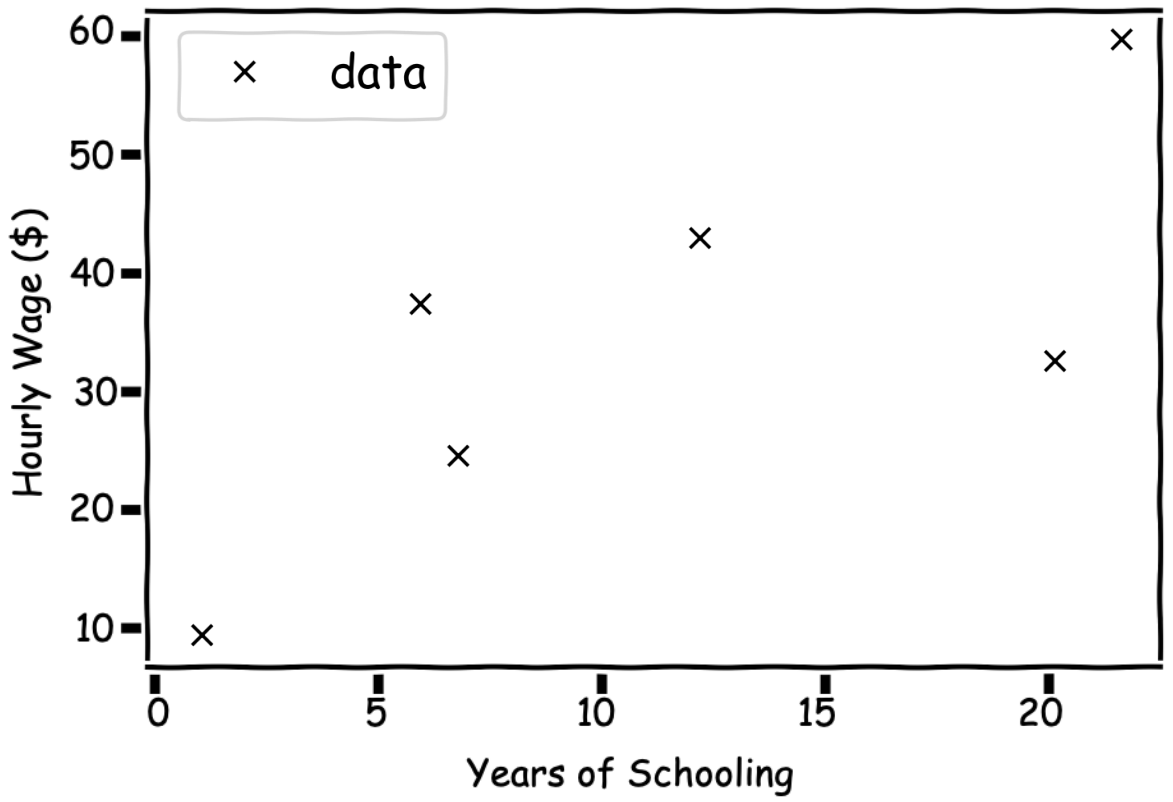
x: Independent Variable

β_o: Intercept

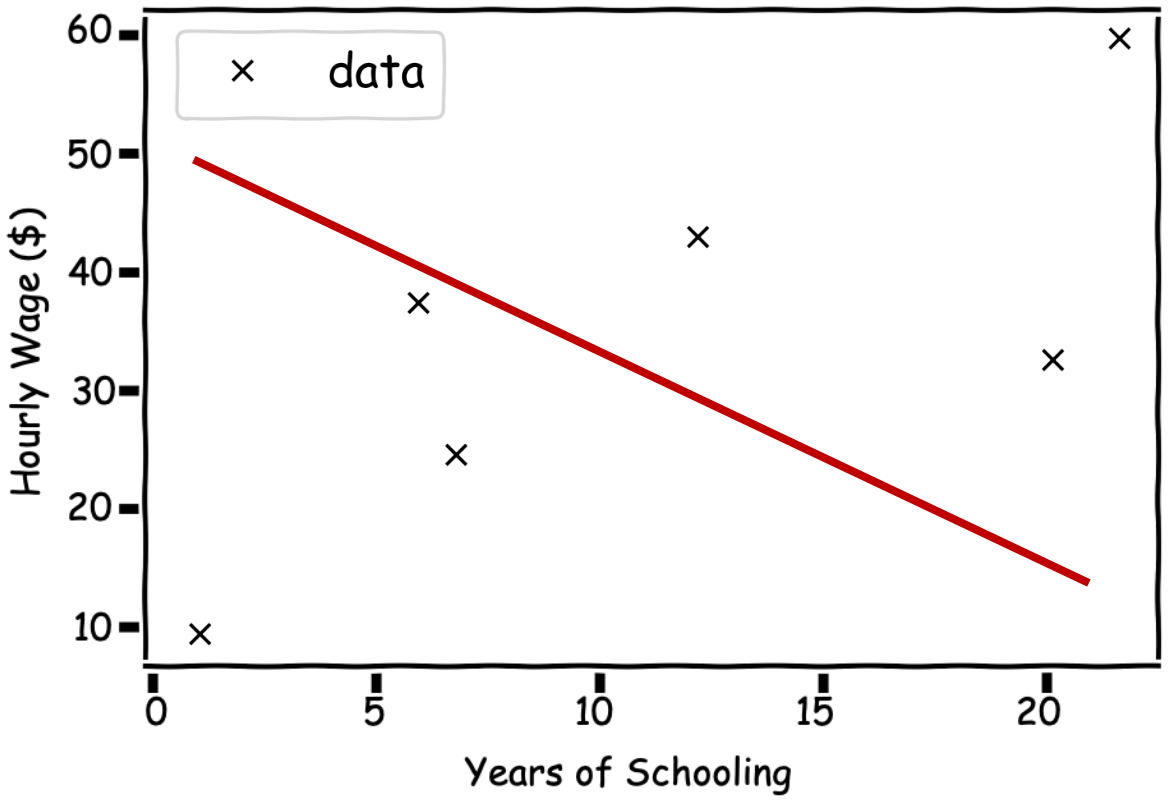
β_1: Slope Coefficient

ε: Error Term

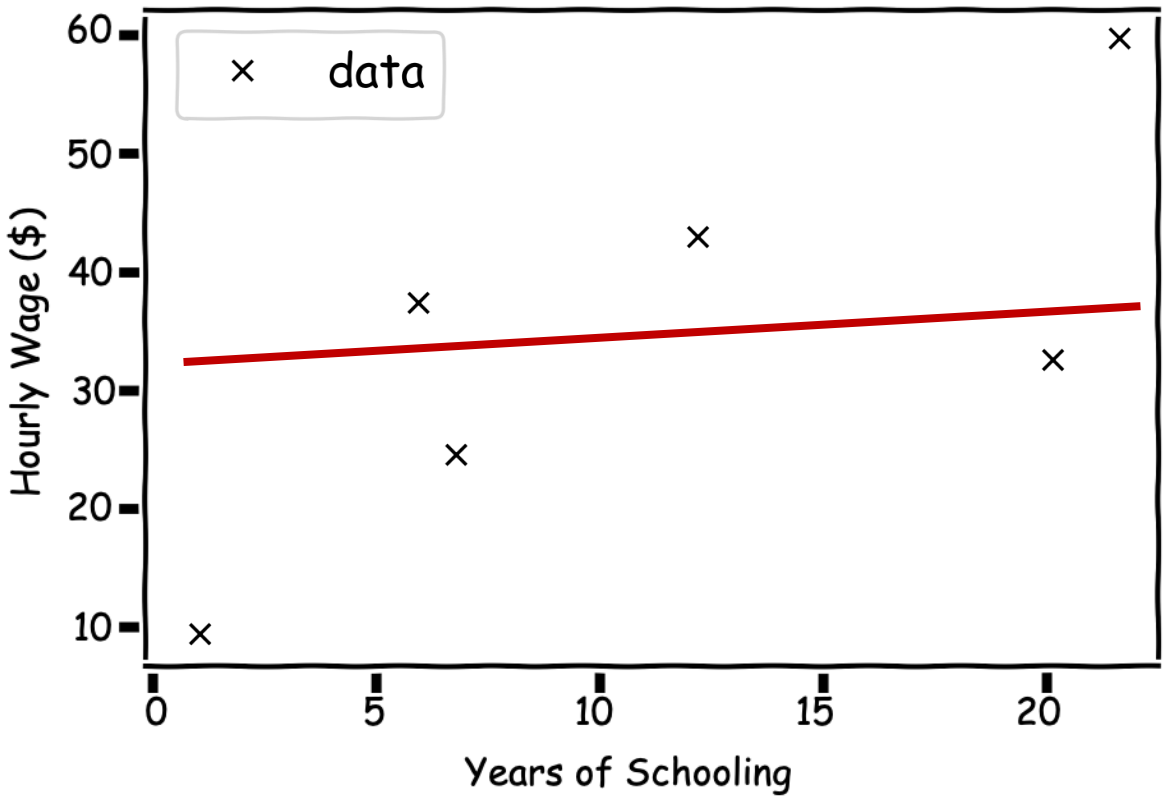
Plotting two continuous variables



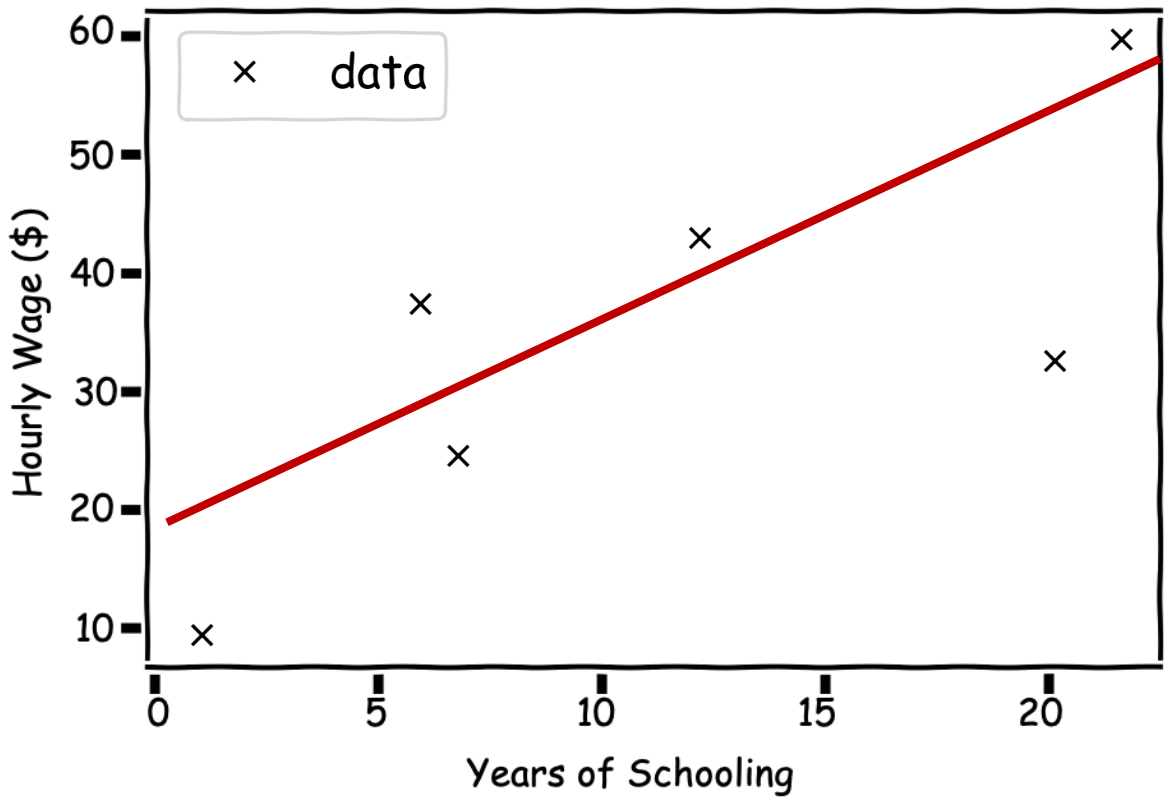
Finding the Line of Best Fit



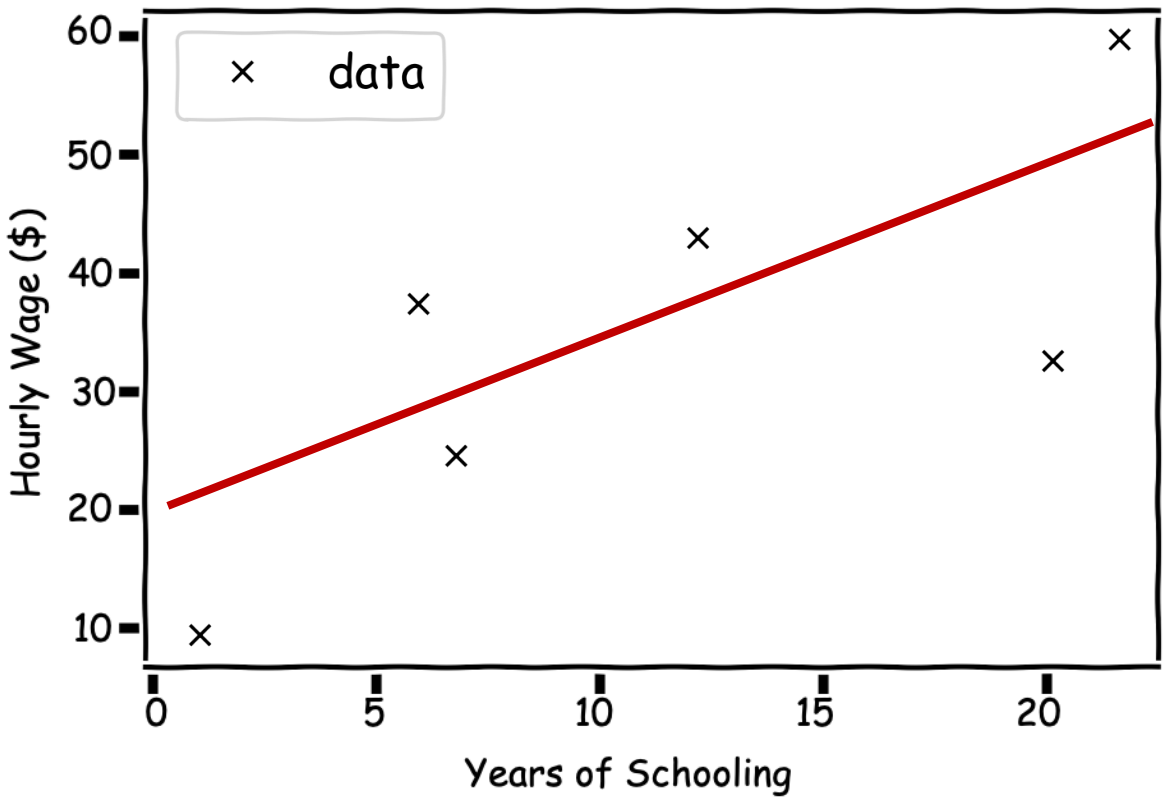
Finding the Line of Best Fit



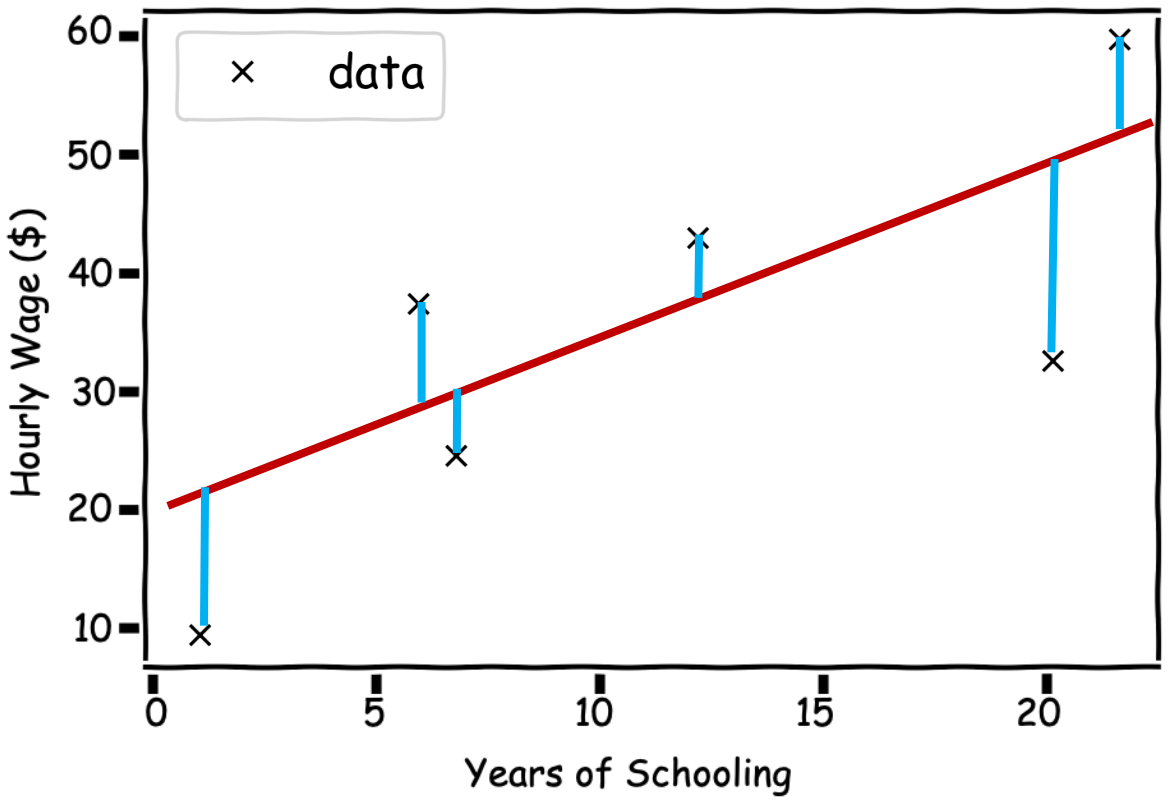
Finding the Line of Best Fit



Finding the Line of Best Fit



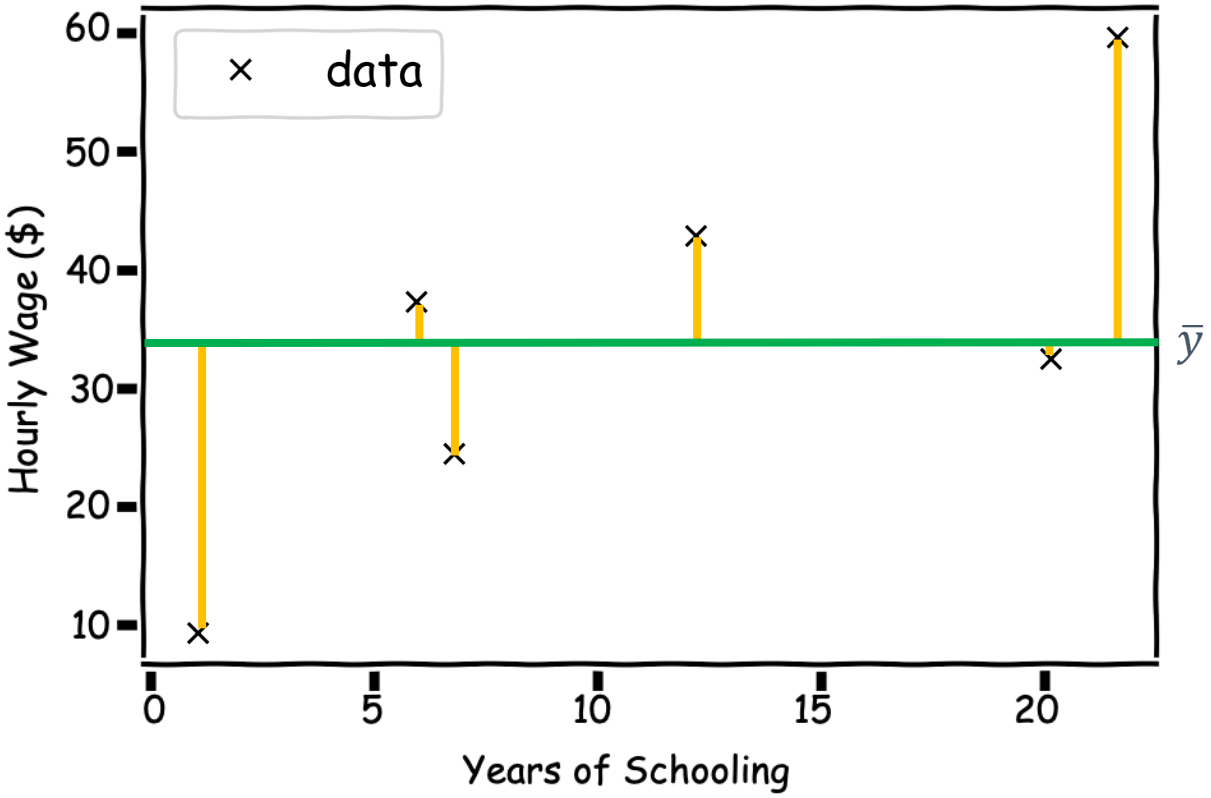
Assessing Fit using Residuals



2.1 Model Fit

The Total Sum of Squares (TSS)

TSS measures the extent of variability of observed data

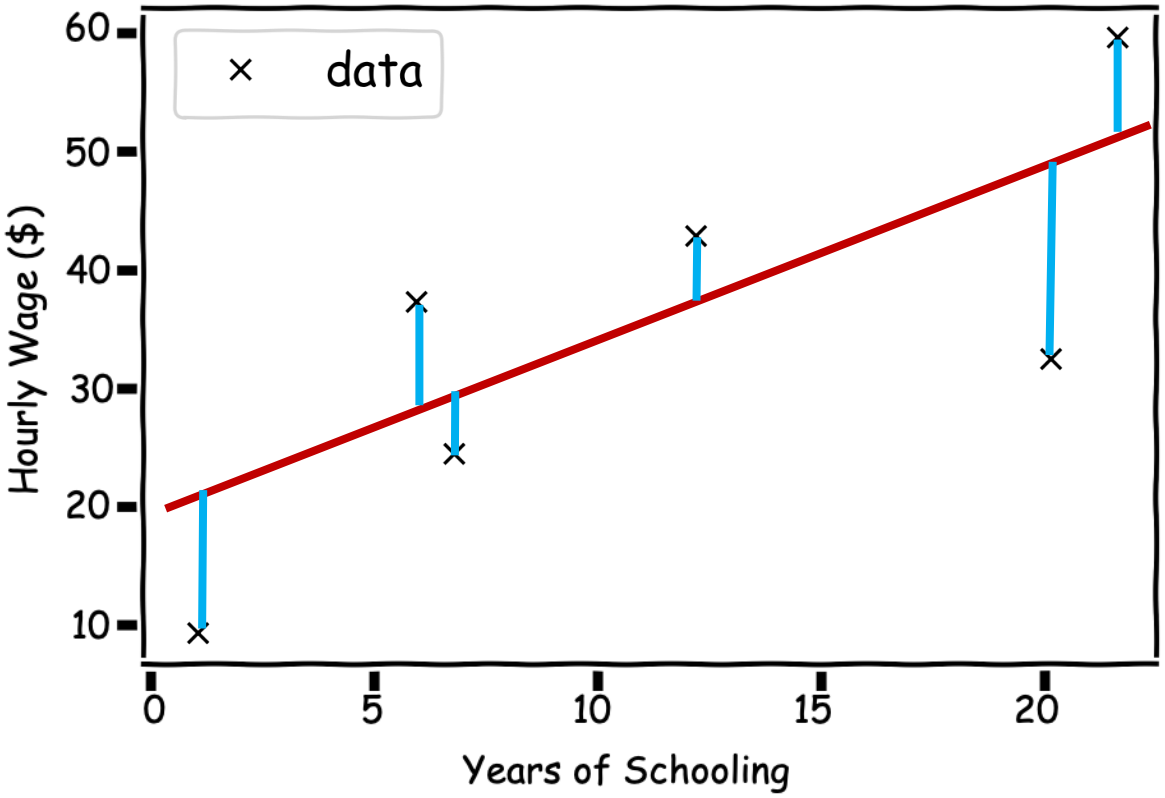


$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$TSS = (\quad + \quad + \quad + \quad + \quad)^2$$

The Residual Sum of Squares (RSS)

RSS measures the extent of variability of observed data not predicted by the model

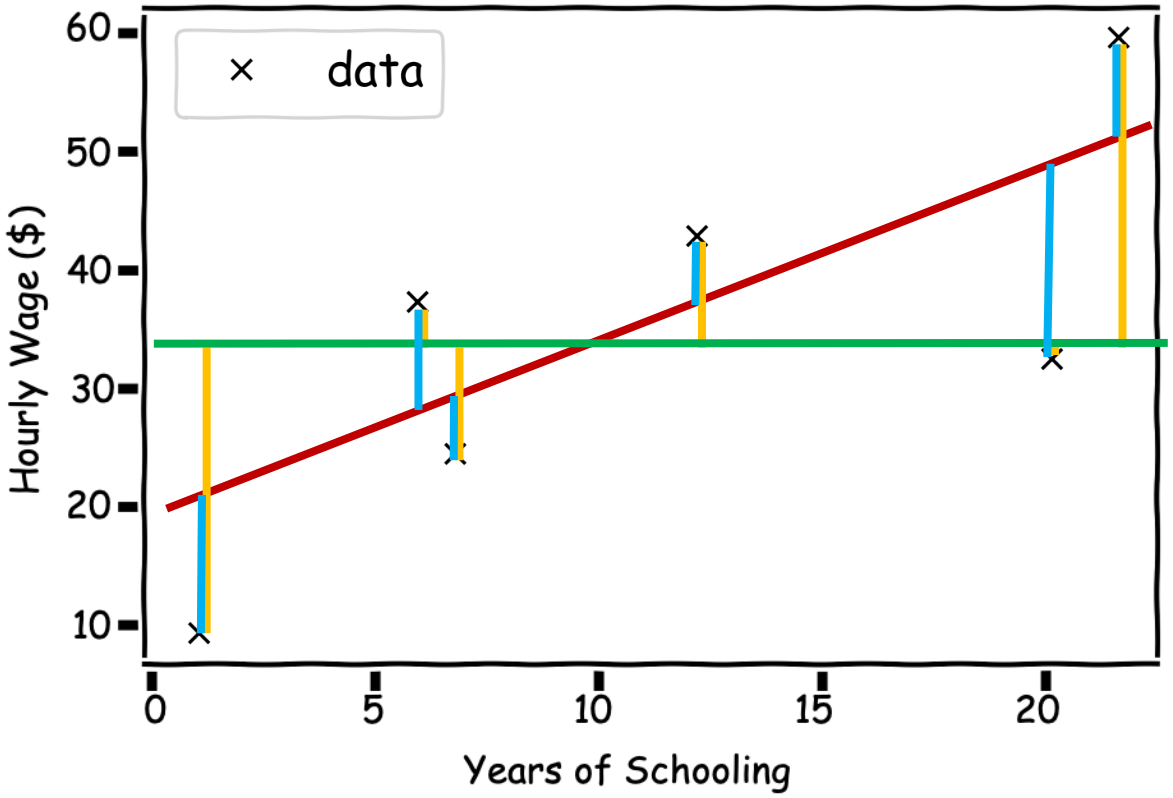


$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$RSS = (|+|+|+|+|+|)^2$$

R^2

R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable



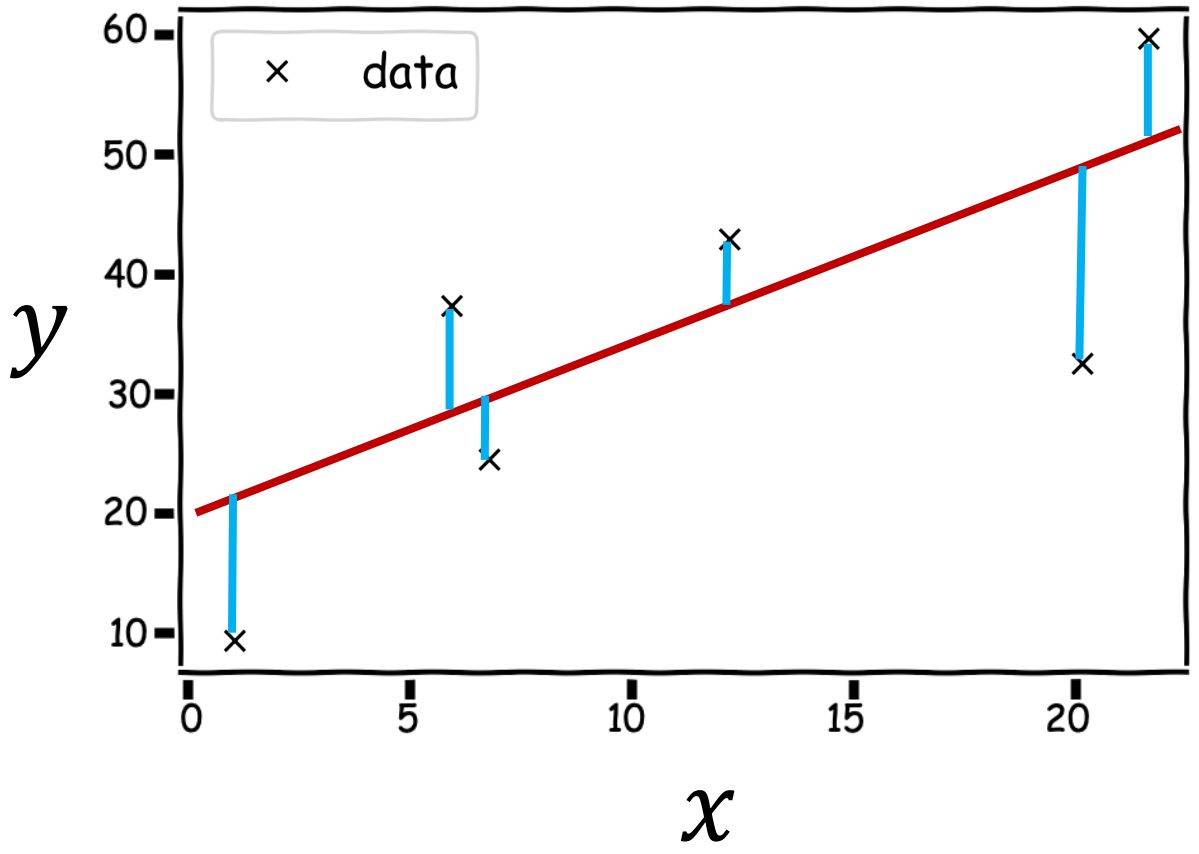
$$R^2 = 1 - \frac{RSS}{TSS}$$

$$R^2 = 1 - \frac{(|+|+|+|+|+|)^2}{(|+|+|+|+|+|)^2}$$

1.3 Regression Coefficients

Estimate of the regression coefficients

$$y = \beta_o + \beta_1 x + \varepsilon$$



Interpretation of Predictors

$$y = \beta_o + \beta_1 x + \varepsilon$$

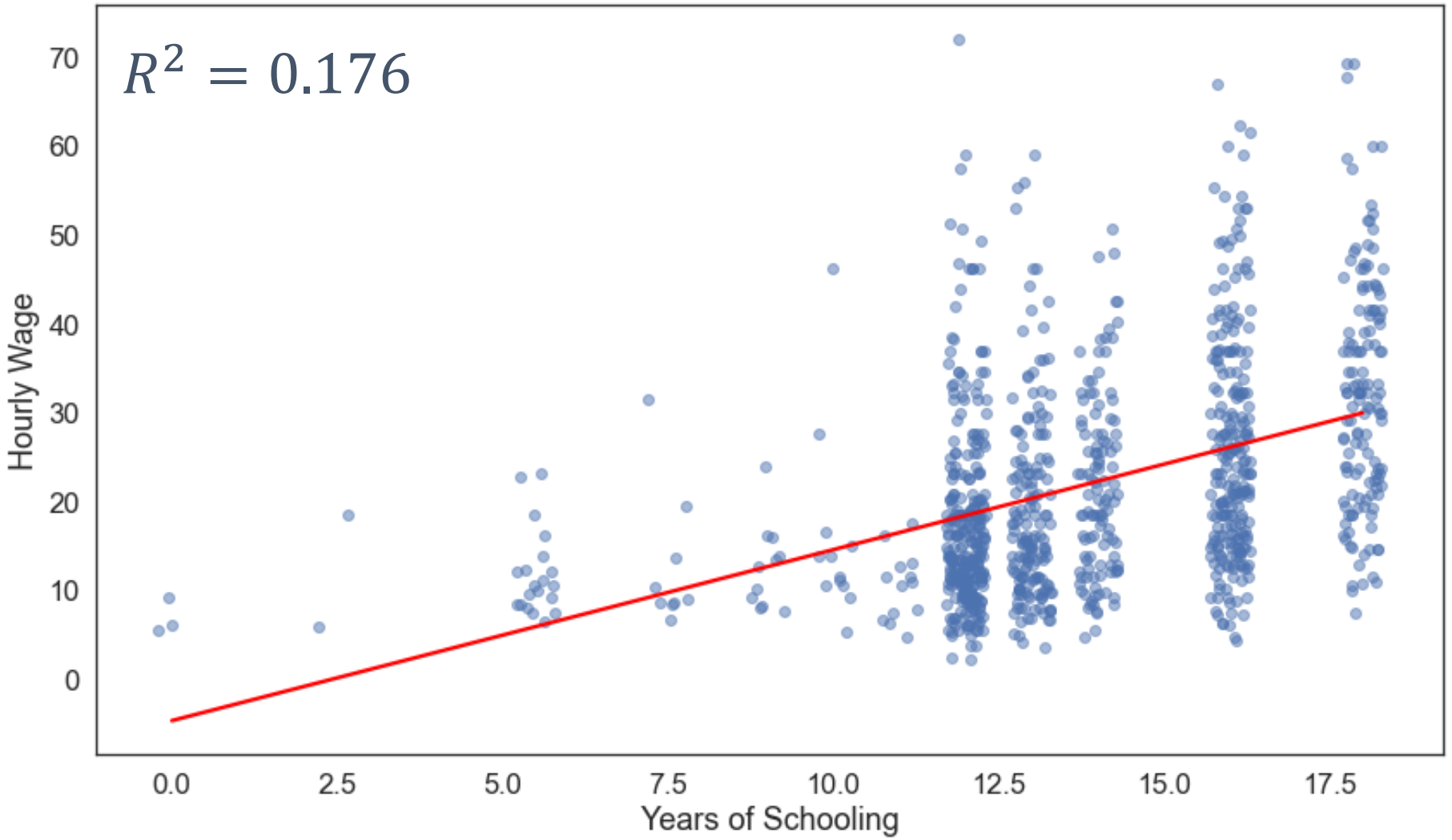
A one-unit increase in $\boldsymbol{x}$ leads to a $\boldsymbol{\beta_1}$ increase in $\boldsymbol{y}$.

According to the model, an observation with a value of x_i is predicted to have a y value of $(\beta_o + \beta_1 x_i)$.

Compared to an observation with a value of x_a, an observation with a value of x_b is associated with a $(\beta_o + \beta_1 x_a)$-$(\beta_o + \beta_1 x_b)$ change in y.

R2 percent of the variation in y can be explained by our model.

$$Hourly\ Wage = -5.51 + 1.93 \times Years\ of\ Schooling$$



Interpretation of Predictors

$$R^2 = 0.176$$

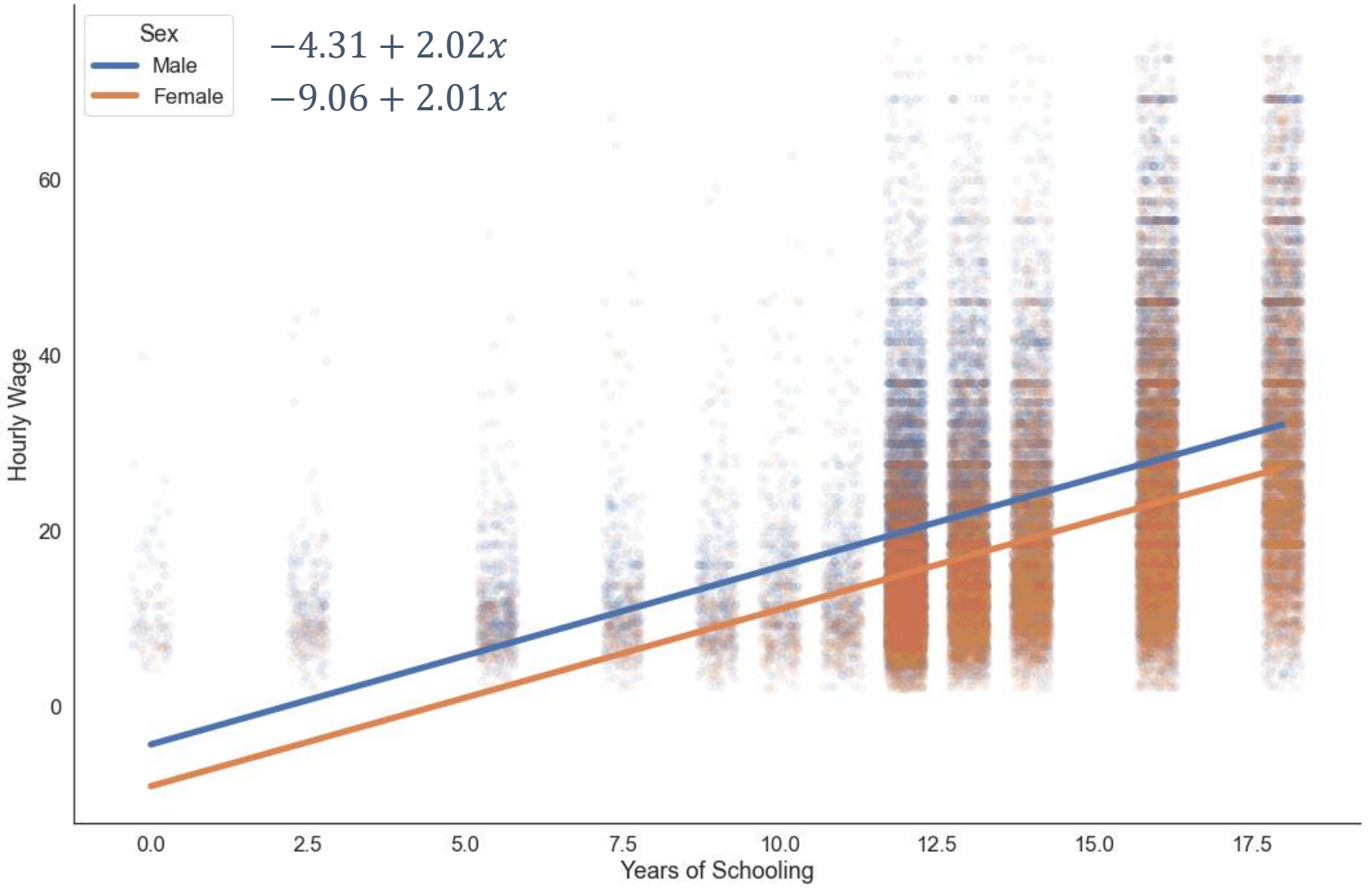
$$Hourly\ Wage = -5.51 + 1.93 \times Years\ of\ Schooling$$

Every additional year of schooling leads to a $\$1.93$ increase in hourly wages.

According to the model, an individual with 18 years of schooling is predicted to have an hourly wage of \$29.23. $(-5.51 + 1.93 \times 18)$

On average, a person with an Undergraduate degree makes \$7.72 more per hour than someone with a High School degree. $(-5.51 + 1.93 \times 12) - (-5.51 + 1.93 \times 16)$

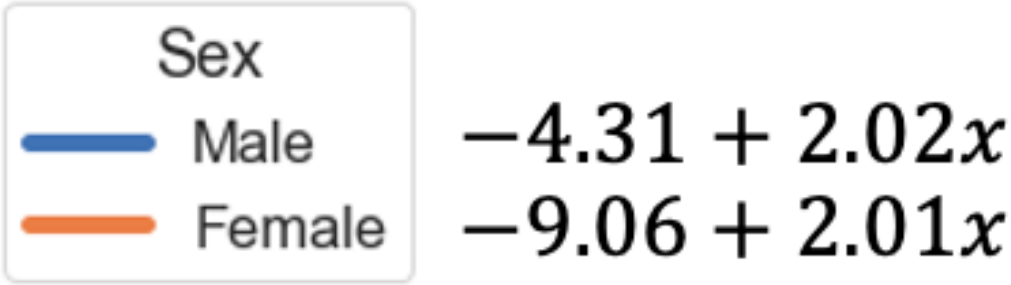
17.6 percent of the variation in hourly income can be explained by an individual's years of schooling.

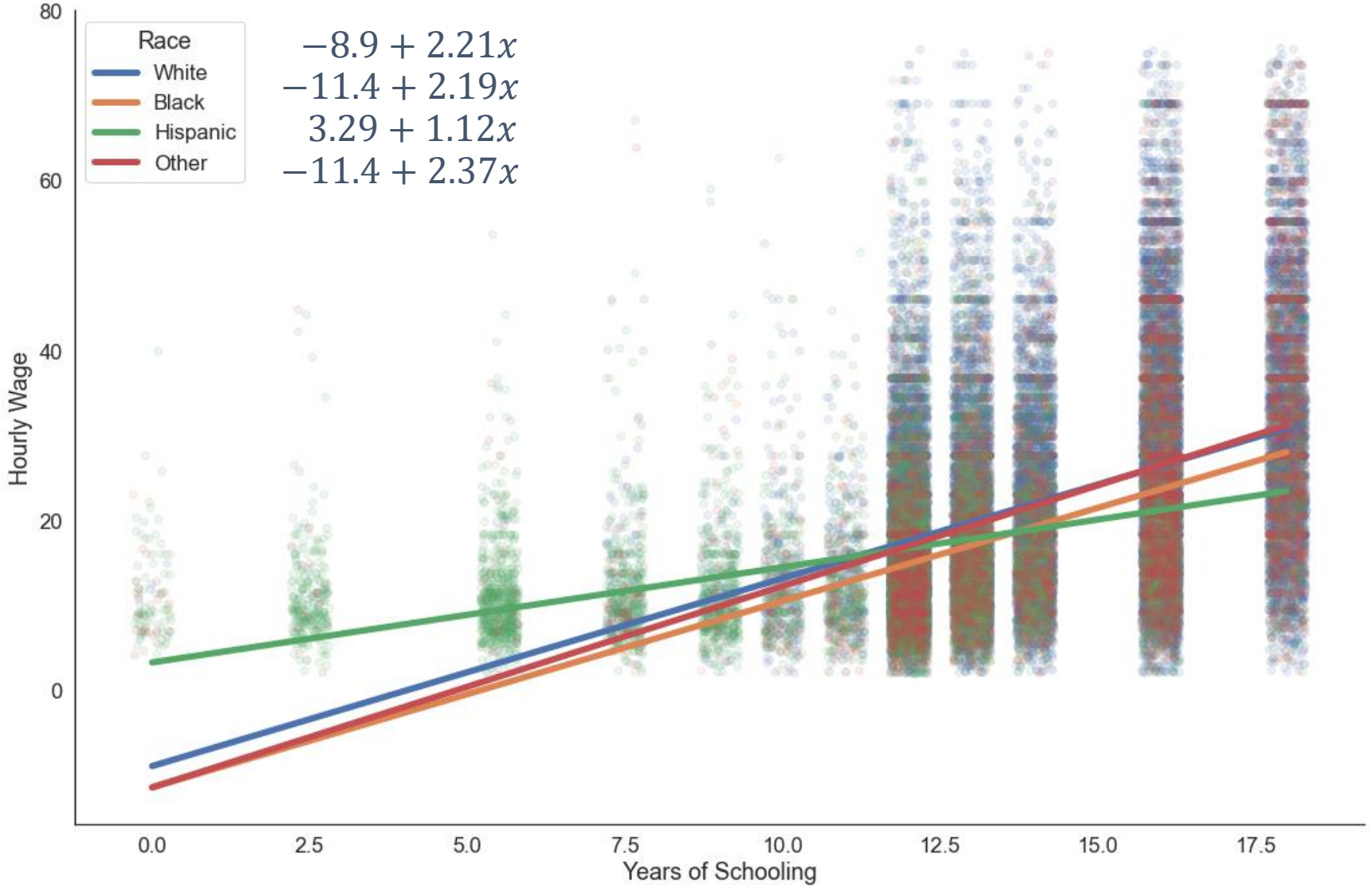

Sex
Male
Female
$-4.31 + 2.02x$
$-9.06 + 2.01x$
Hourly Wage
60
40
20
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Years of Schooling

Making Predictions

On average, a man with a high school diploma makes $19.93 per hour.
$(-4.31 + 2.02 \times 12)$

On average, a woman with a high school diploma makes $15.06 per hour.
$(-9.06 + 2.01 \times 12)$



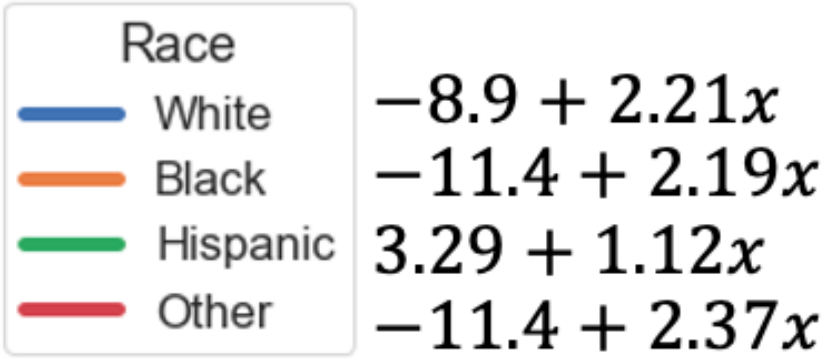

Race
White
Black
Hispanic
Other
$-8.9 + 2.21x$
$-11.4 + 2.19x$
$3.29 + 1.12x$
$-11.4 + 2.37x$
Hourly Wage
80
60
40
20
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Years of Schooling

Making Predictions

Compared to those with a high school diploma, the hourly income gains associated with getting a PhD are:

- 75.25% for white Americans. $\frac{(-8.9+2.21\times18)-(-8.9+2.21\times12)}{(-8.9+2.21\times12)}$
- 40.16% for Hispanic Americans. $\frac{(3.29+1.12\times18)-(3.29+1.12\times12)}{(3.29+1.12\times12)}$

Hispanic Americans experience lower returns to education than other racial groups. (smallest β_1)



Hispanics and the Future of America.

< Prev Next >

Show details

Contents Hardcopy Version at National Academies Press

Search this book

6 Barriers to Educational Opportunities for Hispanics in the United States

Barbara Schneider, Sylvia Martinez, and Ann Owens.

For Hispanics in the United States, the educational experience is one of accumulated disadvantage. Many Hispanic students begin formalized schooling without the economic and social resources that many other students receive, and schools are often ill equipped to compensate for these initial disparities. For Hispanics, initial disadvantages often stem from parents' immigrant and socioeconomic status and their lack of knowledge about the U.S. education system. As Hispanic students proceed through the schooling system, inadequate school resources and their weak relationships with their teachers continue to undermine their academic success. Initial disadvantages continue to accumulate, resulting in Hispanics having the lowest rates of high school and college degree attainment, which hinders their chances for stable employment. The situation of Hispanic educational attainment is cause for national concern.

2.4 Standard Errors

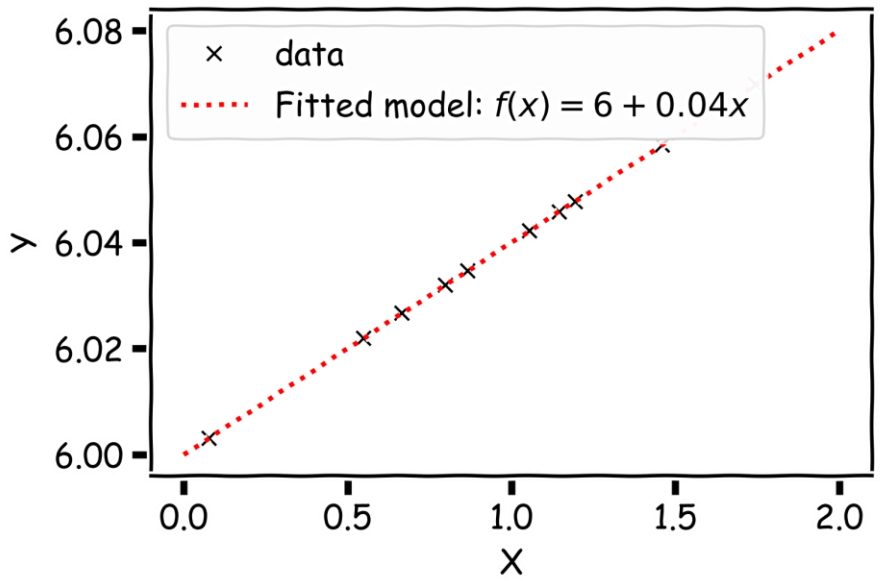
Confidence intervals for the coefficients

We interpret the ε term in our observation

$$y = f(x) + \varepsilon$$

to be noise introduced by random variations in natural systems or imprecisions of our scientific instruments.

If we knew the exact form of $f(x)$, for example, $f(x) = \beta_0 + \beta_1 x$, and there was no ε , then estimating the $\hat{\beta}'s$ would have been exact.



Confidence intervals for the coefficients

However, three things happen, which result in mistrust of the values of $\hat{\beta}'s$:

1. $\boldsymbol{\varepsilon}$ is always there
2. we do not know the exact form of $f(x)$
3. limited sample size

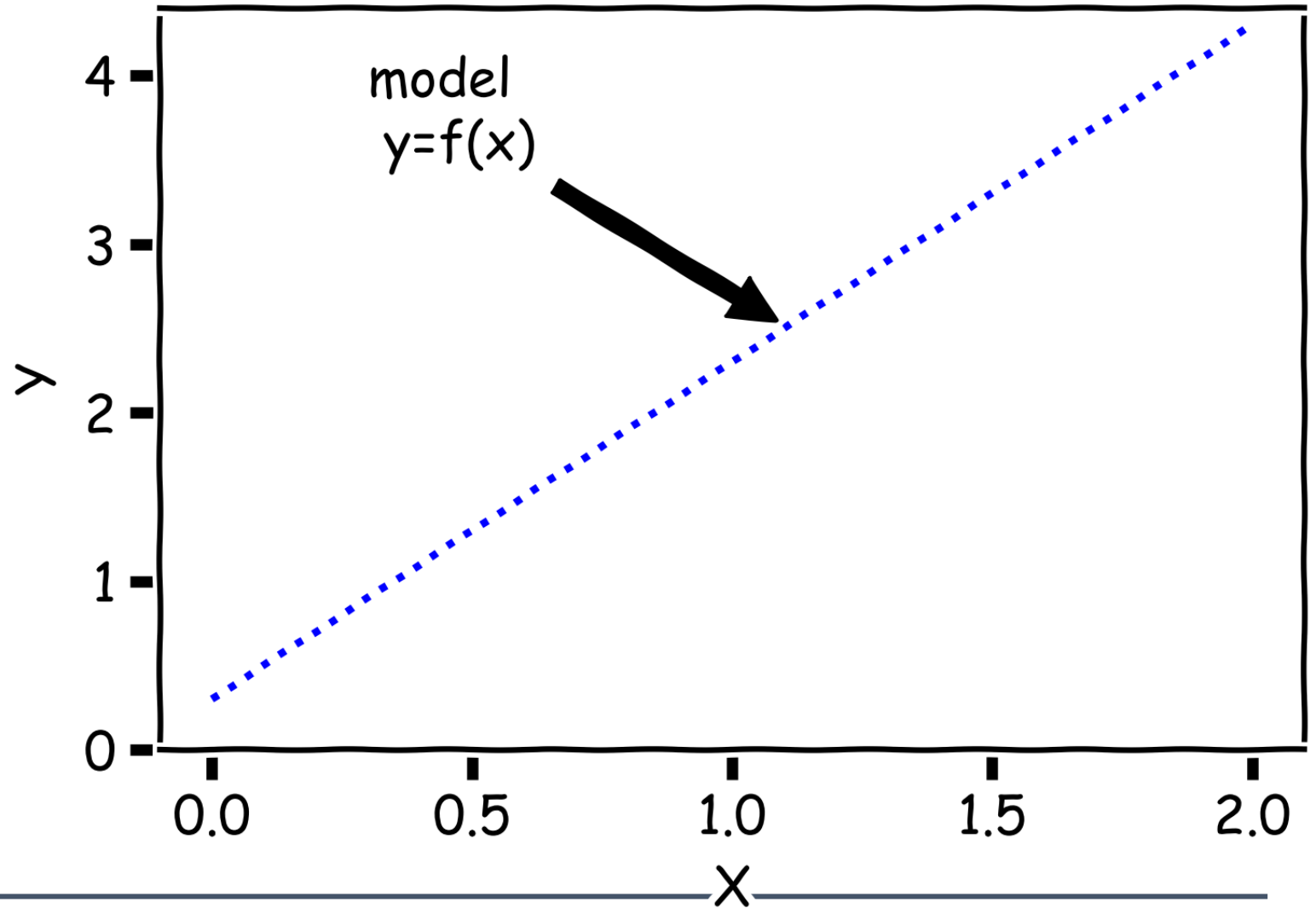
We will first address ε

We call ε the measurement error or ***irreducible error***. Since even predictions made with the actual function f will not match observed values of y.

Because of ε, every time we measure the response Y for a fixed value of X, we will obtain a different observation, and hence a different estimate of $\hat{\beta}'s$.

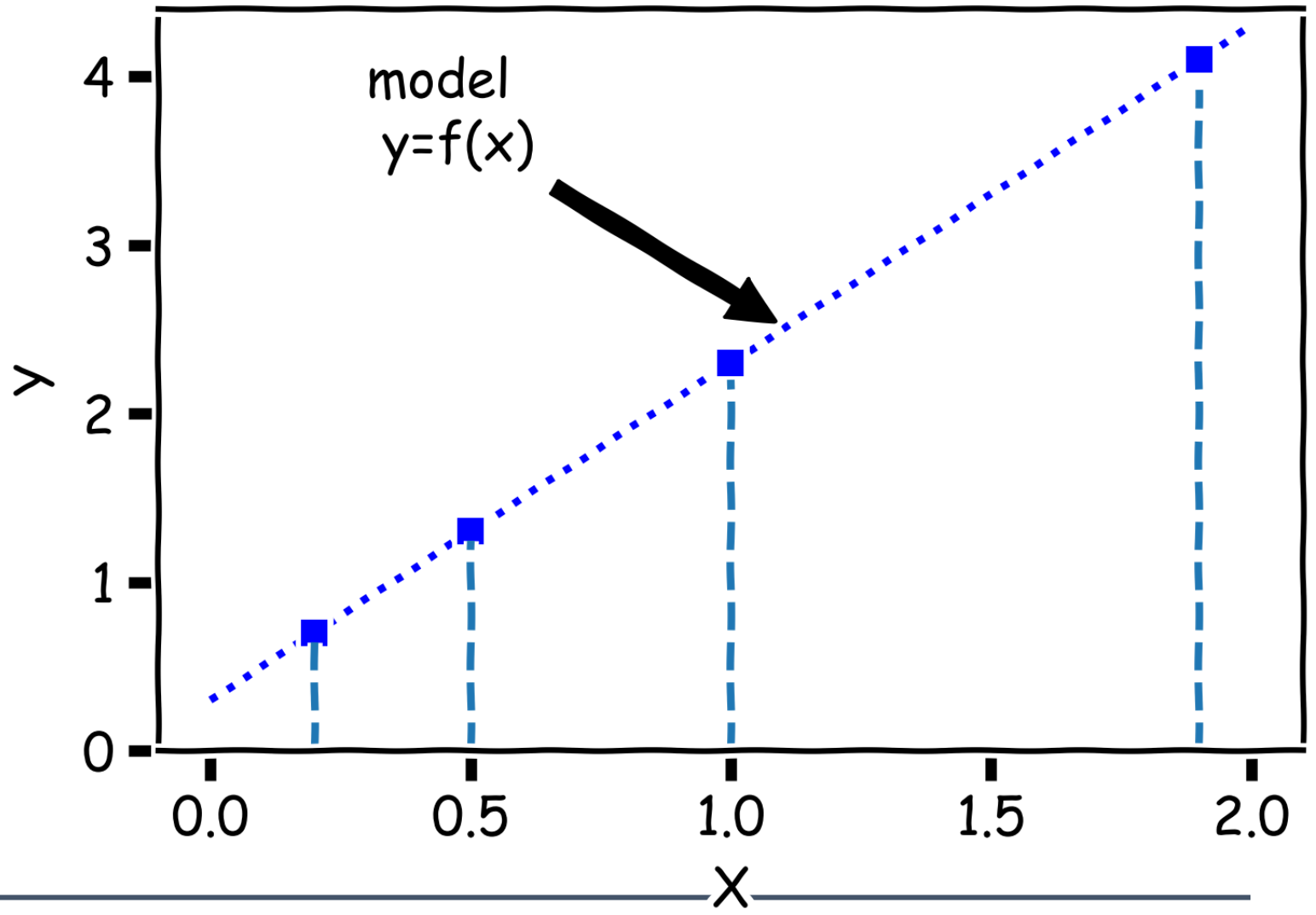
Confidence intervals for the coefficients

Start with a model



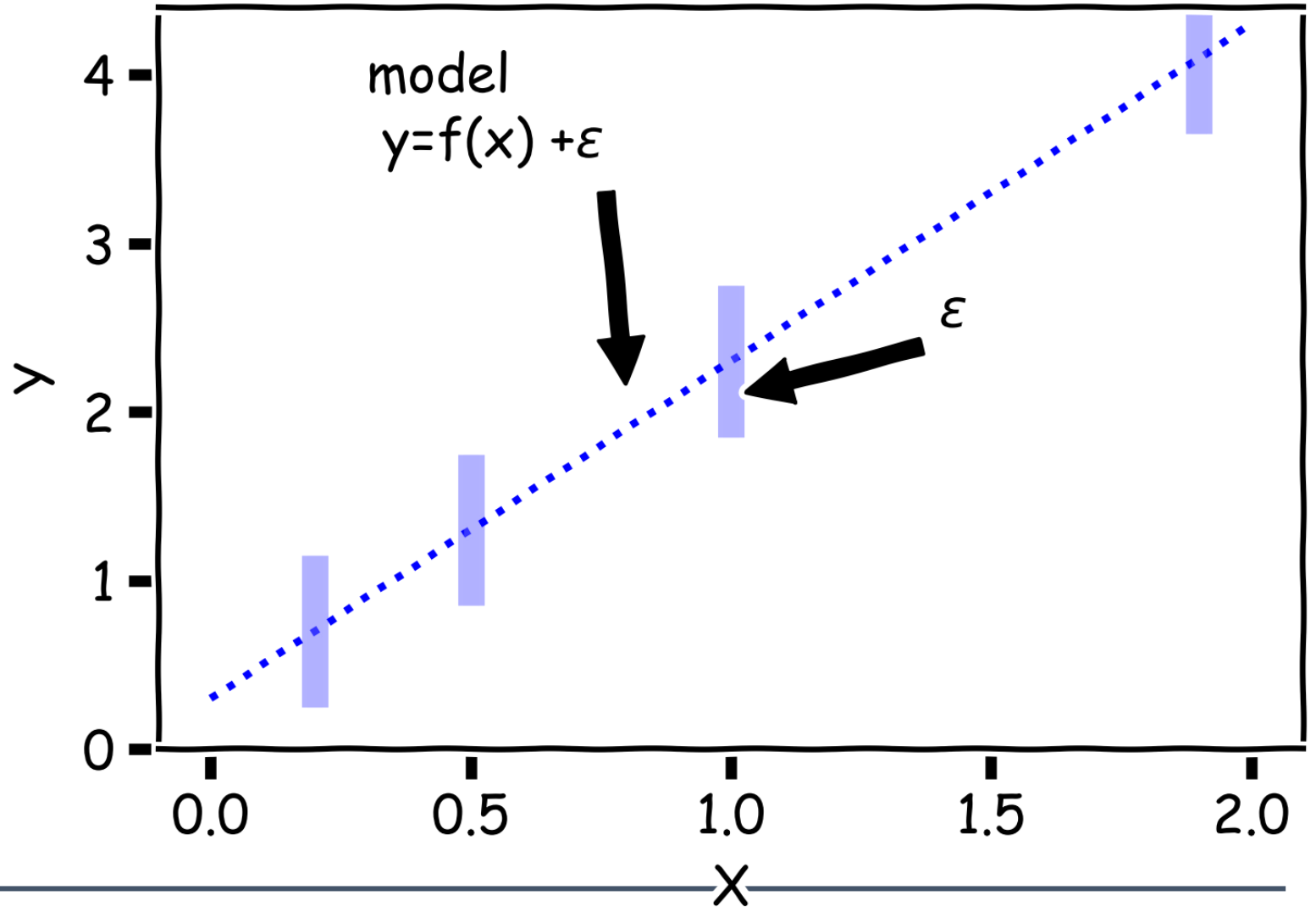
Confidence intervals for the coefficients

For some values of
$X, Y = f(X)$



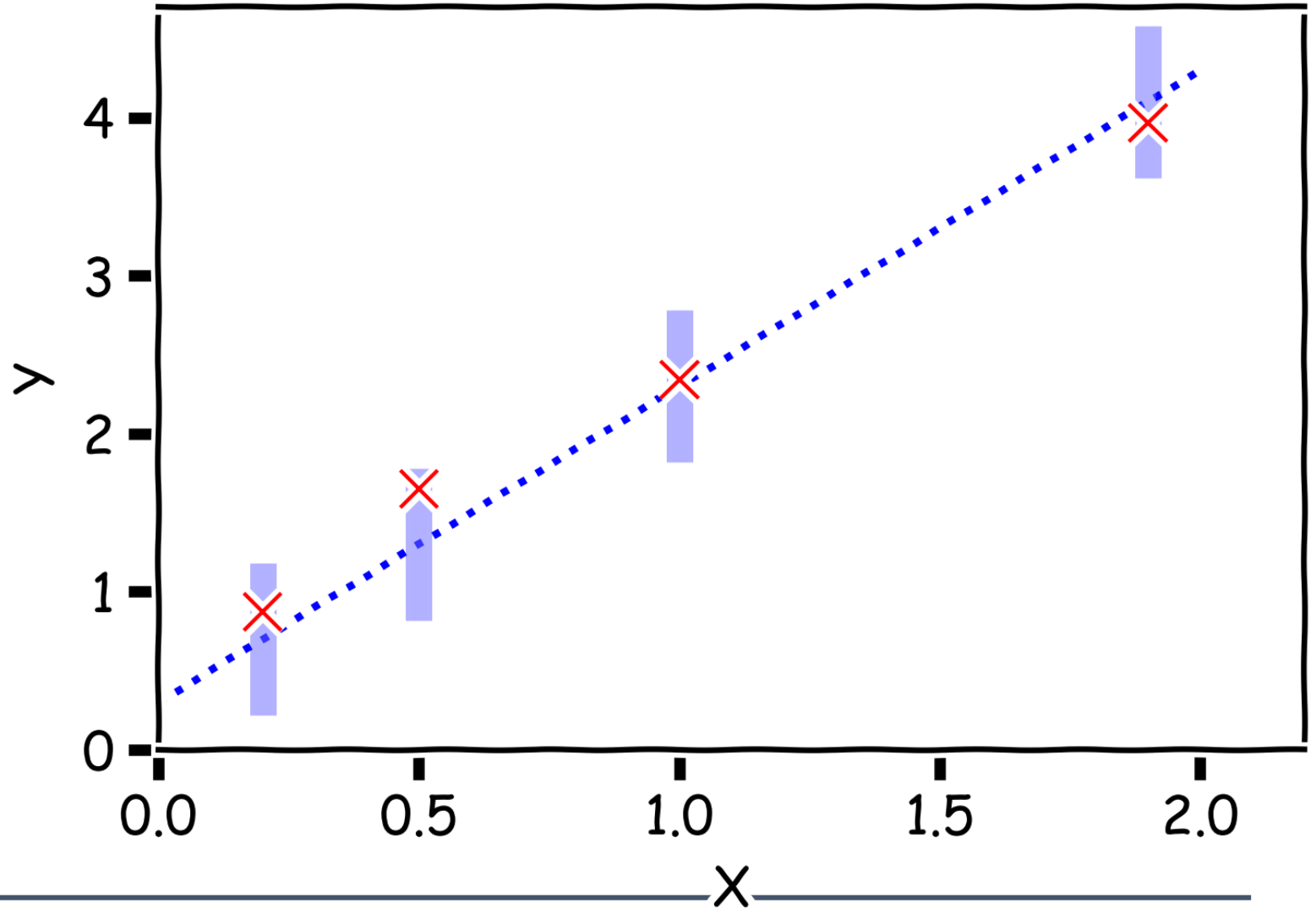
Confidence intervals for the coefficients

But due to error, every time we measure the response *Y* for a fixed value of X we will obtain a different observation.



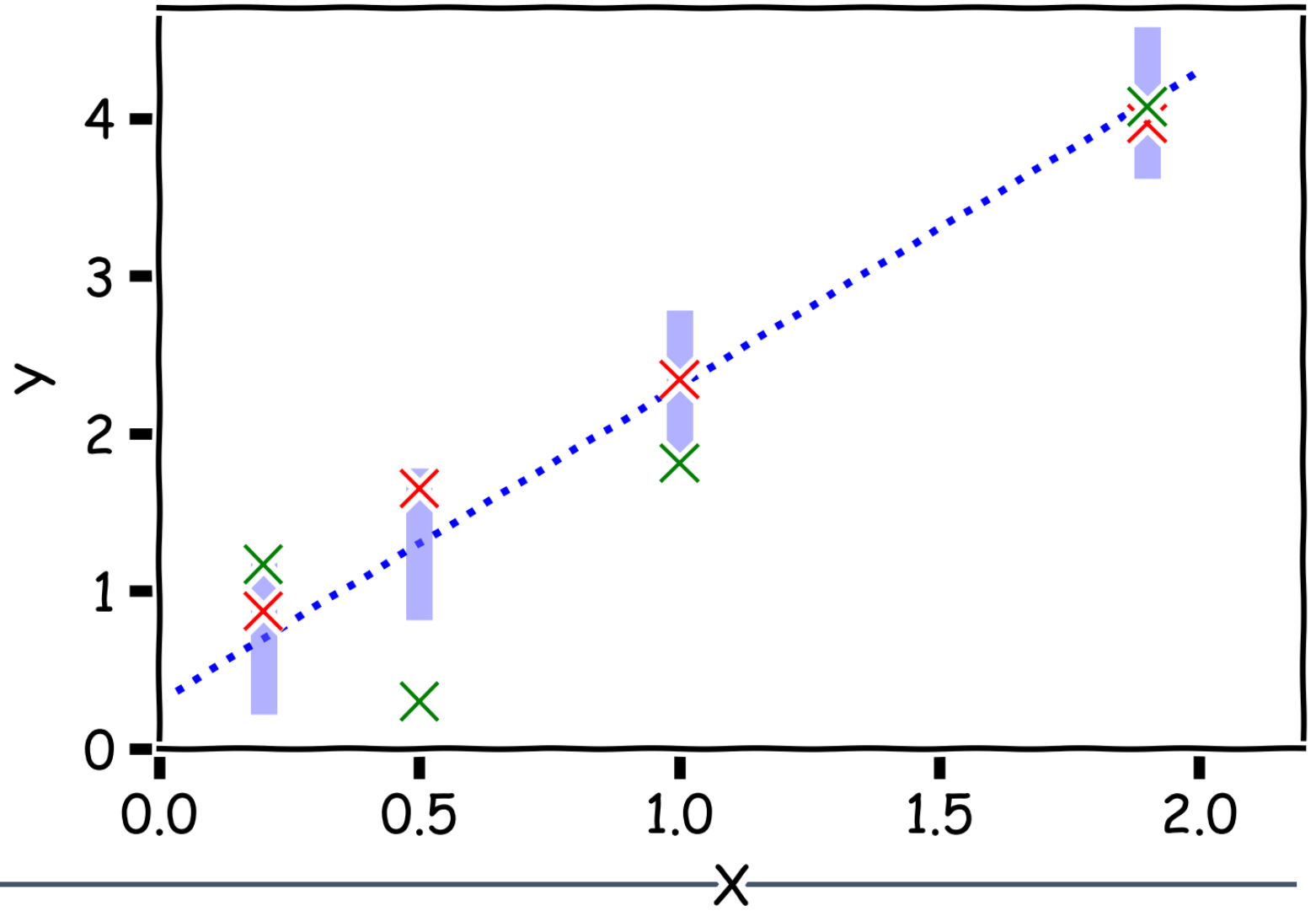
Confidence intervals for the coefficients

One set of observations, "one realization" we obtain one set of *Y*s (red crosses).



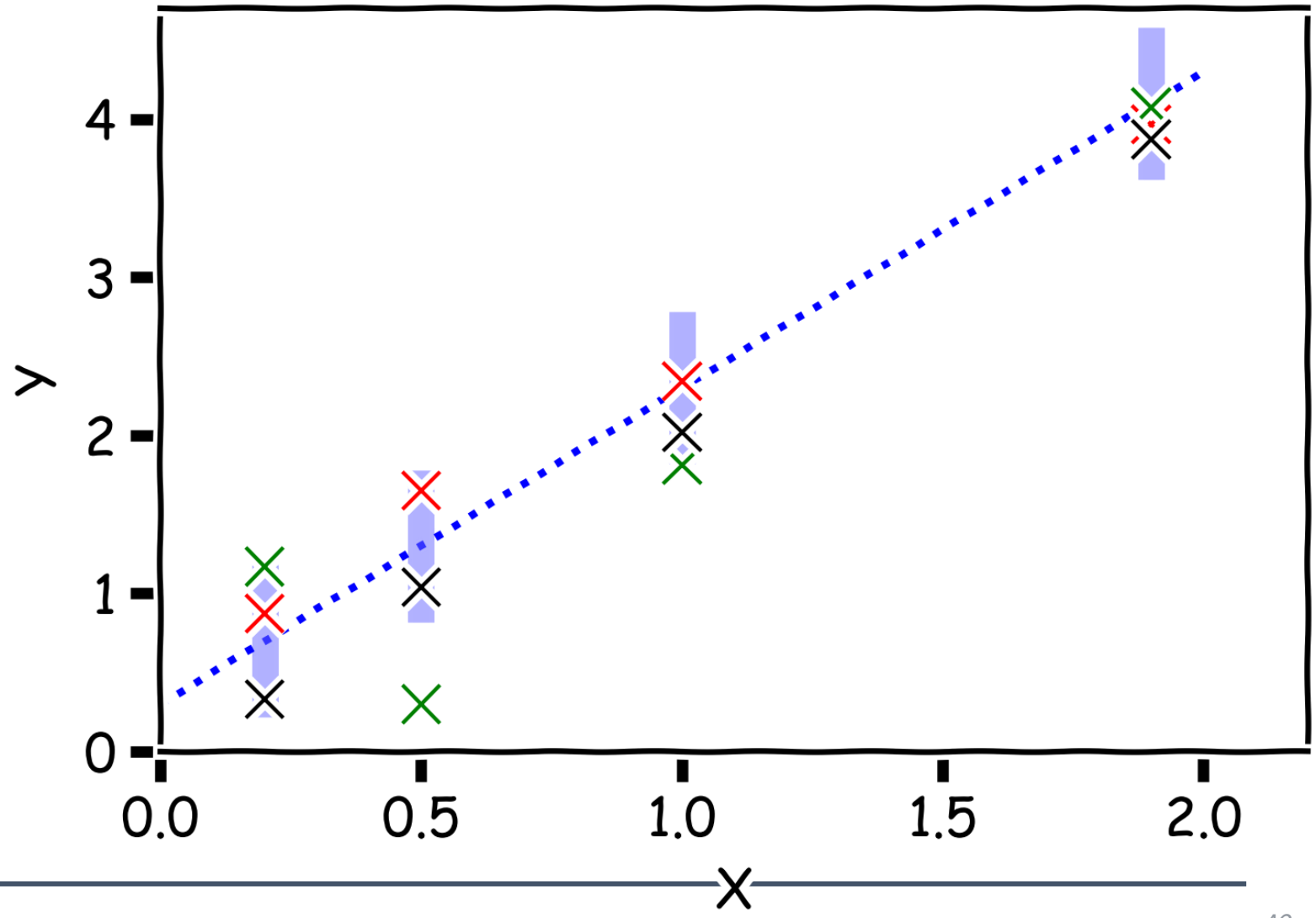
Confidence intervals for the coefficients

Another set of observations, “another realization” we obtain another set of *Y*s (green crosses).



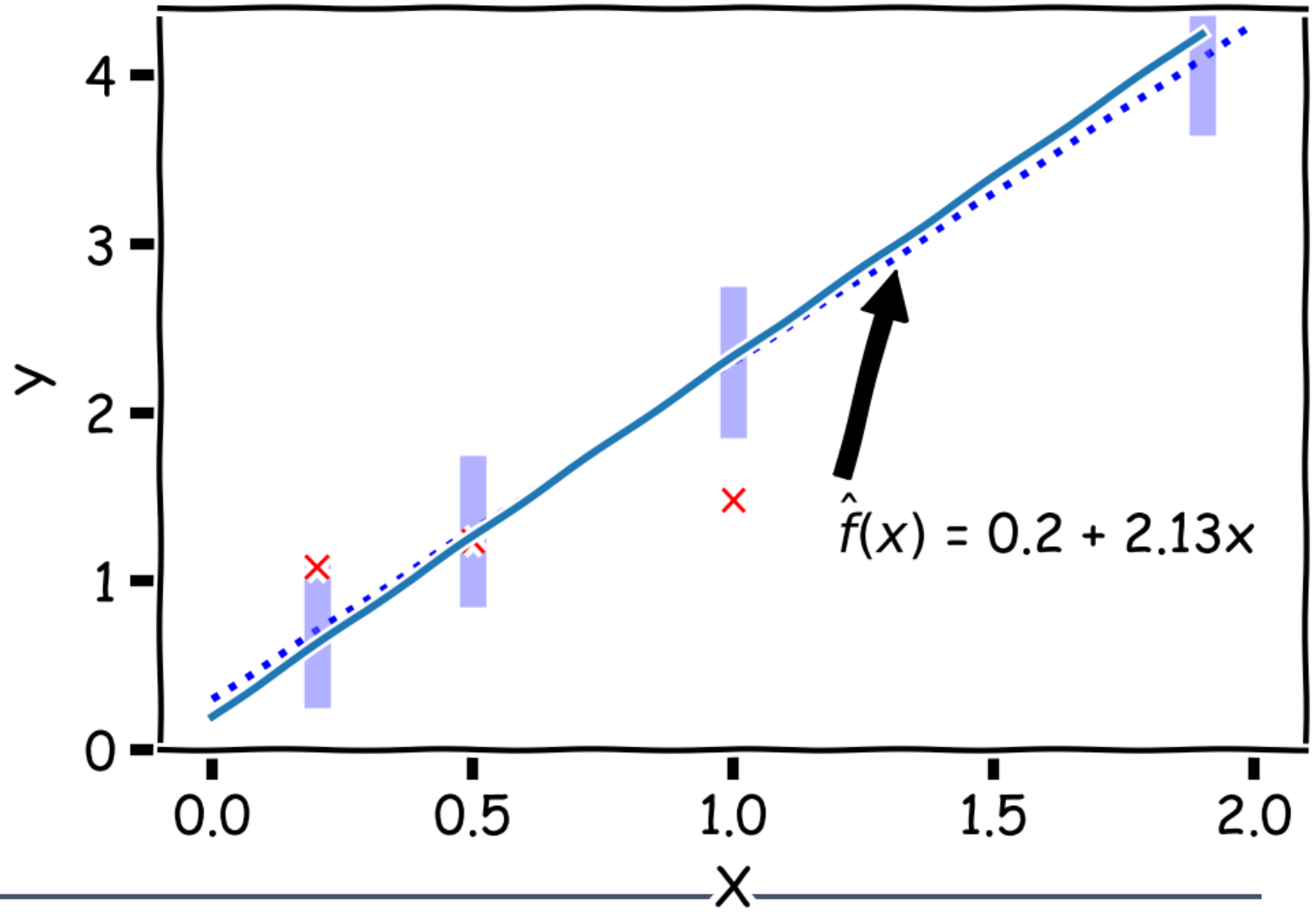
Confidence intervals for the coefficients

Another set of observations, “another realization” we obtain another set of *Y*s (black crosses).



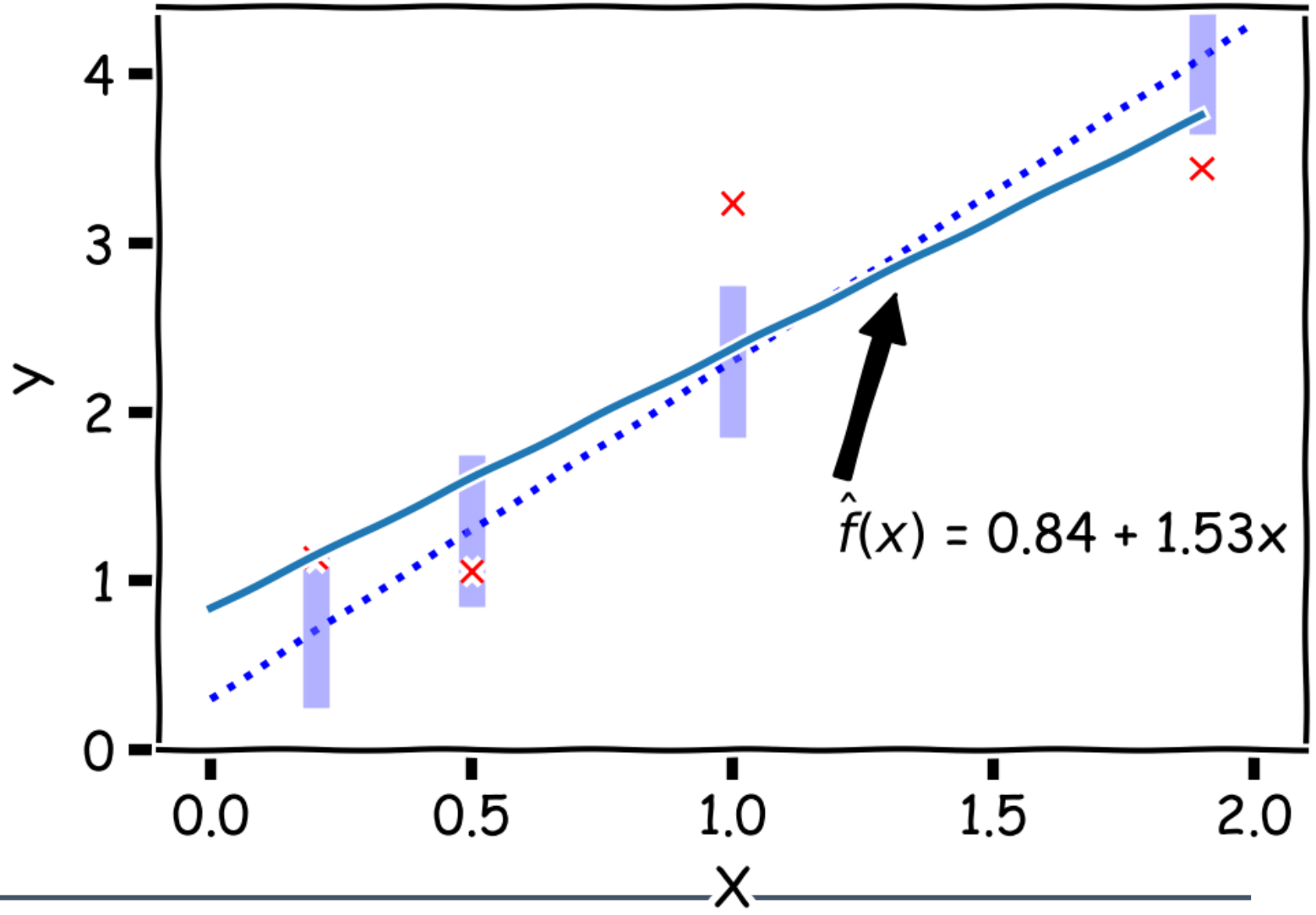
Confidence intervals for the coefficients

For each one of those "realizations", we could fit a model and estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.



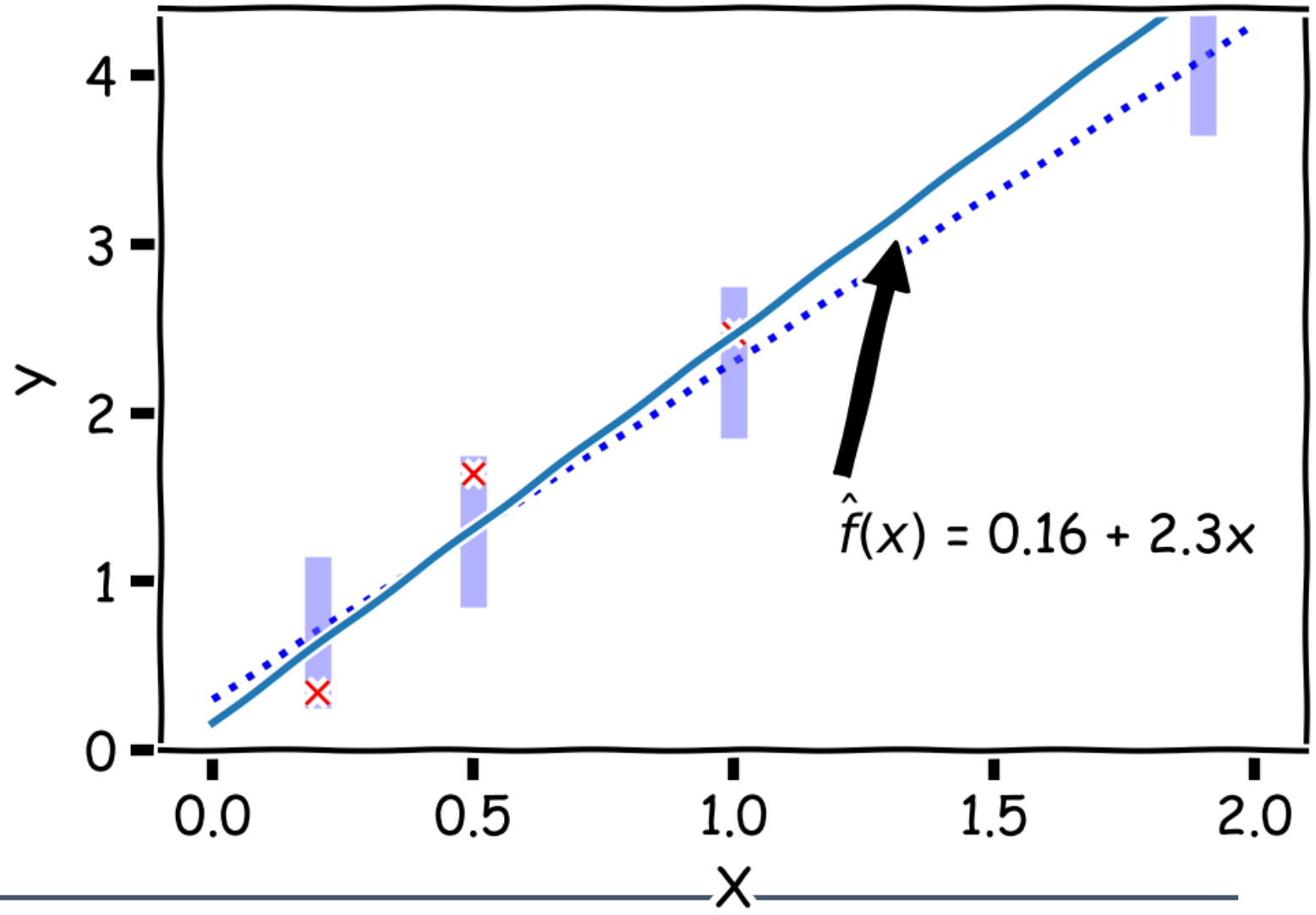
Confidence intervals for the coefficients

For each one of those "realizations", we could fit a model and estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$.



Confidence intervals for the coefficients

For each one of those "realizations", we could fit a model and estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$.

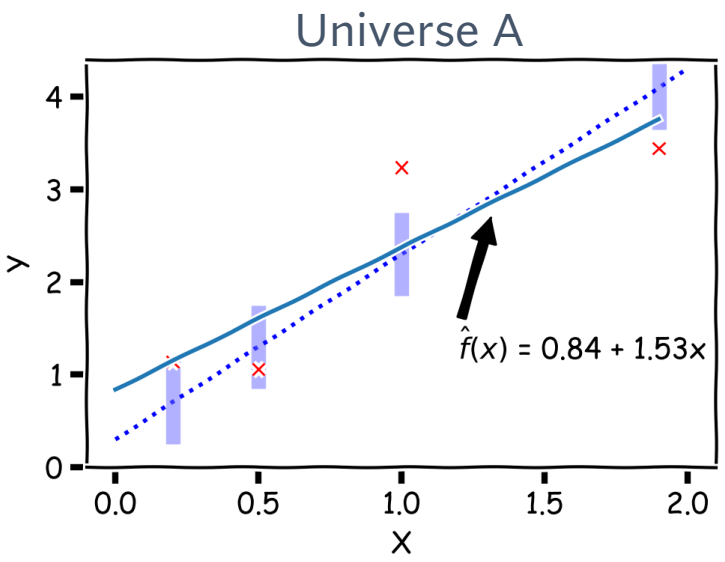


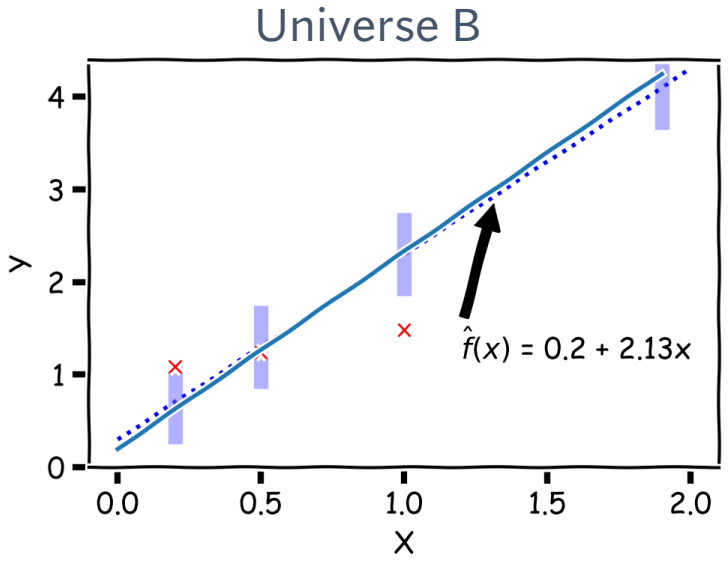
Confidence intervals for the coefficients

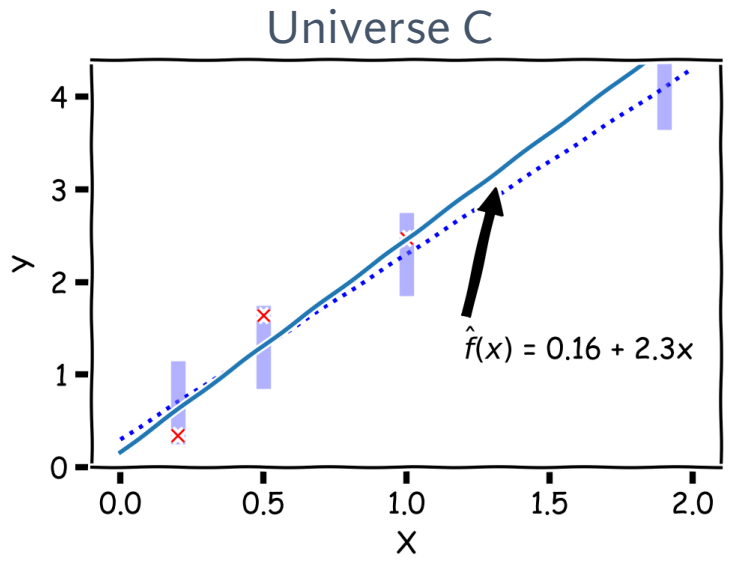
So if we just have one set of measurements of $\{X, Y\}$, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are just for this particular realization.

Question: If this is just one realization of the reality how do we know the truth? How do we deal with this conundrum?

Imagine (magic realism) we have parallel universes and we repeat this experiment on each of the other universes.







Bootstrapping for Estimating Sampling Error

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties by, for example, sampling from the observed data.

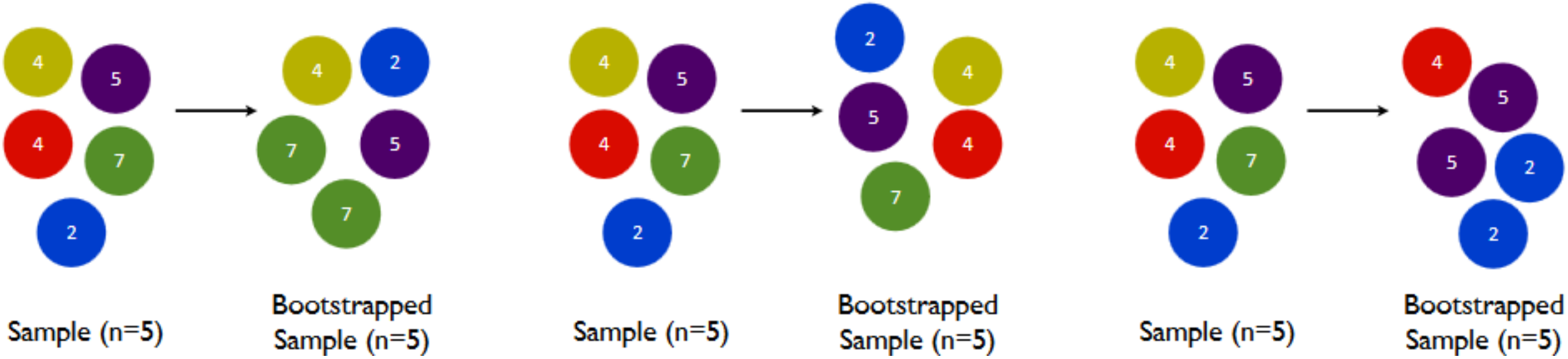
For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly sampling from our data set. We then use the variance of our multiple estimates to approximate the true variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Bootstrapping with replacement

Imagine a bowl full of marbles.

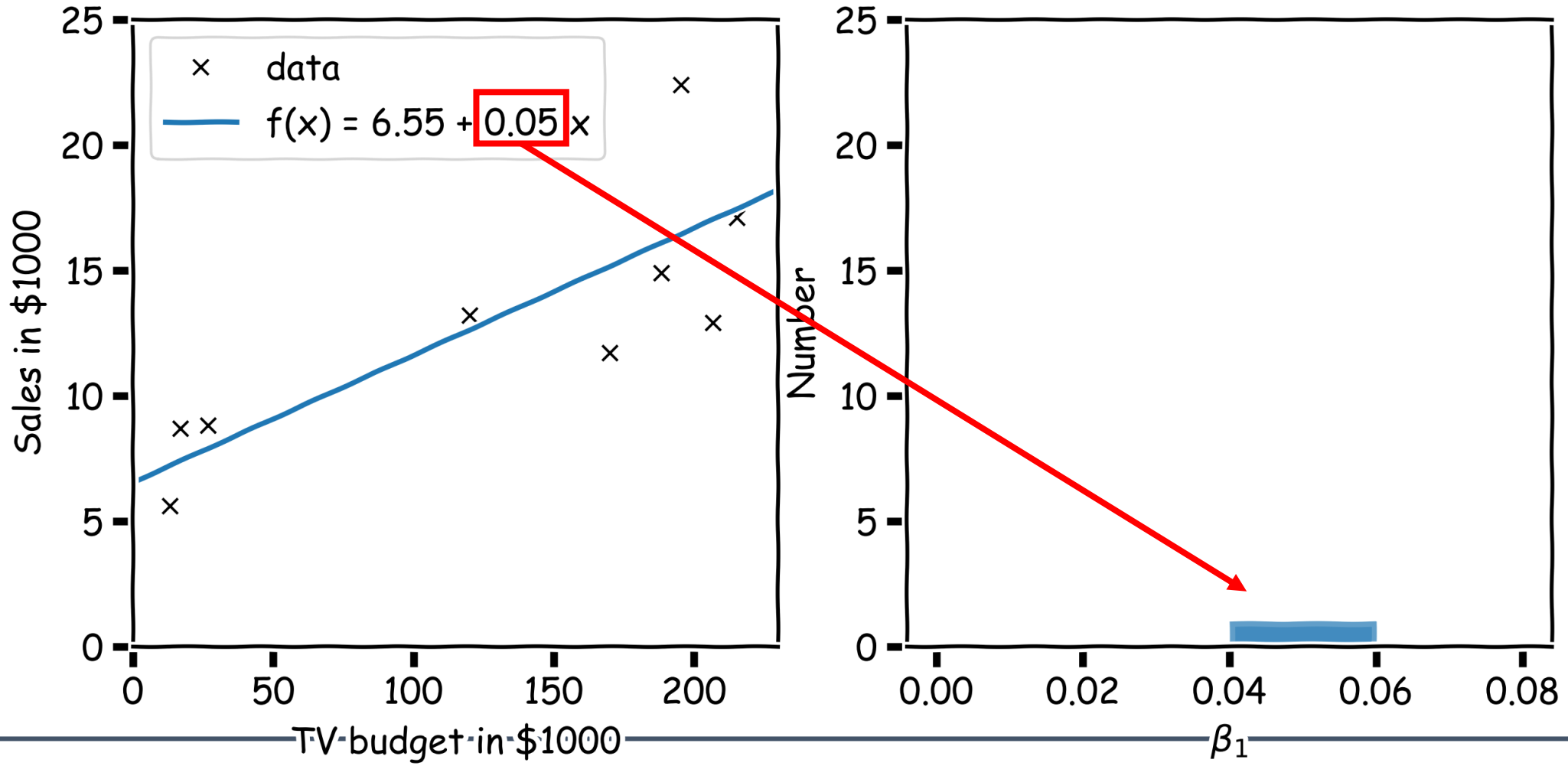
Sampling the marbles one-by-one, taking single marble out of the bowl, signing down it's color in your notebook and then *returning it back* to the bowl.

So when sampling with replacement the same marble can be sampled multiple times.

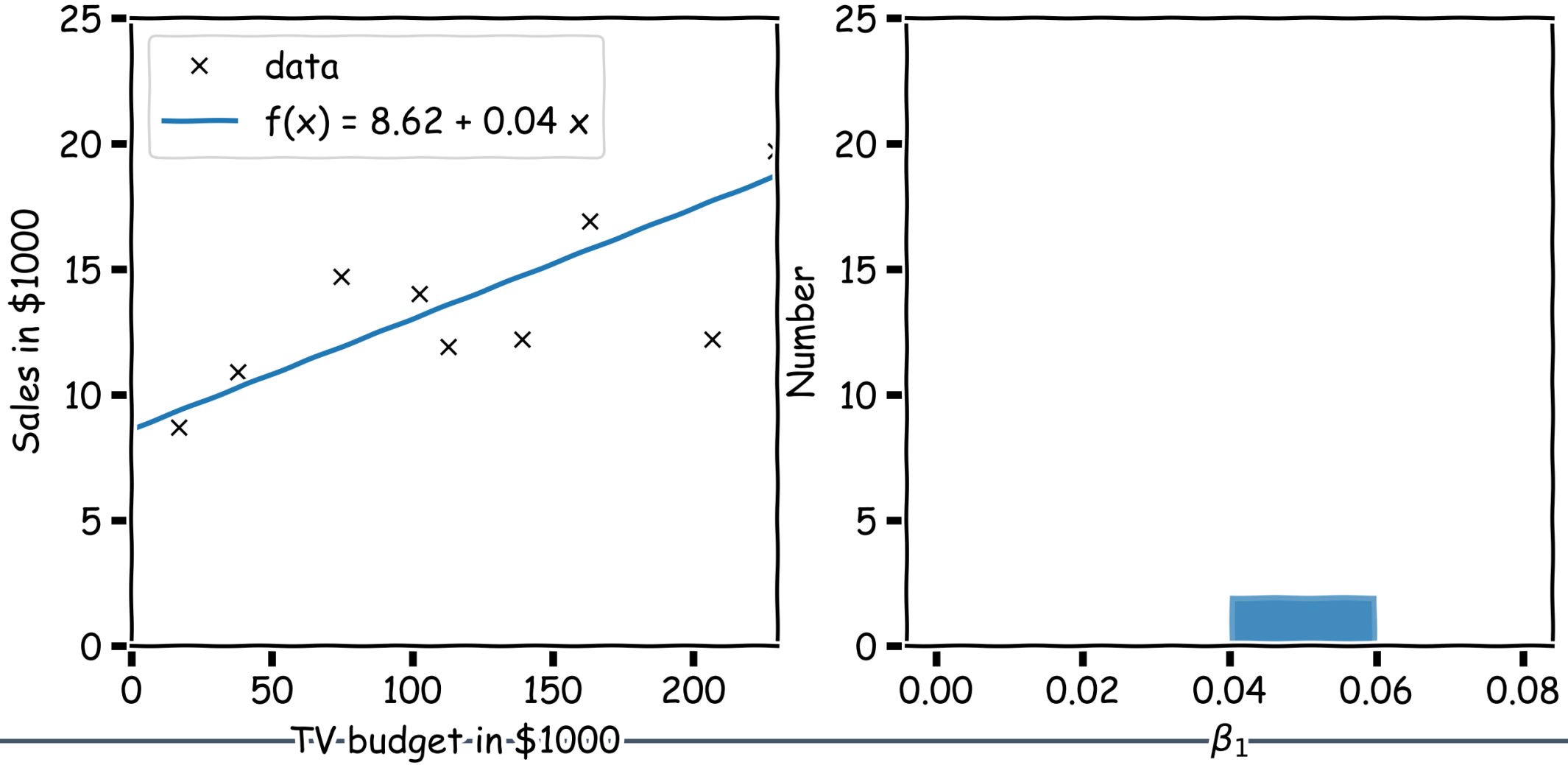


Confidence intervals for the coefficients

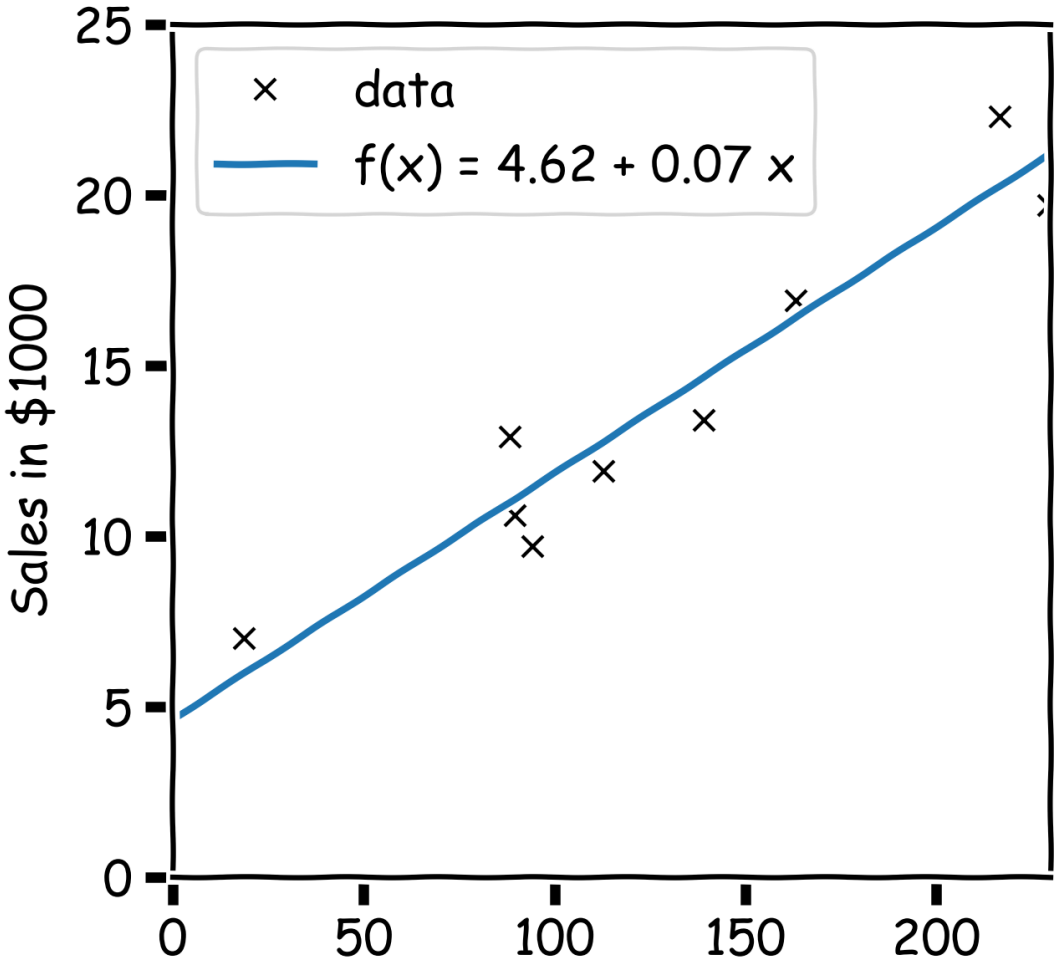
Let's start by taking one bootstrapped sample of our data and fitting a model

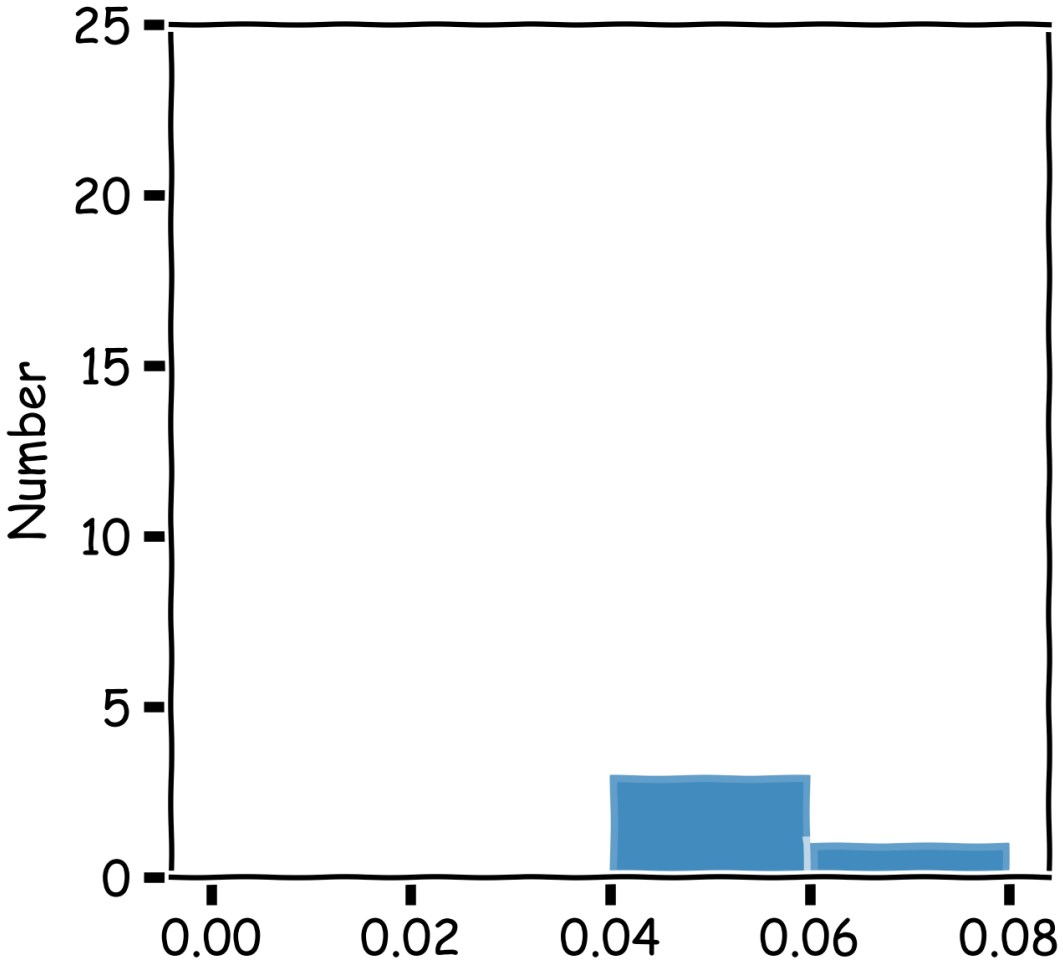


Now let's take another sample

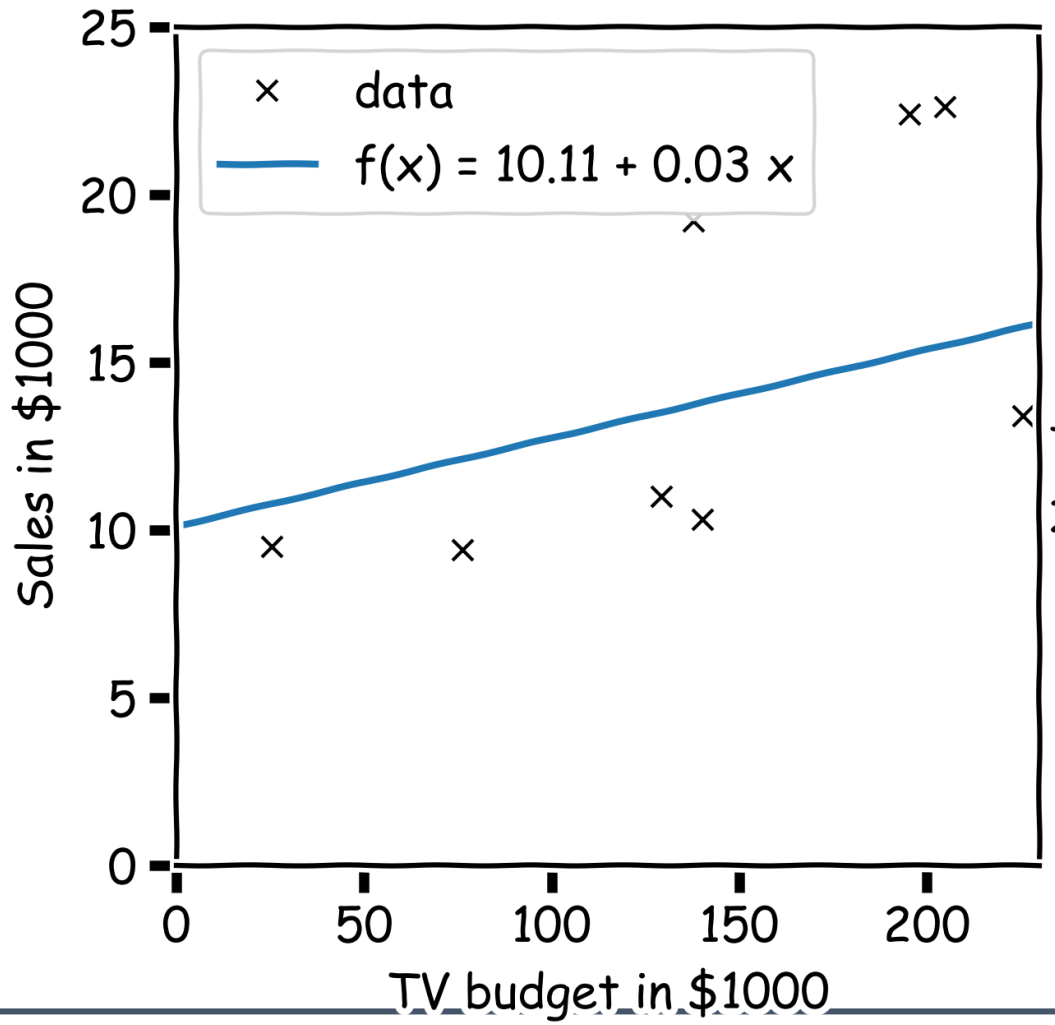


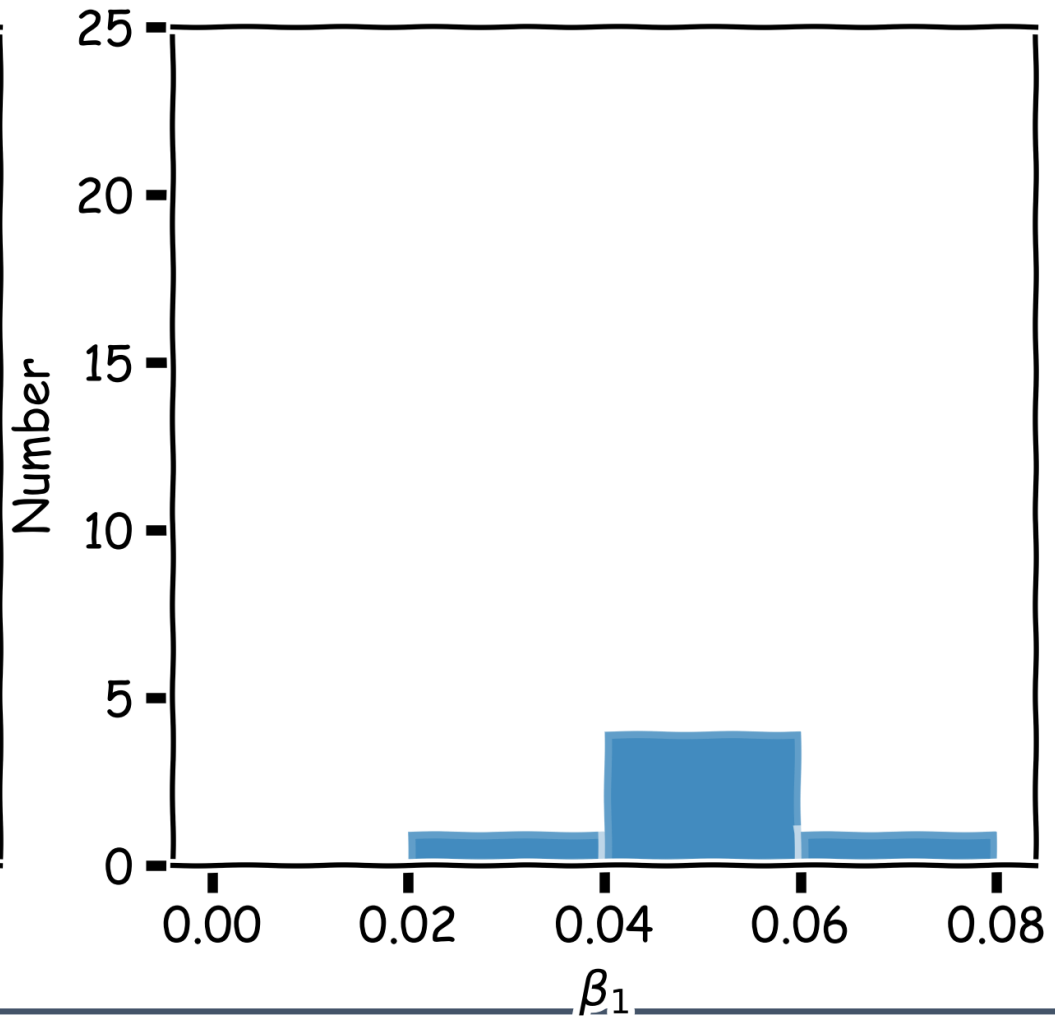
And another


data
f(x) = 4.62 + 0.07 x
Sales in $1000
TV-budget-in-$1000
0
5
10
15
20
25
0
50
100
150
200

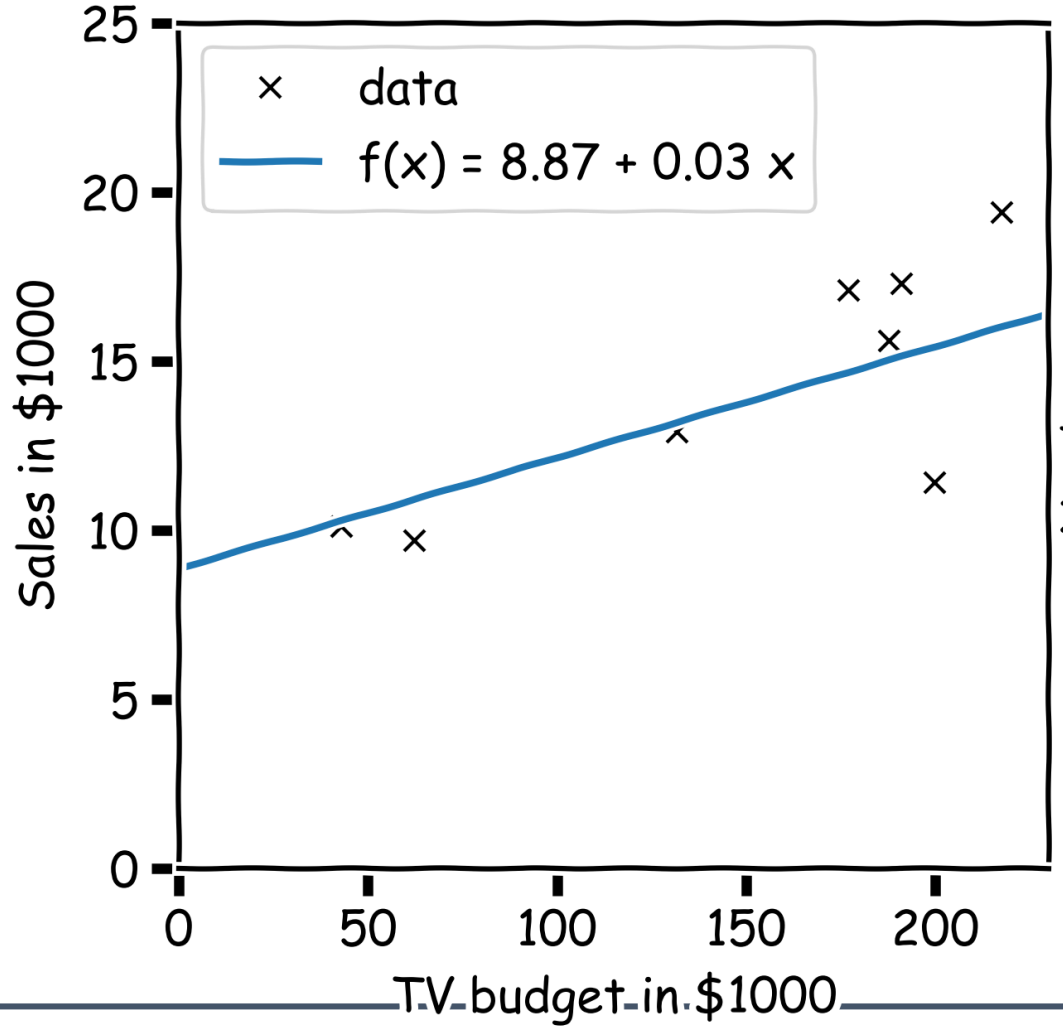

Number
β_1
0
5
10
15
20
25
0.00
0.02
0.04
0.06
0.08

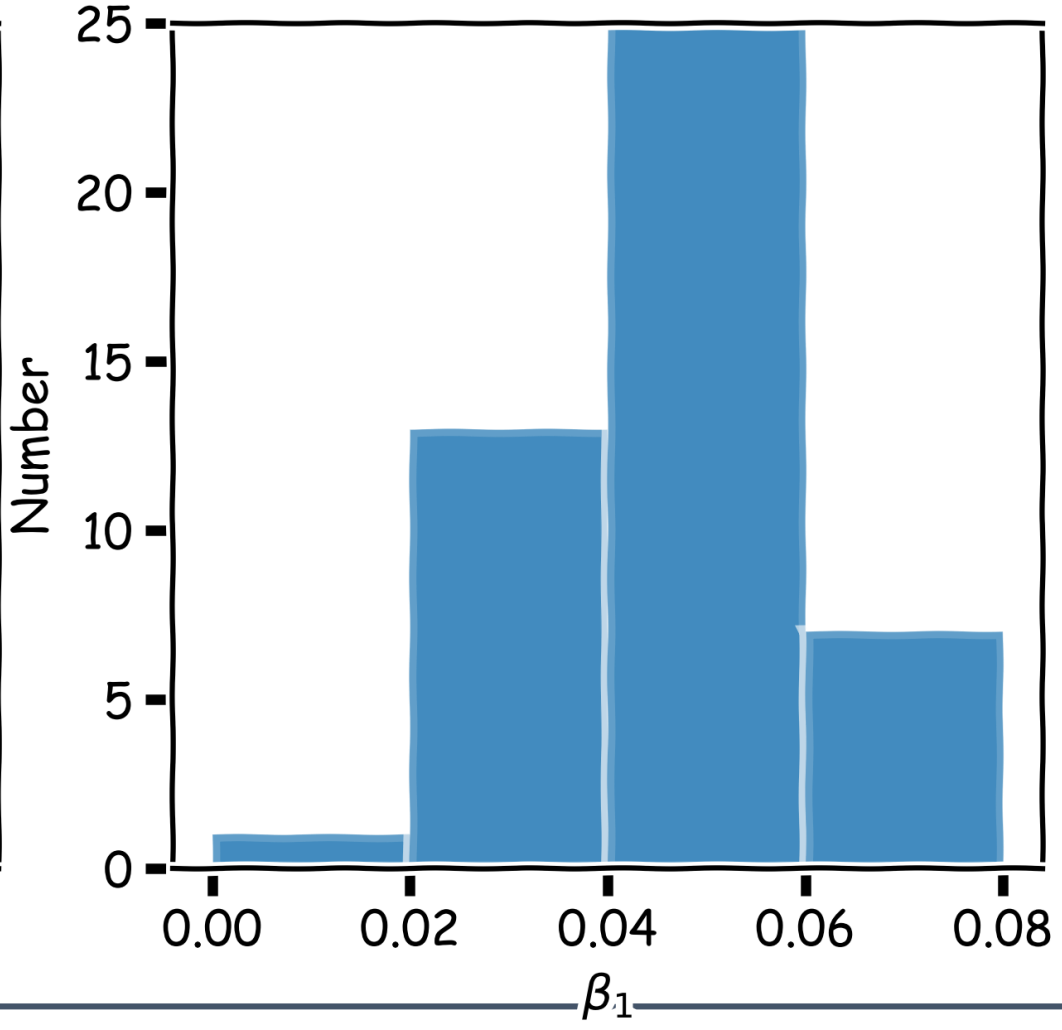
And another


data
f(x) = 10.11 + 0.03 x
Sales in $1000
TV budget in $1000
0
5
10
15
20
25
0
50
100
150
200


Number
β_1
0
5
10
15
20
25
0.00
0.02
0.04
0.06
0.08

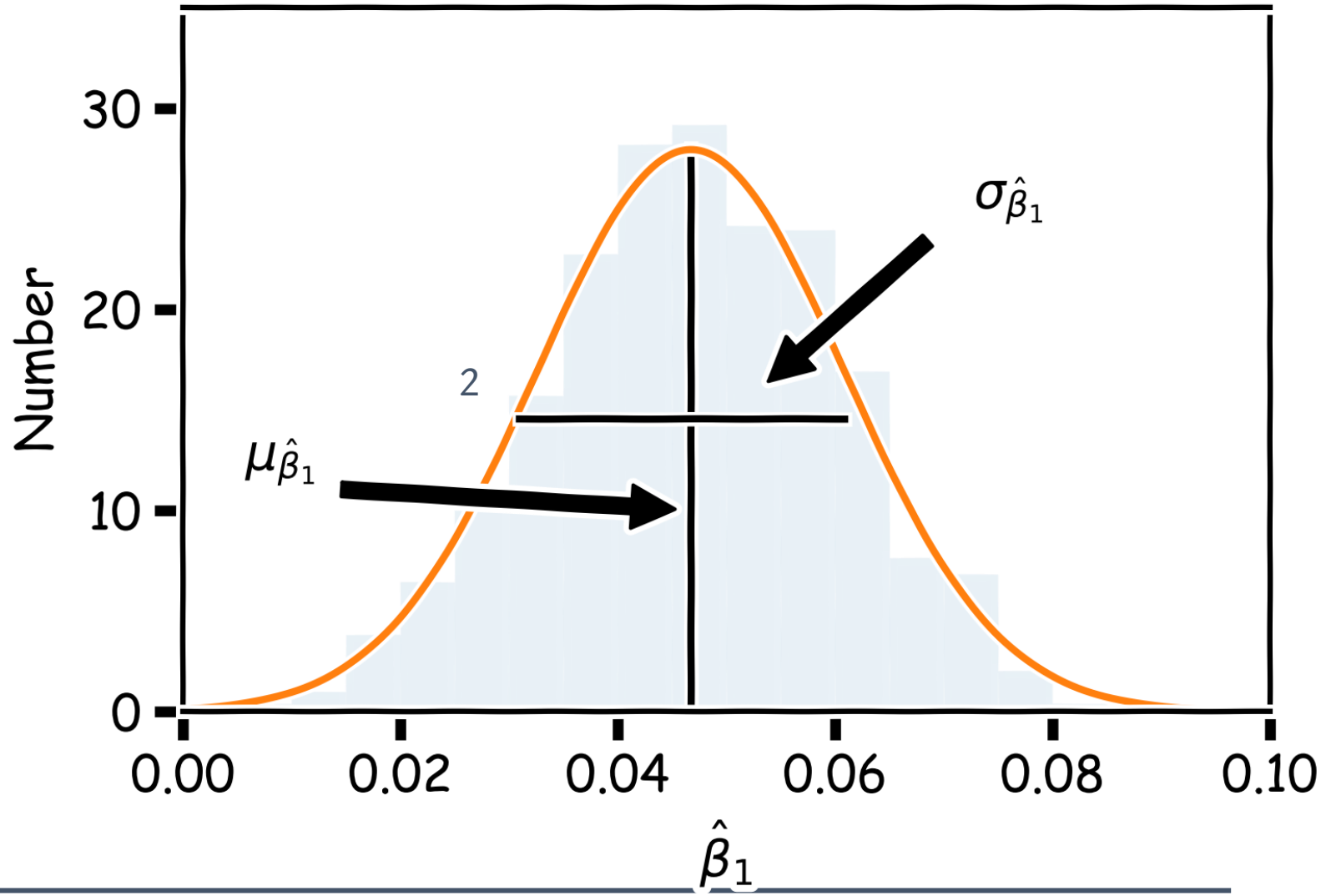
Repeat this 100 times





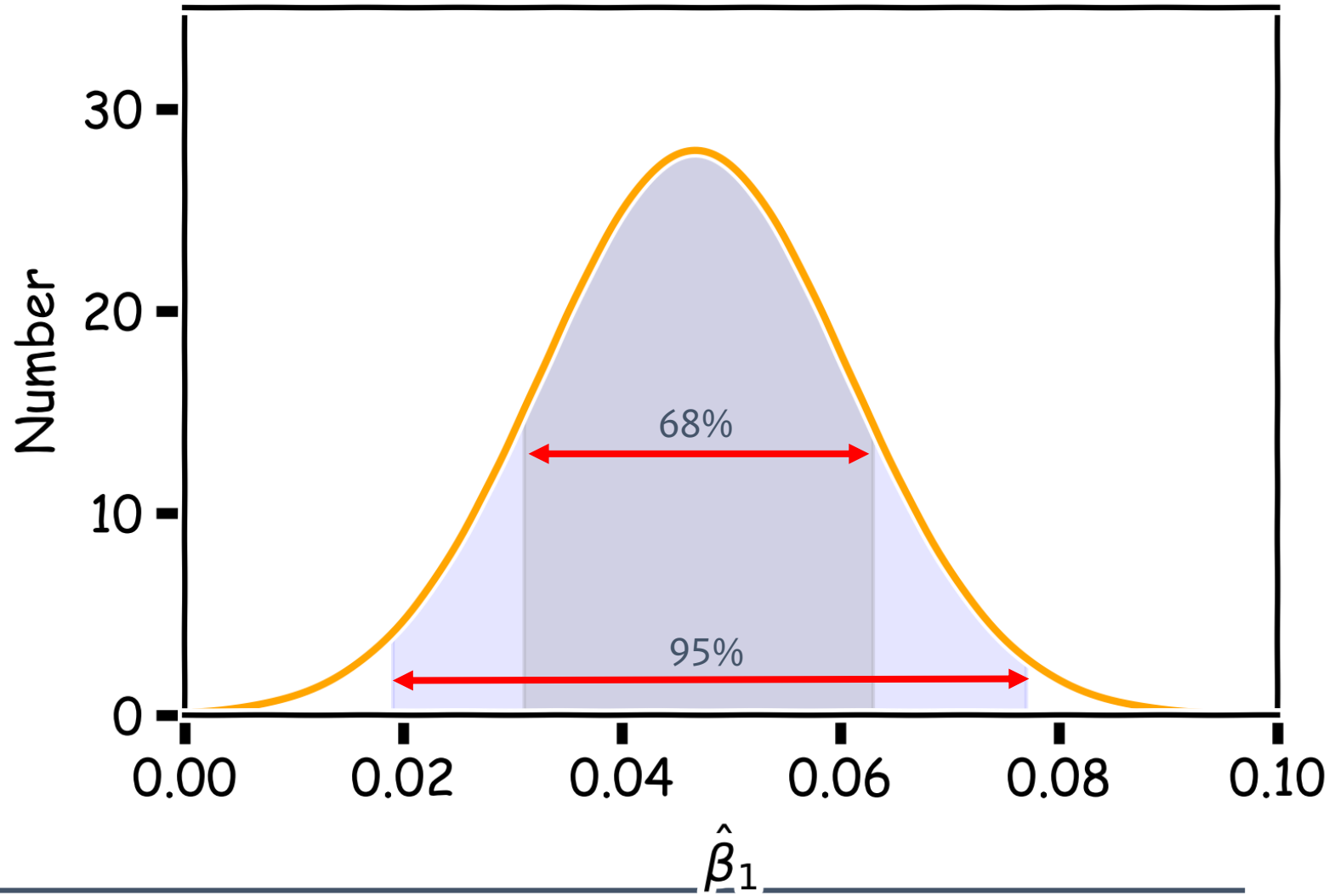
Confidence intervals for the coefficients

- We can now estimate the mean and standard deviation of all the estimates $\hat{\beta}_1$.
- The deviation of $\hat{\beta}_0$ and $\hat{\beta}_1$ are also called their ***standard errors***:
- $SE(\hat{\beta}_0)$
- $SE(\hat{\beta}_1)$



Confidence intervals for the coefficients

Finally we can calculate the confidence intervals, which are the ranges of values such that the **true** value of β_1 is contained in this interval with *n* percent probability.

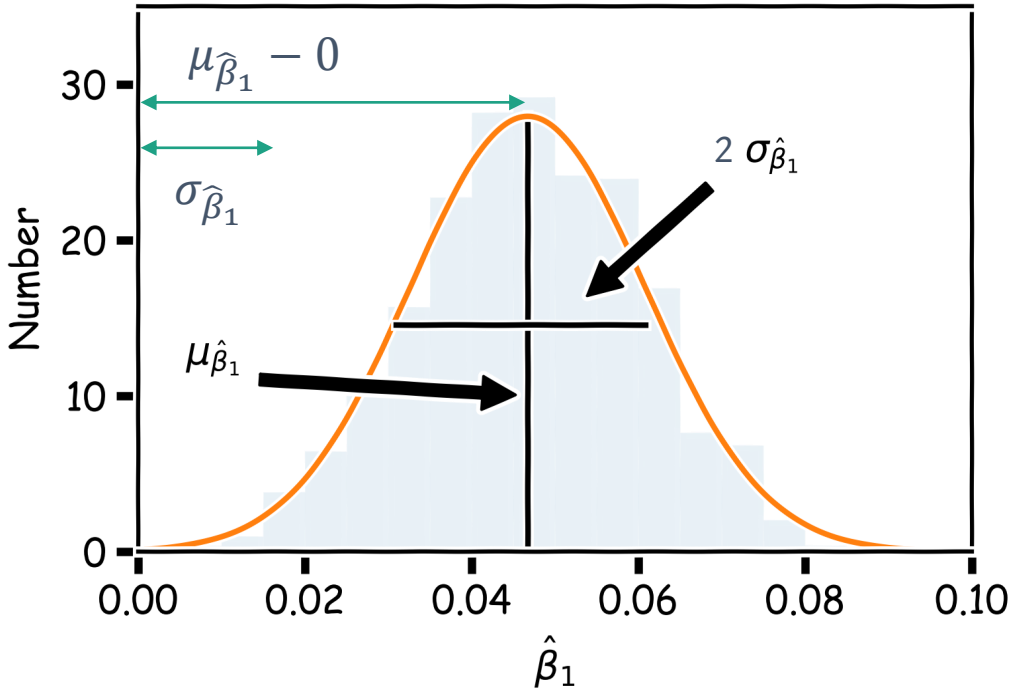


Importance of predictors

- The importance of predictors is evaluated using a T-Test:

$$t = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

- Which measures the distance from zero in units of standard deviation.



- We evaluate how often a particular value of *t* can occur by accident. We expect that *t* will have a *t-distribution with n-2 degrees* of freedom.

- To compute the probability of observing any value equal to $|t|$ or larger, assuming $\hat{\beta}_1 = 0$ is easy. We call this probability the **p-value**.
  - ***a small p-value (<0.05) indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance. We call such results "statistically significant" at the relevant level***

Coefficient Plot

Predicting Gas Mileage

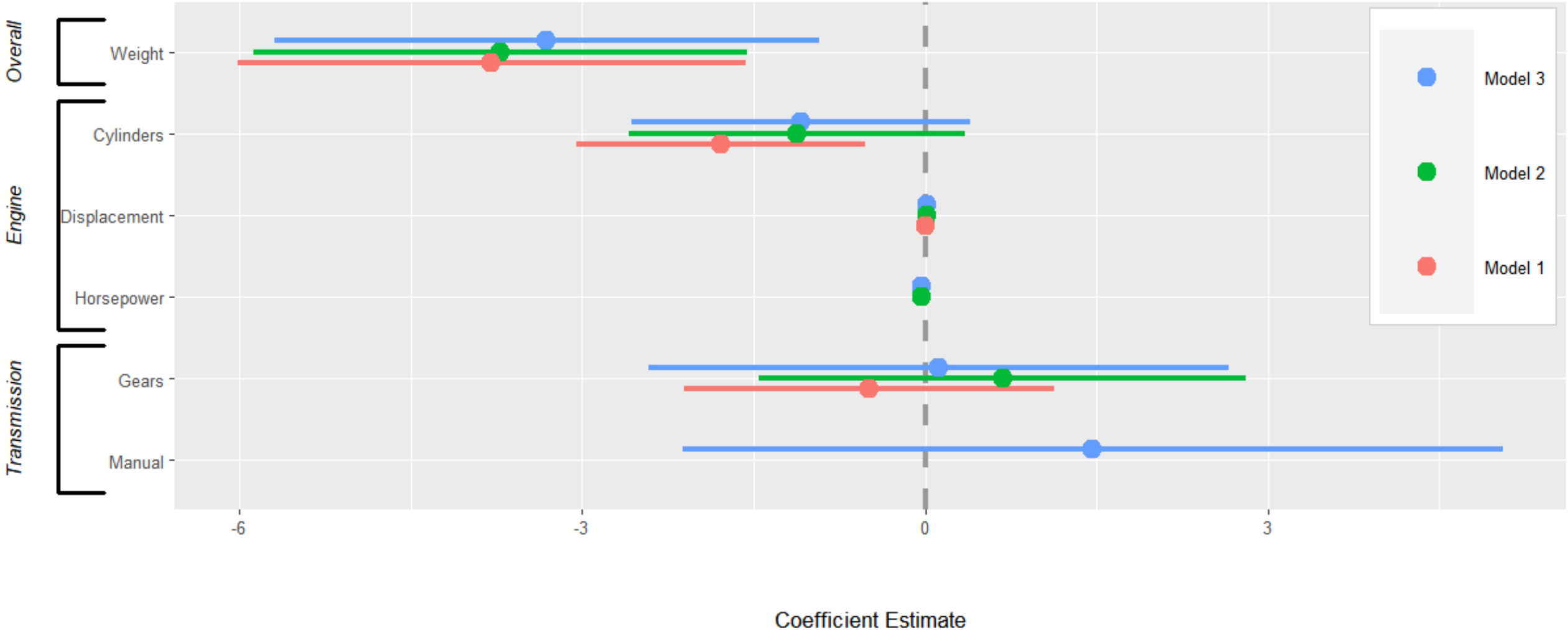


Table 1: Effect of Demographic Variables on Hourly Wages

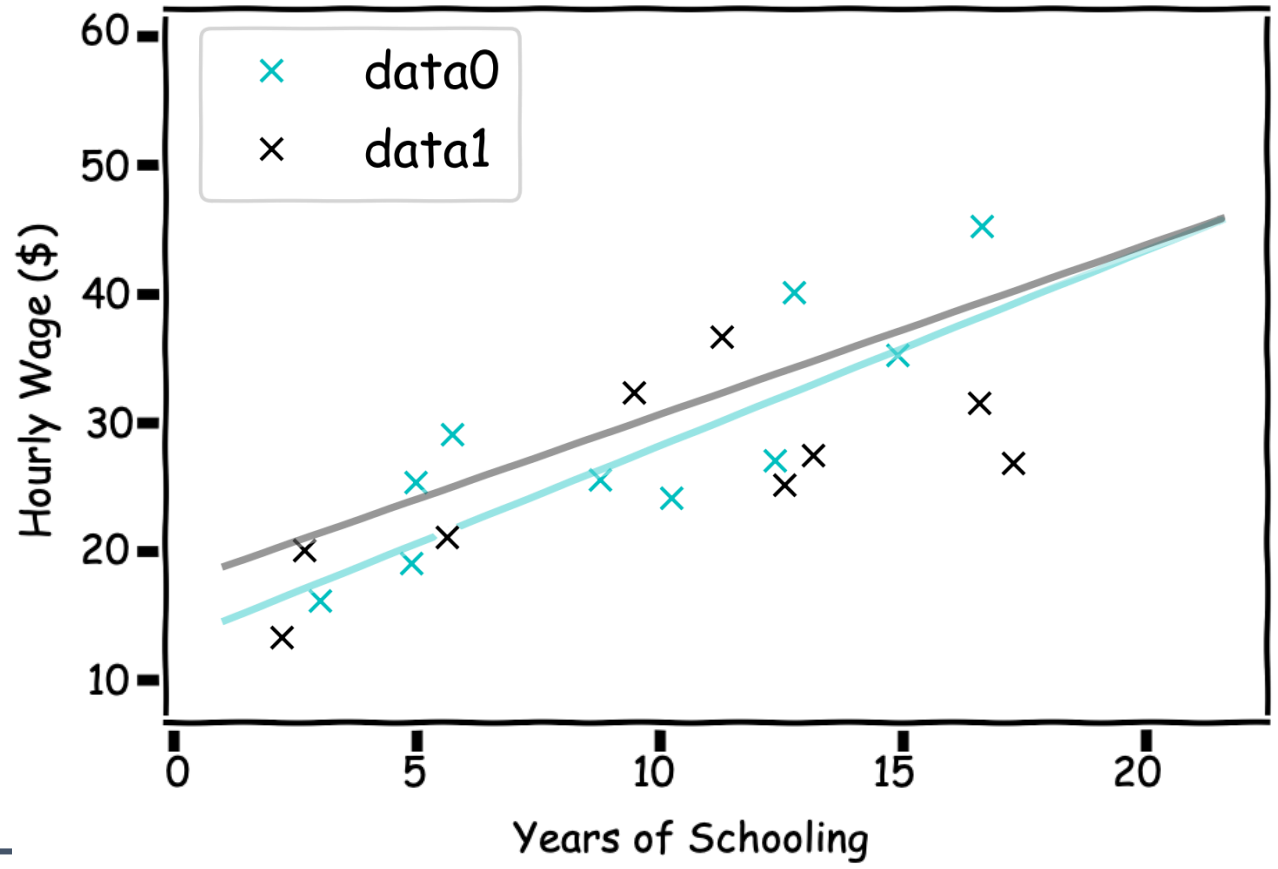
	All	Production	Farmers	Bankers	Doctors & Lawyers
Schooling	1.972***	0.757***	0.435***	1.905***	2.376*
	(0.019)	(0.056)	(0.134)	(0.176)	(1.395)
Age	0.156***	0.109***	0.087**	0.203***	0.635***
	(0.005)	(0.013)	(0.044)	(0.031)	(0.075)
Female	-4.881***	-4.893***	-1.480	-4.849***	1.443
	(0.096)	(0.306)	(1.026)	(0.670)	(1.536)
Not Union Member	-0.669***	-0.507	-1.663	-0.064	-3.401
	(0.139)	(0.394)	(1.304)	(0.875)	(2.147)
Union Member	1.813***	4.159***	5.750	0.640	4.840
	(0.320)	(0.864)	(8.288)	(3.022)	(4.721)
Black	-2.498***	-2.702***	-3.130	-1.731	-0.770
	(0.160)	(0.467)	(1.971)	(1.056)	(3.024)
Hispanic	-1.209***	-2.426***	-2.379**	-1.844*	-0.872
	(0.139)	(0.360)	(1.113)	(1.089)	(3.289)
Other	0.046	-2.273***	-2.633	-1.342	-1.127
	(0.178)	(0.510)	(2.376)	(1.036)	(2.042)
Intercept	-9.772***	5.942***	6.363**	-8.088**	-30.641
	(0.348)	(0.978)	(2.653)	(3.244)	(25.584)
R-squared	0.236	0.192	0.102	0.161	0.163
N	52366	3452	351	1267	452

Standard errors in parentheses. $*p < .1, **p < .05, ***p < .01$

2.4 Confidence Interval on Predicted Values

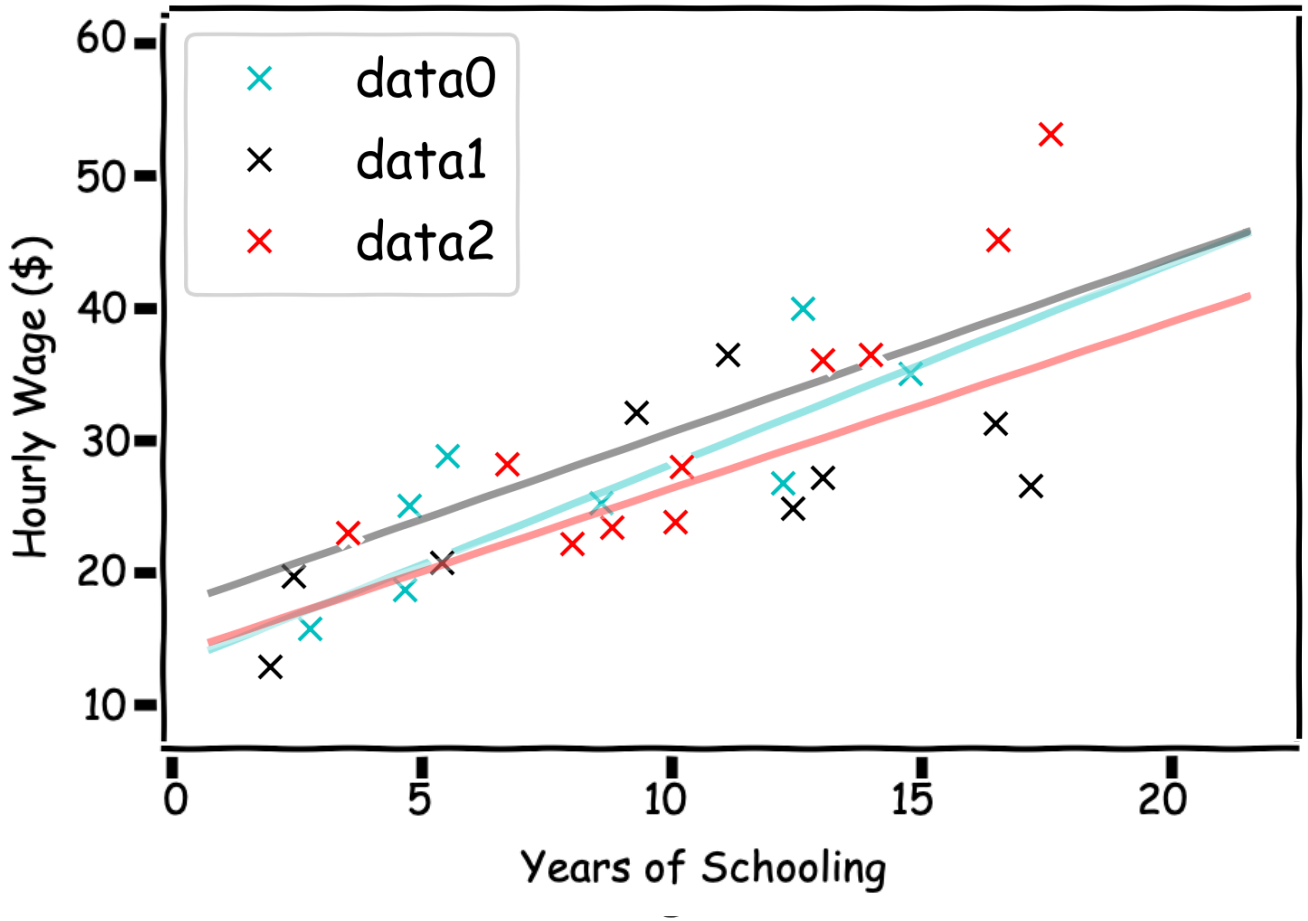
How well do we know $\hat{f}$?

Here we show two difference set of models given the fitted coefficients.



How well do we know $\hat{f}$?

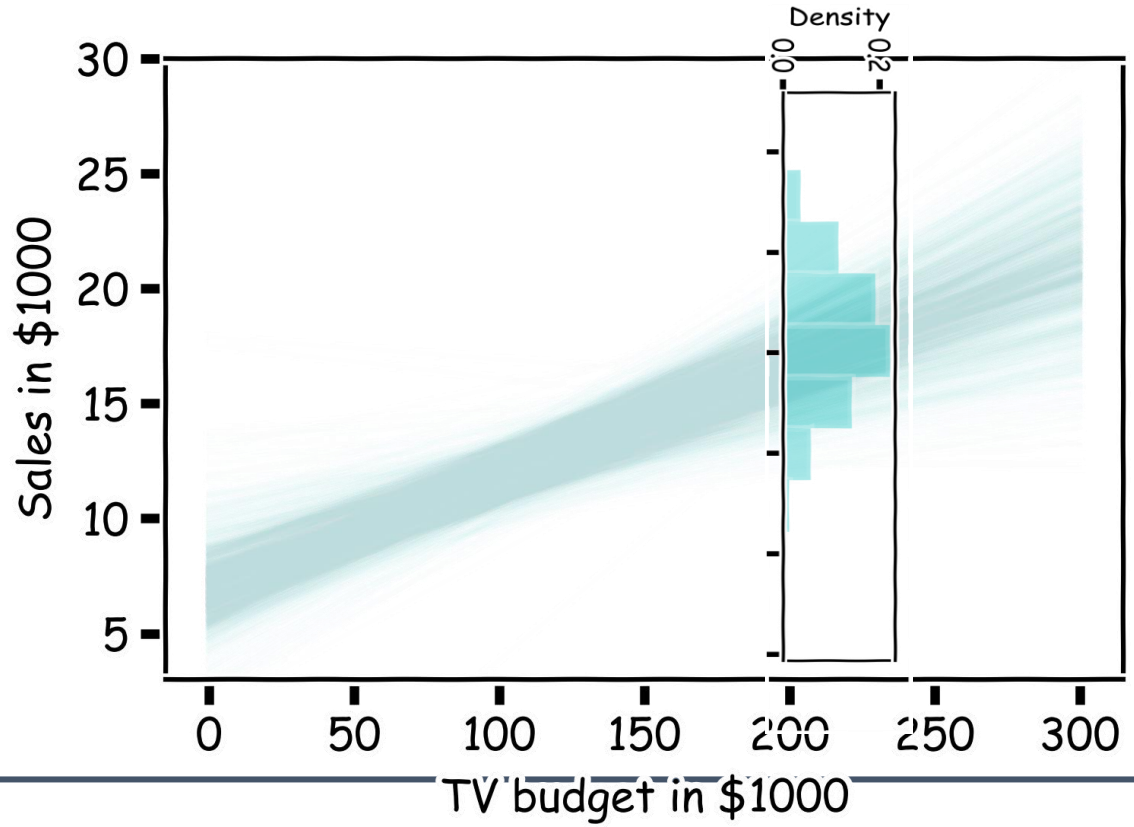
There is one such regression line for every bootstrapped sample.



How well do we know $\hat{f}$?

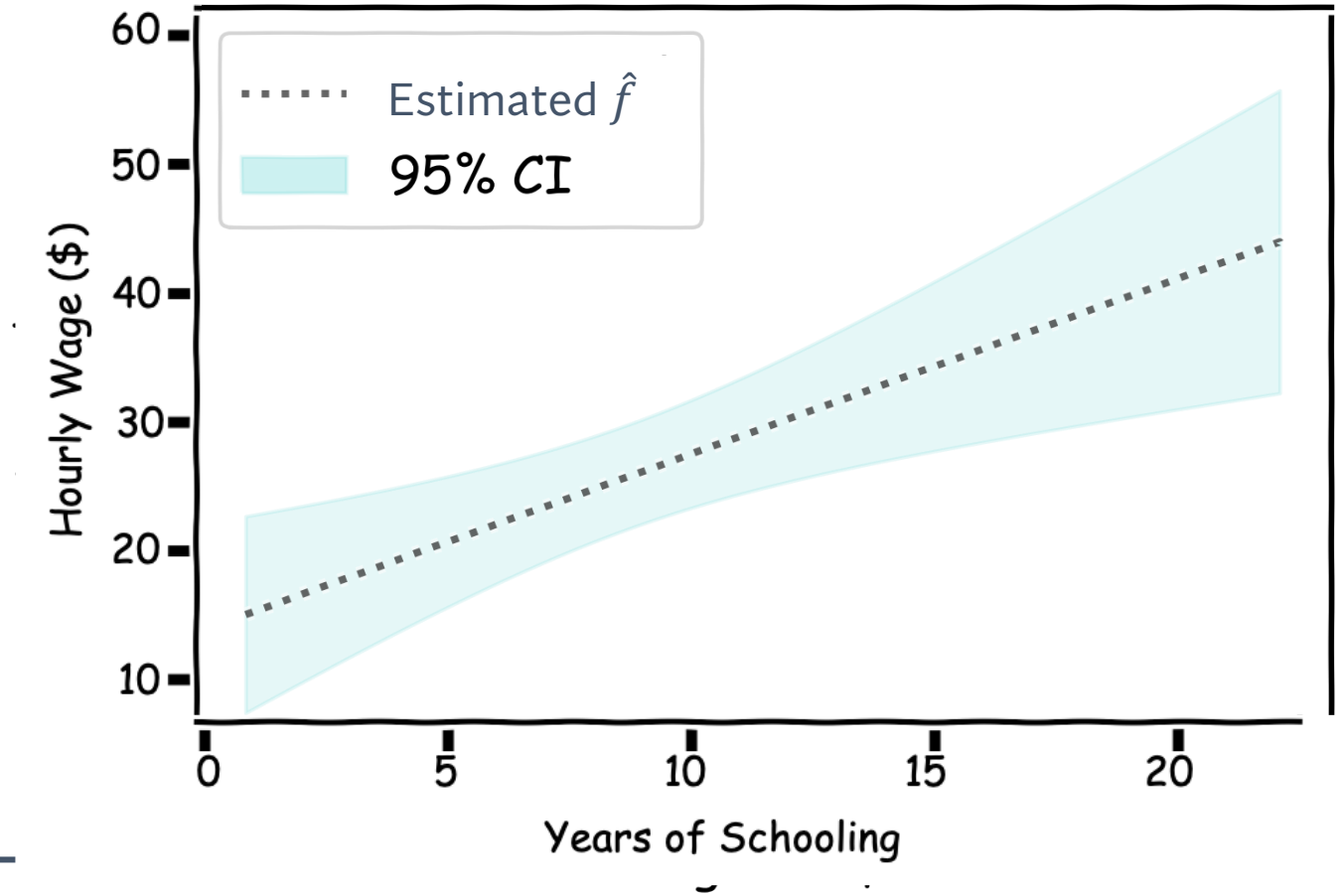
Below we show all regression lines for a thousand of such bootstrapped samples.

For a given x, we examine the distribution of $\hat{f}$, and determine the mean and standard deviation.



How well do we know $\hat{f}$?

For every x, we calculate the mean of the models, $\hat{f}$ (shown with dotted line) and the 95% CI of those models (shaded area).



3. Assumptions

Linear Regression Assumptions

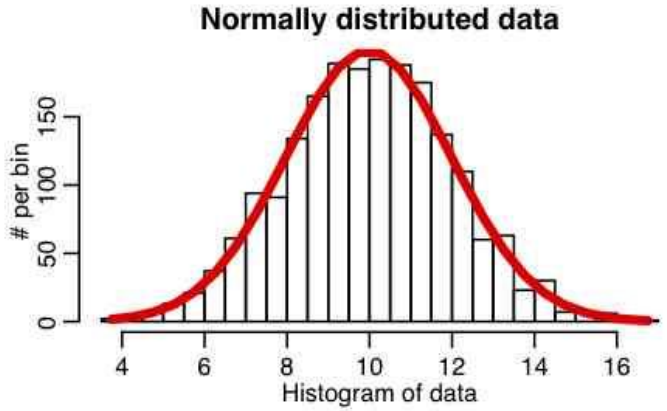
1. Normality
2. Homoscedasticity
3. Independence

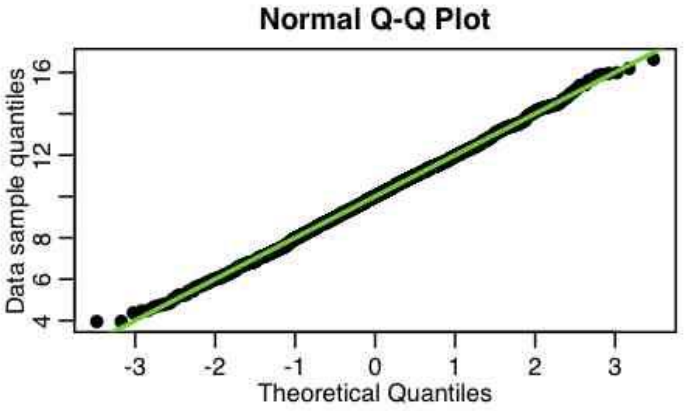
1. Normality

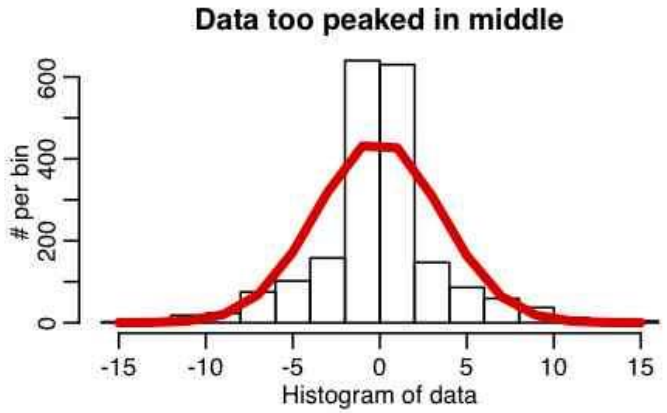
The residuals from the regression should be normally distributed

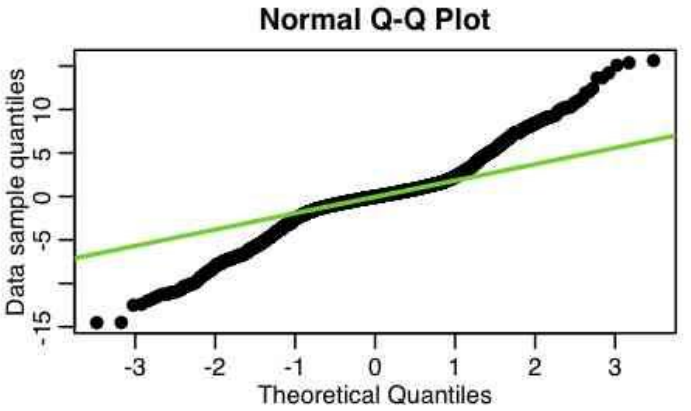
If they are not, our model does not perform consistently across the full range of our data

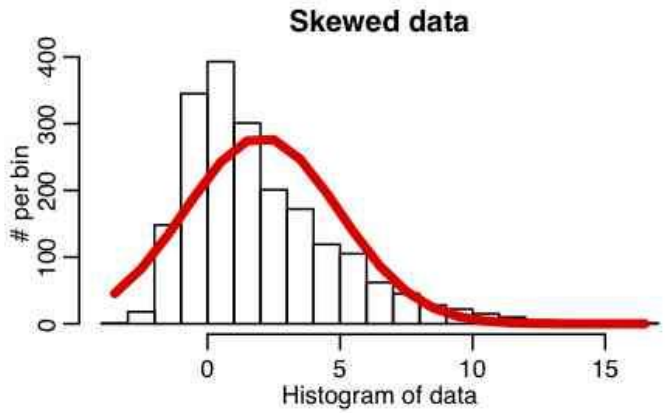
The relationship between the variables may not be linear

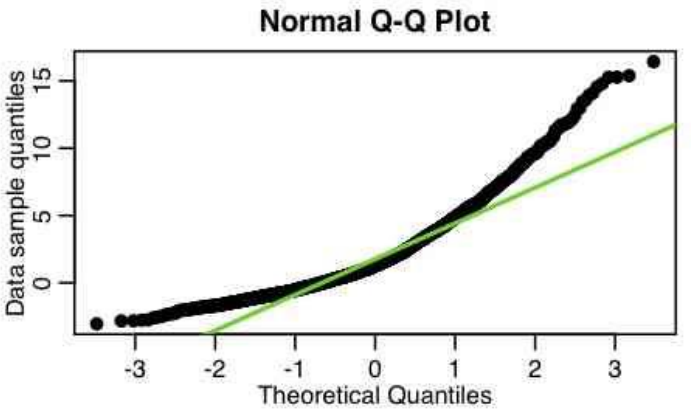










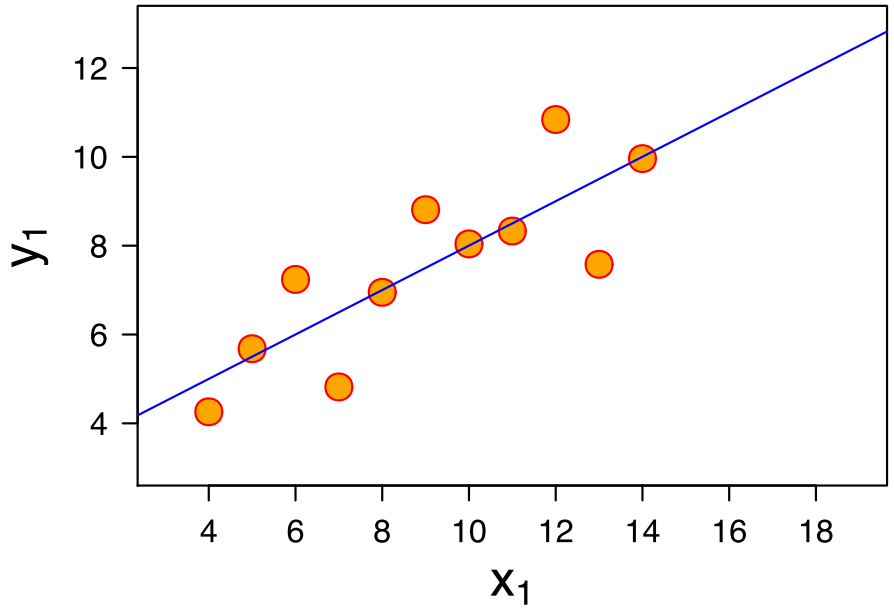


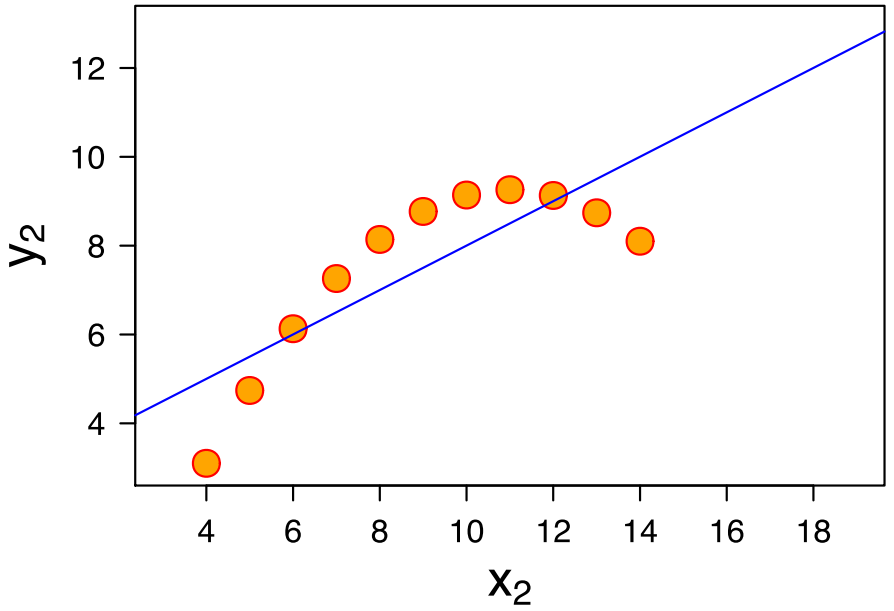
Anscombe's Quartet

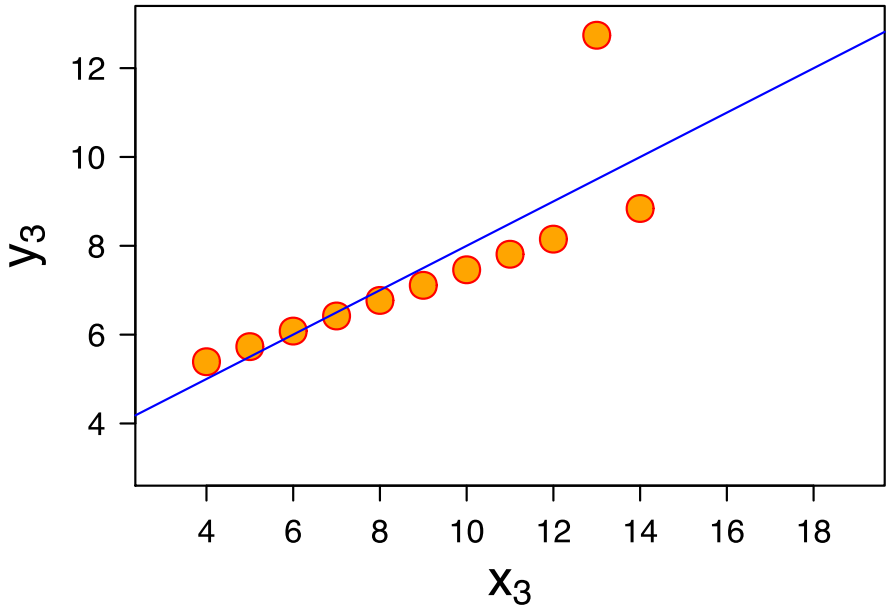
$\bar{x}$	9
S_x^2	11
$\bar{y}$	7.50
S_y^2	4.125

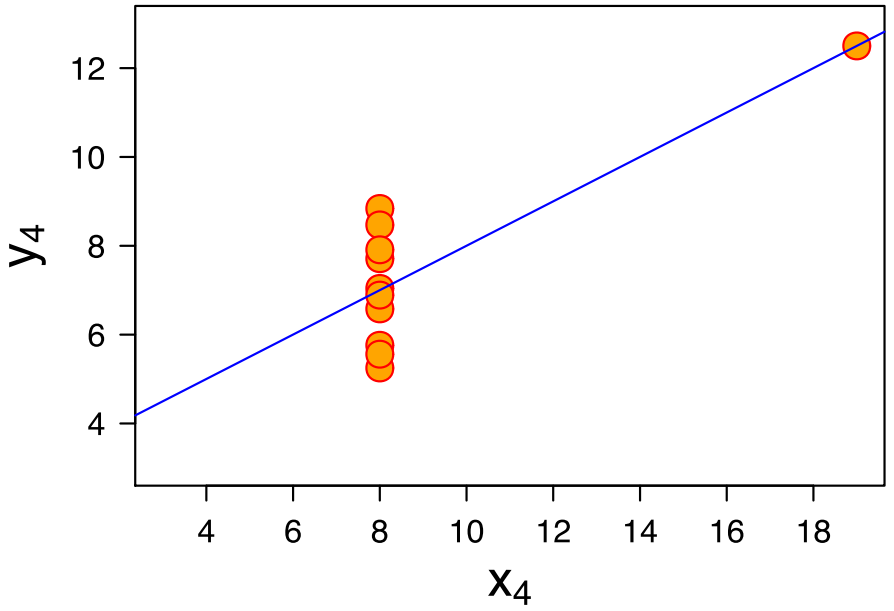
$y = 3.00 + 0.500x$

$R^2 = 0.67$

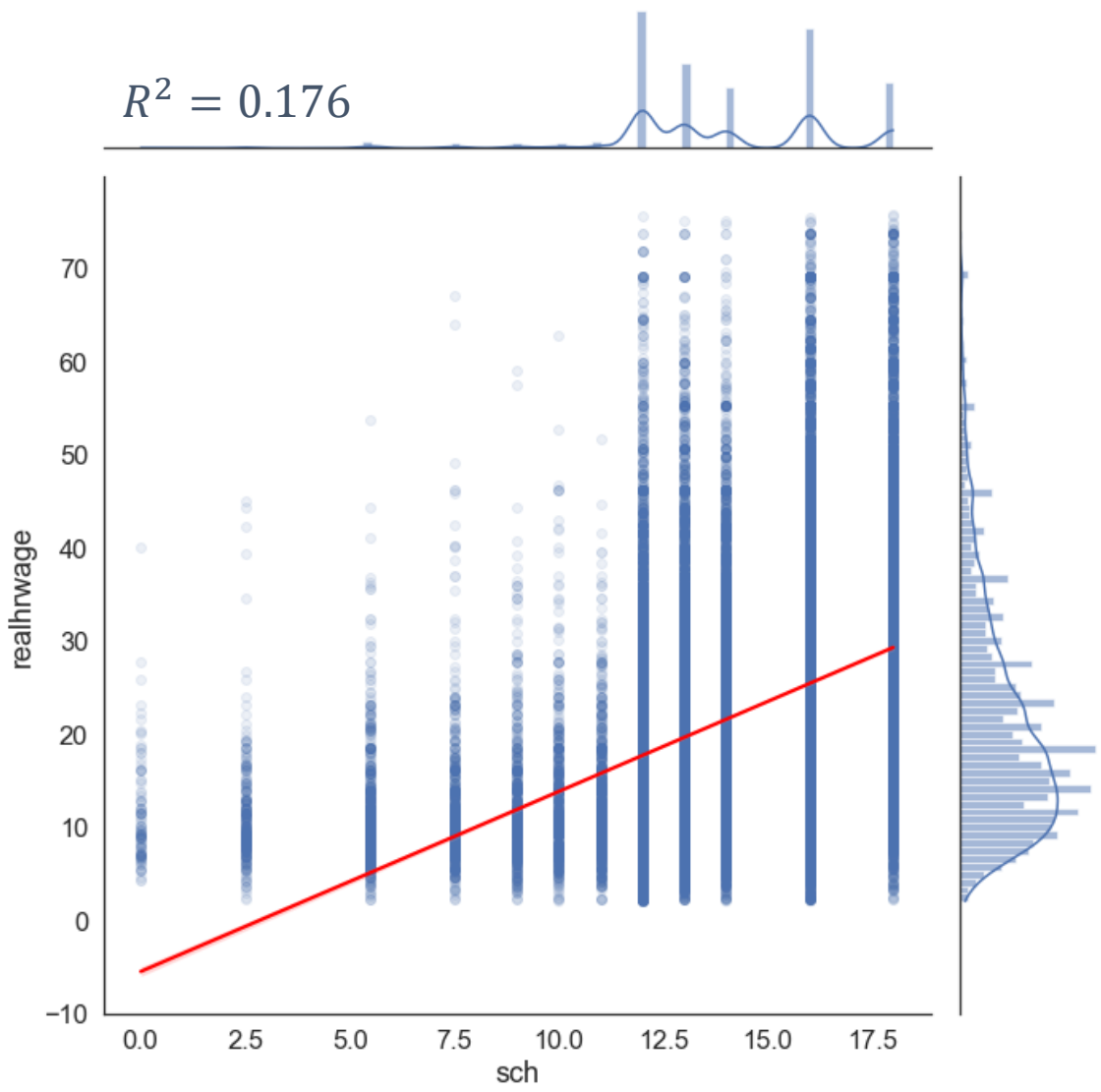


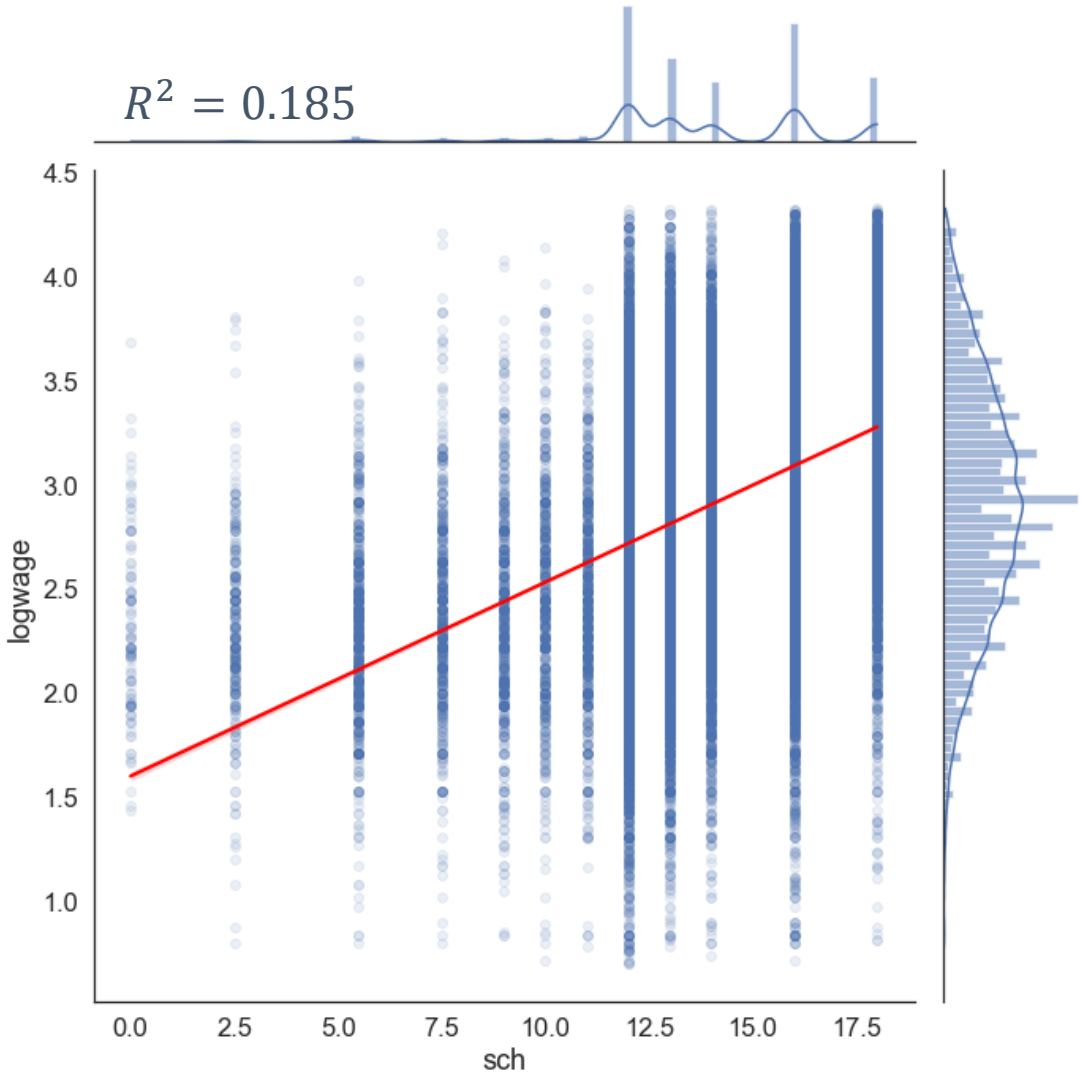






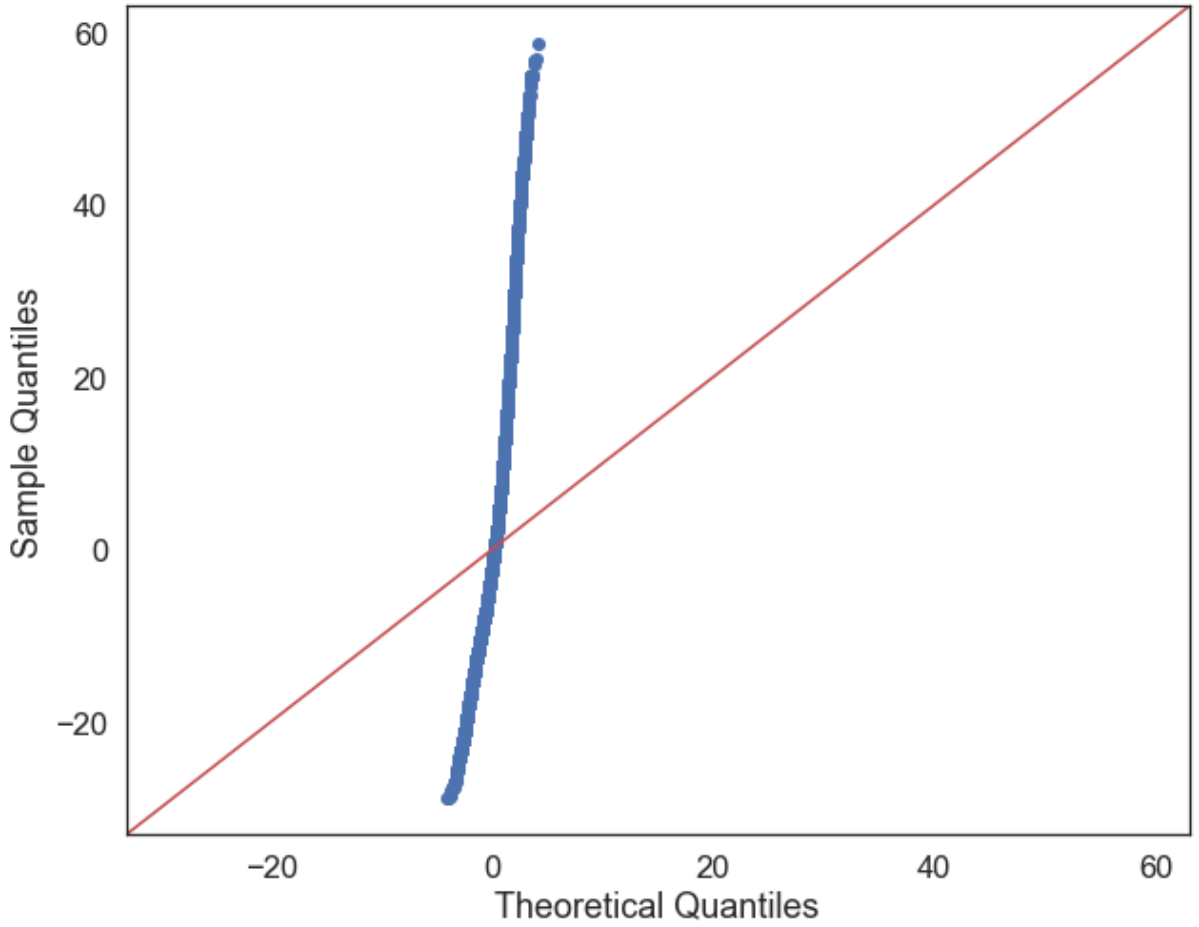
Non-Linear relationships



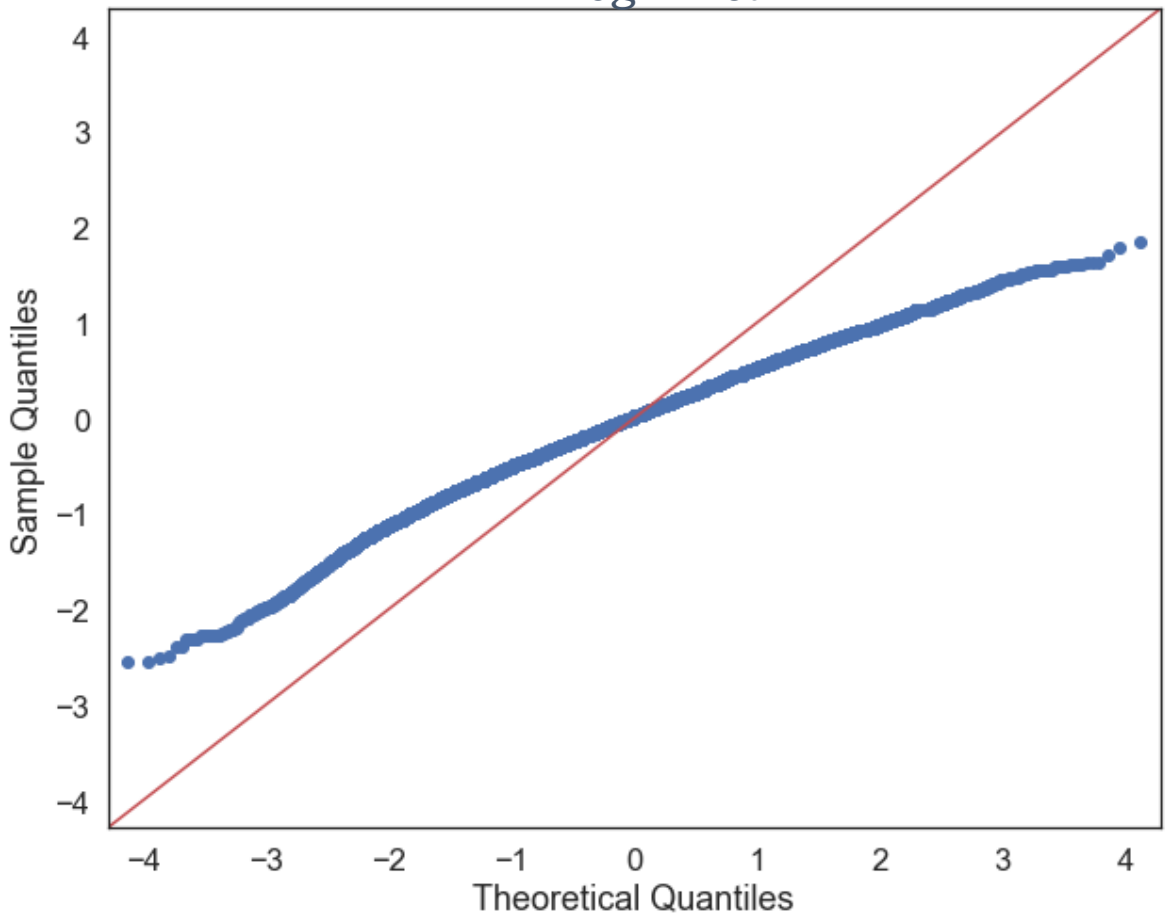


Non-Linear relationships

Linear-Linear

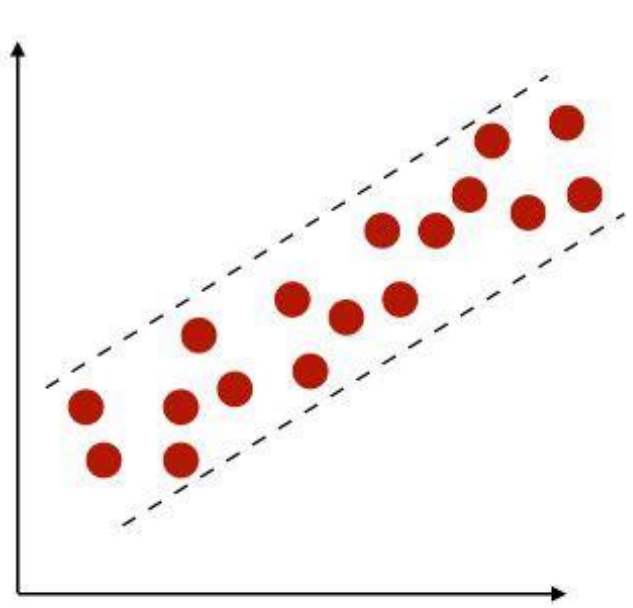


Log-Linear

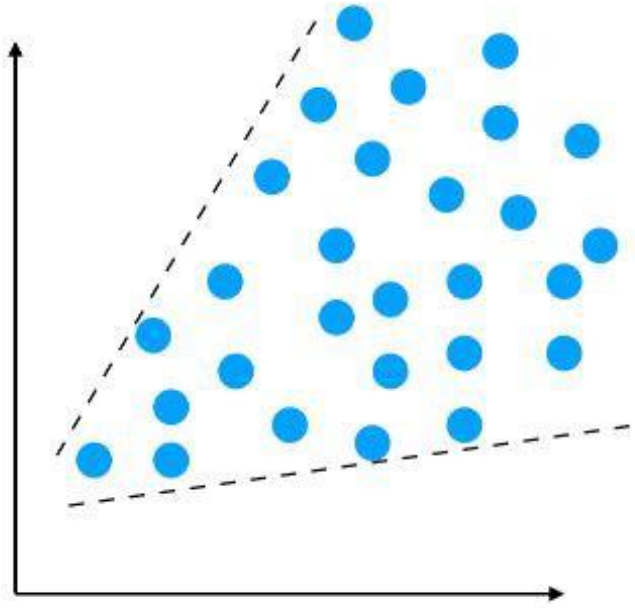


2. Homoscedasticity

The variance of residuals is the same for any value of X

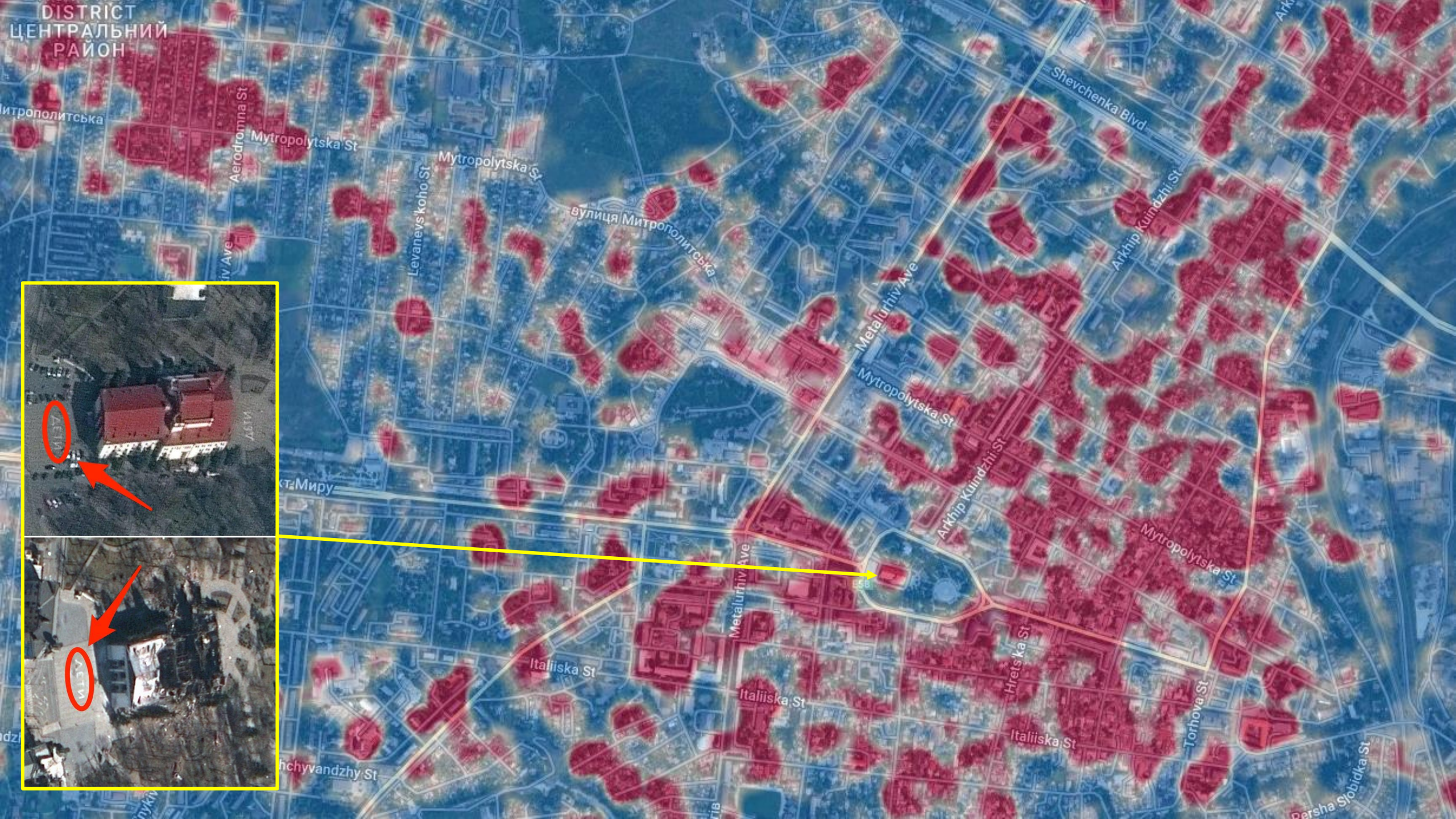


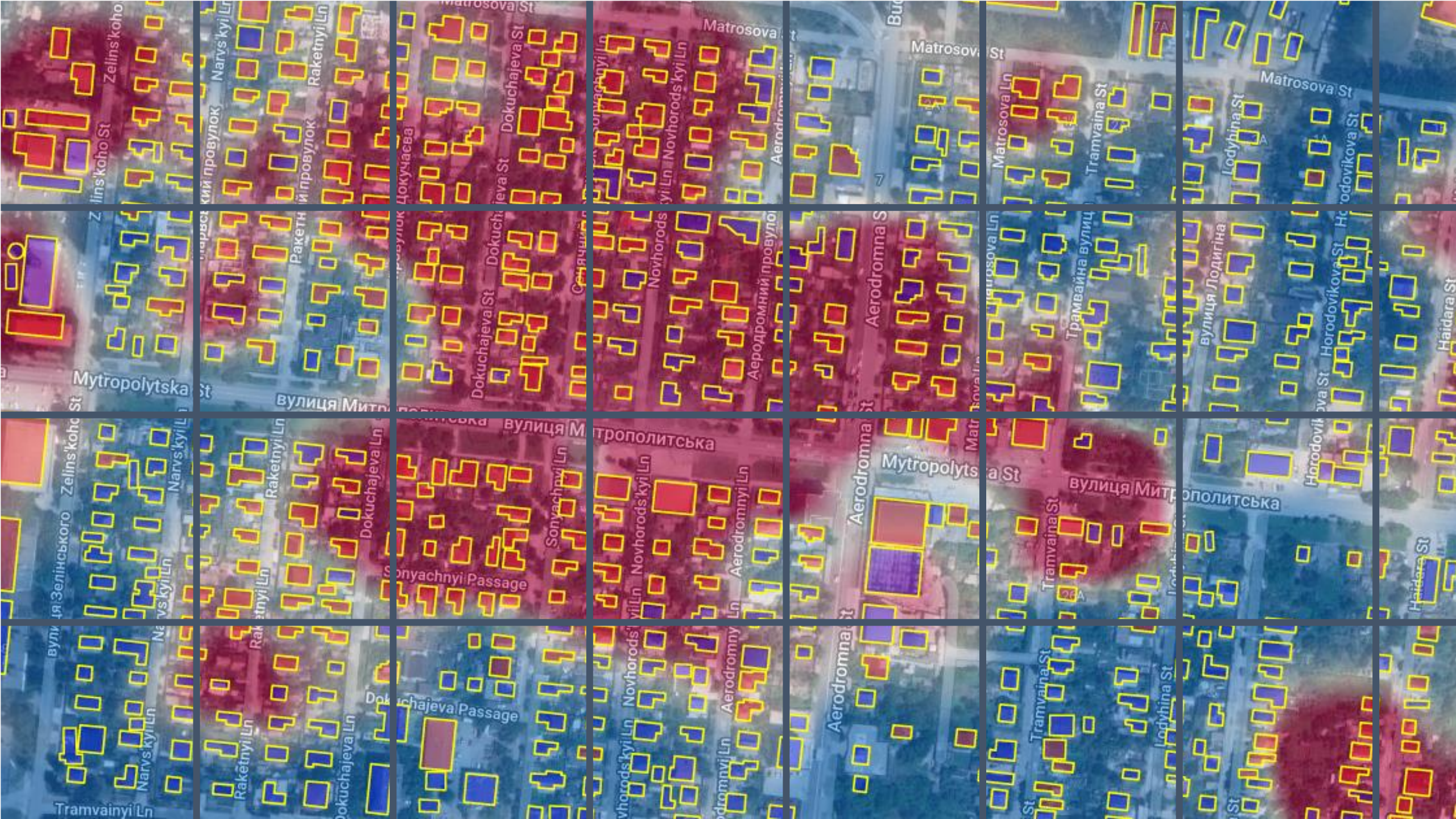
Homoscedasticity



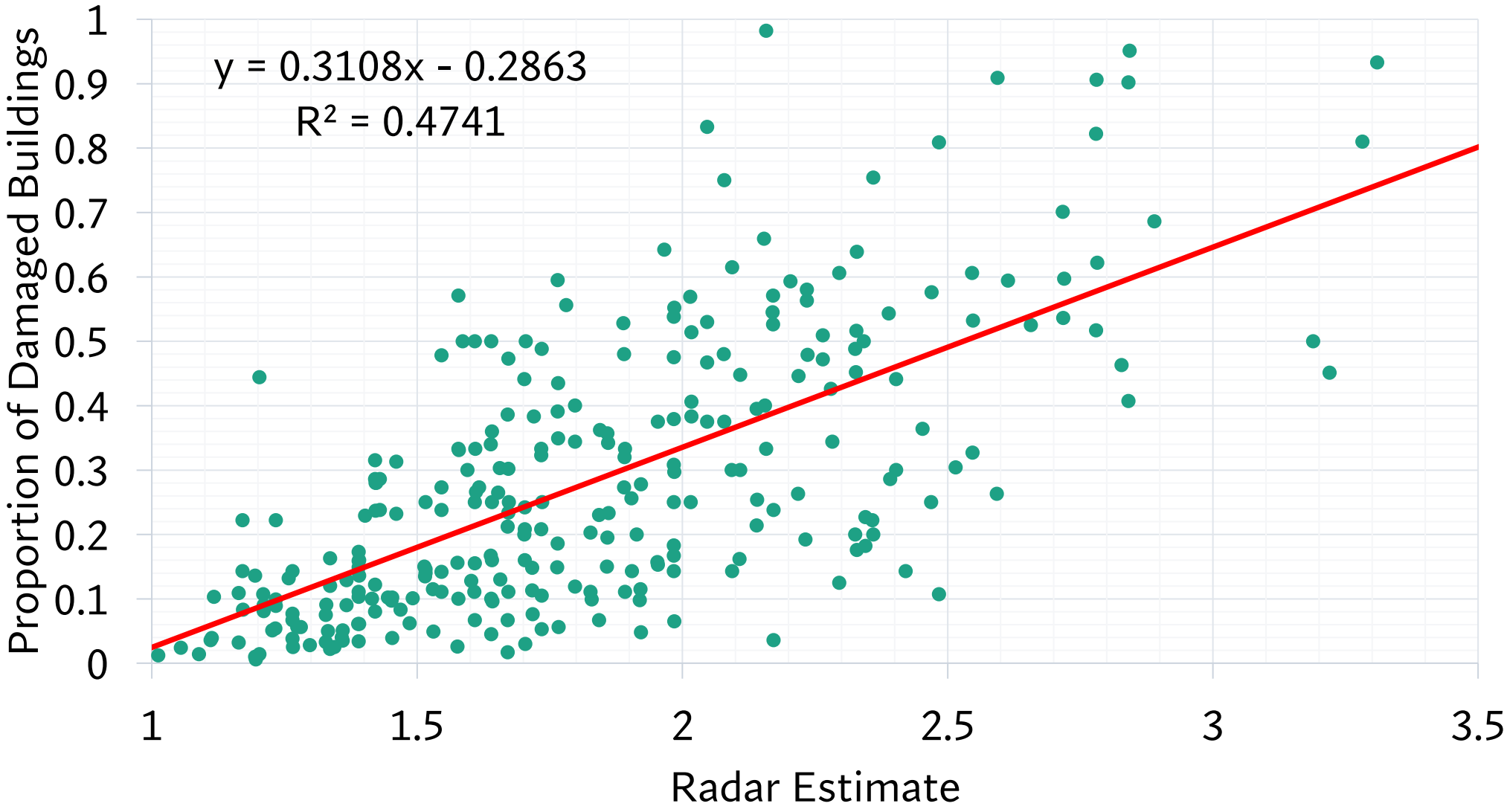
Heteroscedasticity


DISTRICT
ЦЕНТРАЛЬНИЙ
РАЙОН
Митрополитська
Aerodromna St
Mytropolytska St
Levanevs'koho St
Mytropolytska St
вулиця Митрополитська
Budivel'nykiv Ave
Shevchenka Blvd
Arkhip Kuindzhi St
Metalurhiv Ave
Mytropolytska St
Arkhip Kuindzhi St
проспект Миру
Budivel'nykiv Ave
Mytropolytska St
Italiiska St
Metalurhiv Ave
Italiiska St
Hrets'ka St
Italiiska St
Italiiska St
Torhova St
Bakhchyvandzhy St
Persha Slobidka St


DISTRICT
ЦЕНТРАЛЬНИЙ
РАЙОН
Митрополитська
Aerodromna St
Mytropolytska St
Levanevs'koho St
Mytropolytska St
вулиця Митрополитська
Shevchenka Blvd
Arkhip Kuindzhi St
Metalurhiv Ave
Mytropolytska St
Arkhip Kuindzhi St
Mytropolytska St
Metalurhiv Ave
Italiiska St
Italiiska St
Italiiska St
Hrets'ka St
Torhova St
Persha Slobidka St
Миру
hchyvandzhy St


Matrosova St
Dokuchajeva St
Raketnyi Ln
Narvs'kyi Ln
Zelins'koho St
Novhorods'kyi Ln
Aerodromna St
Tramvaina St
Lodyhina St
Horodovikova St
Mytropolytska St
Вулиця Митрополитська
Вулиця Зелінського
Sonyachnyi Passage
Aerodromnyi Ln
Tramvainyi Ln

Heteroscedasticity



3. Independence

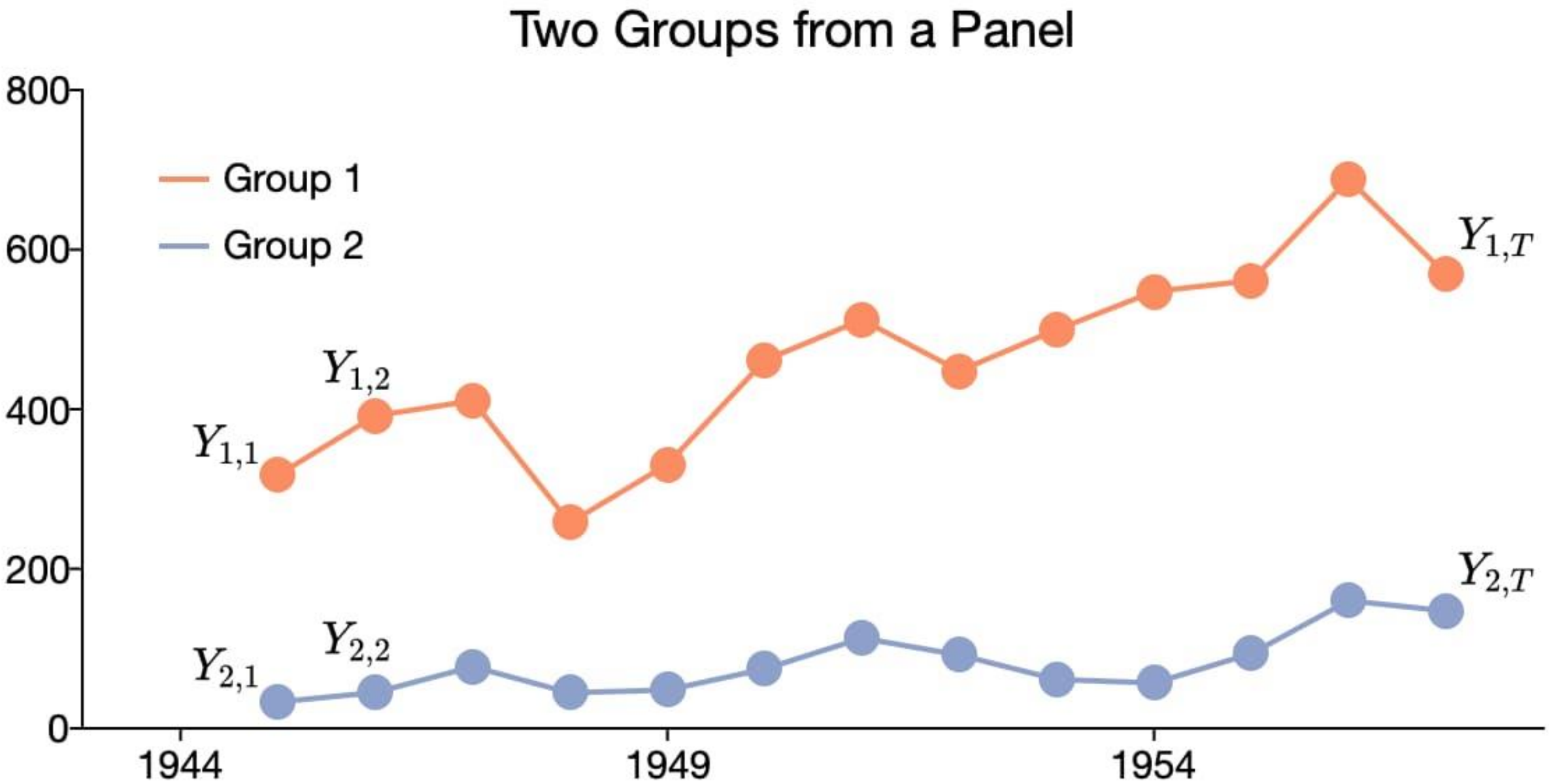
Observations in your sample must be independent from each other

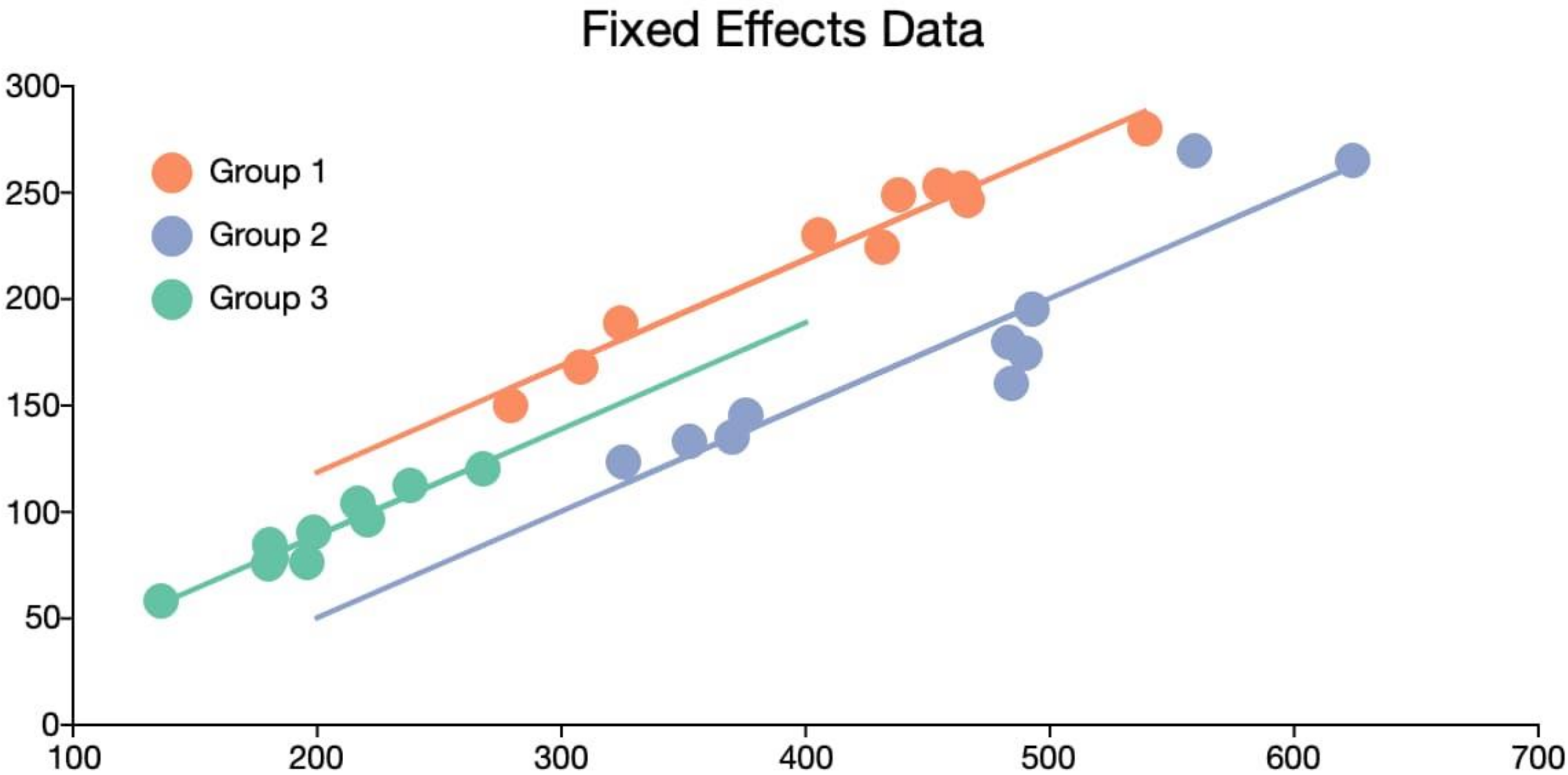
- i.e., The measurements for each sample subject are *in no way influenced by or related to the measurements of other subjects*.

If we have repeat observations of the same individual, we have *two* kinds of variation:

- Within group
- Across group

We need to conduct a Panel Regression


Two Groups from a Panel
800
600
400
200
0
Group 1
Group 2
$Y_{1,1}$
$Y_{1,2}$
$Y_{1,T}$
$Y_{2,1}$
$Y_{2,2}$
$Y_{2,T}$
1944
1949
1954


Fixed Effects Data
Group 1
Group 2
Group 3
300
250
200
150
100
50
0
100
200
300
400
500
600
700