Statistics&Probability II

Tian Zheng
Department of Statistics
Data Science Institute
Columbia University

Conditional Probability and Bayes' formula

Definition of Conditional Probability

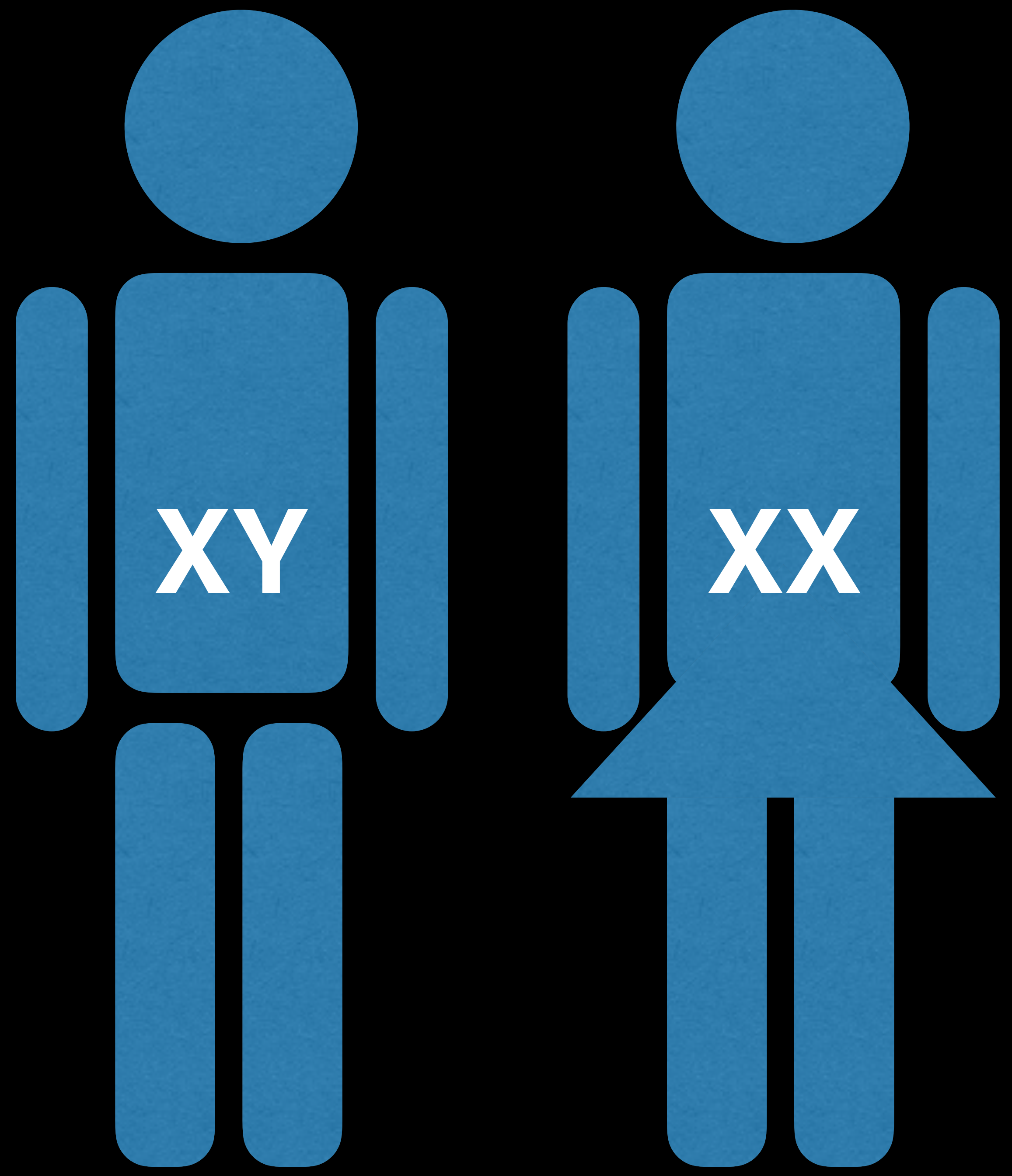
- For a family with 2 children, if we know that there is at least one girl in this family, what is the probability that the other child is also a girl?

Definition of
Conditional Probability P(A | B)

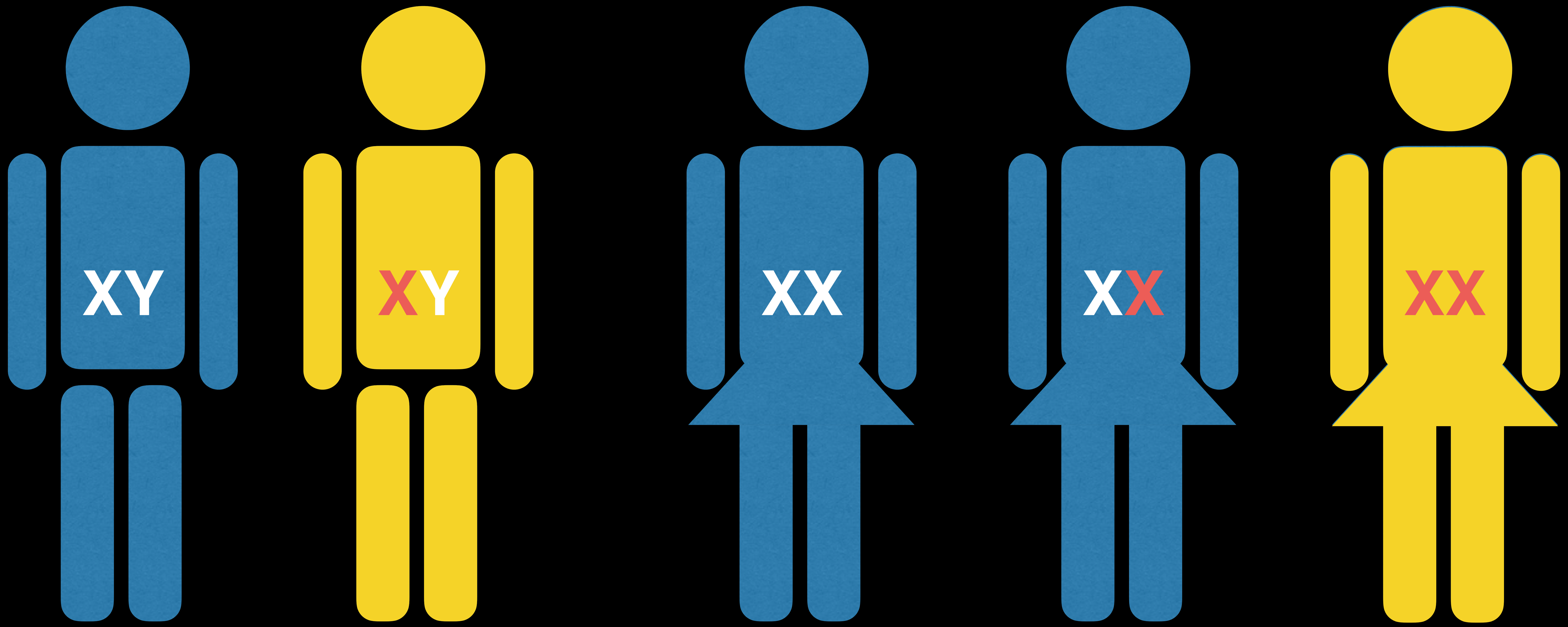
- For a family with 2 children, if we know that there is at least one girl in this family, what is the probability that the other child is also a girl?

Bayes' formula

Hemophilia example

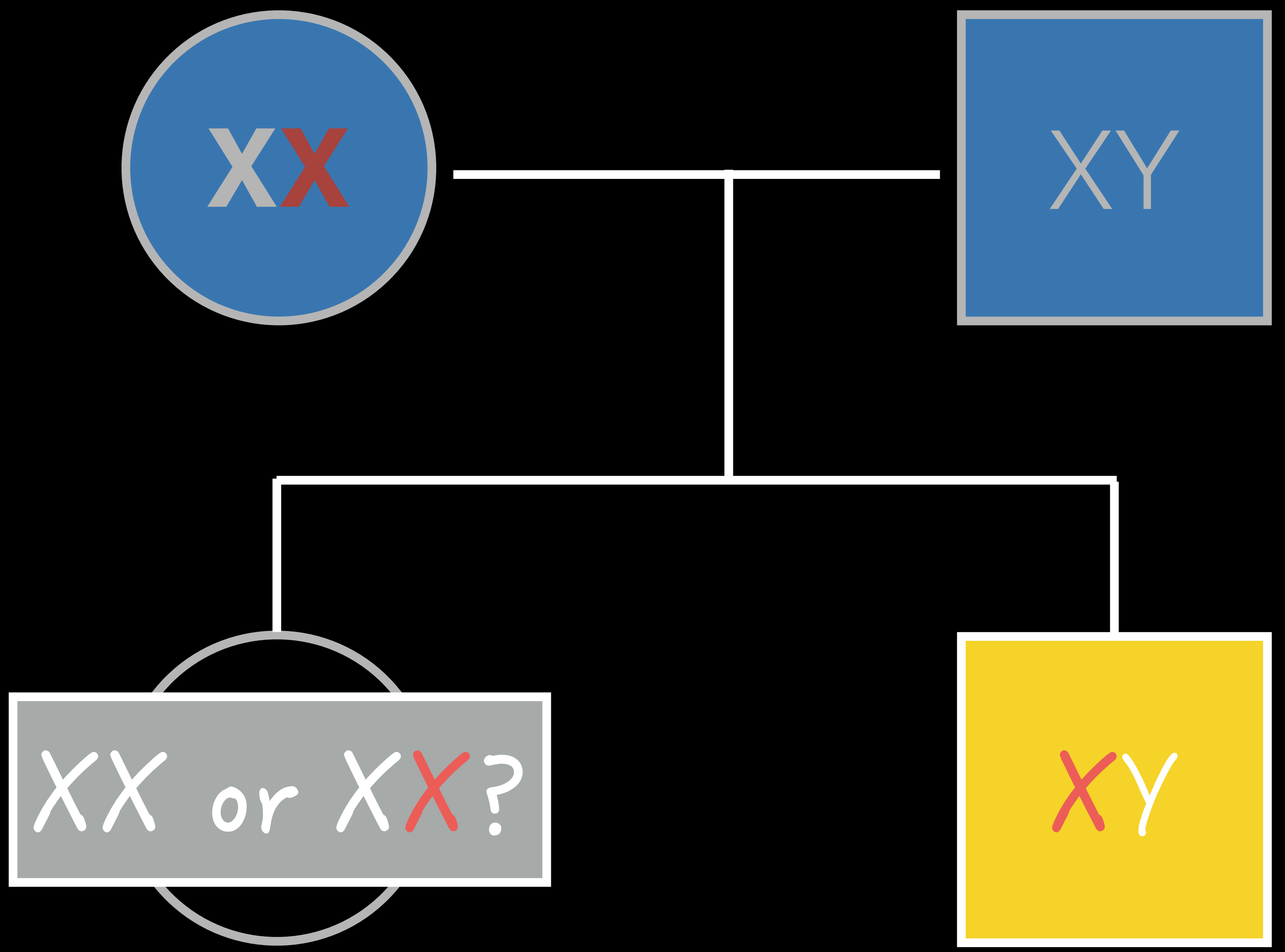


Hemophilia example



A woman with
an affected brother

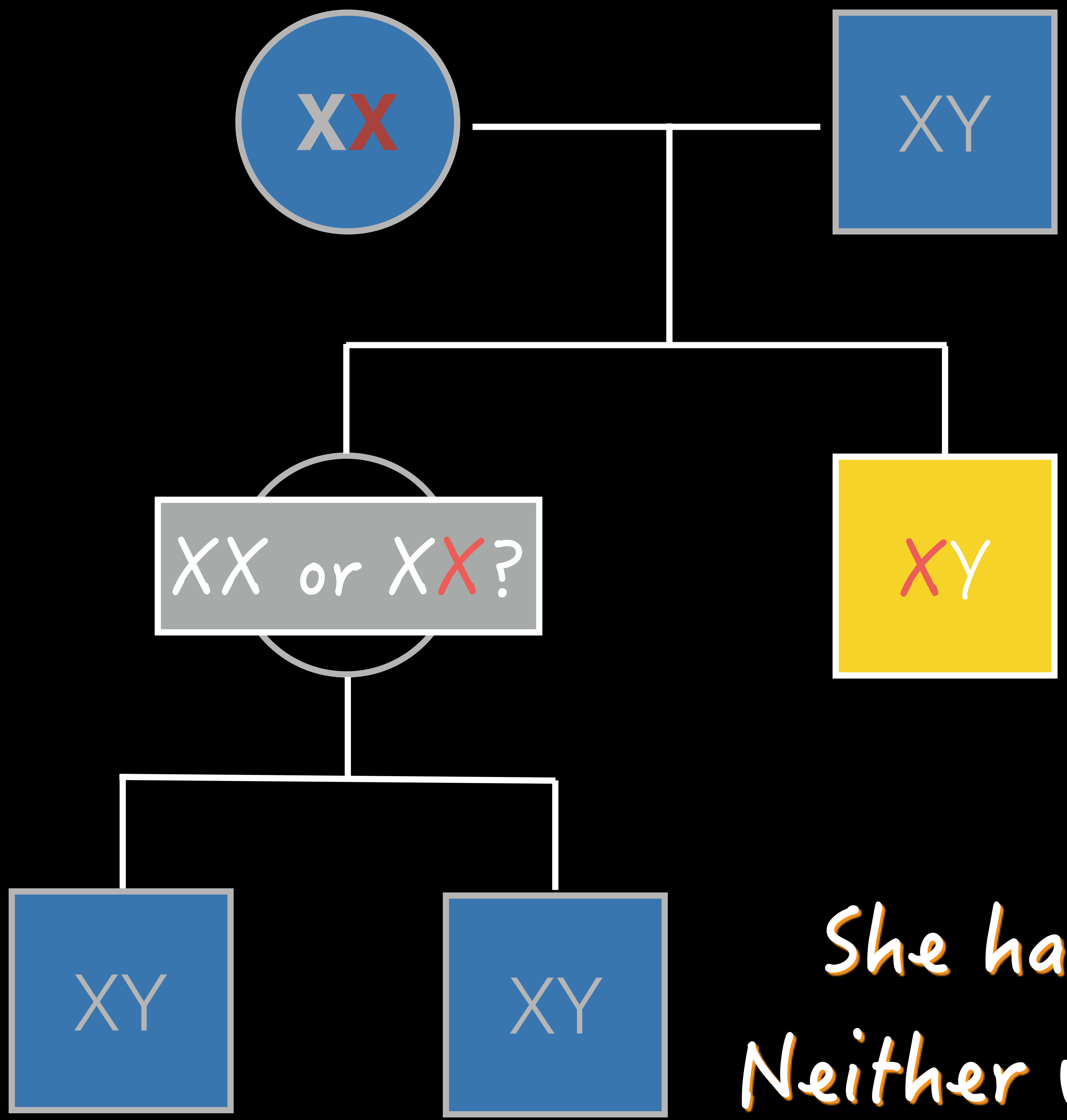
Neither parent was affected



$$P(XX)=P(XX)=0.5$$

Now we observe
some data

A: XX
"not A": XX
B: 2 sons with XY
P(B|A) = 1
P(B|not A) = (1/2)(1/2)
= 1/4



She has two sons.
Neither was affected

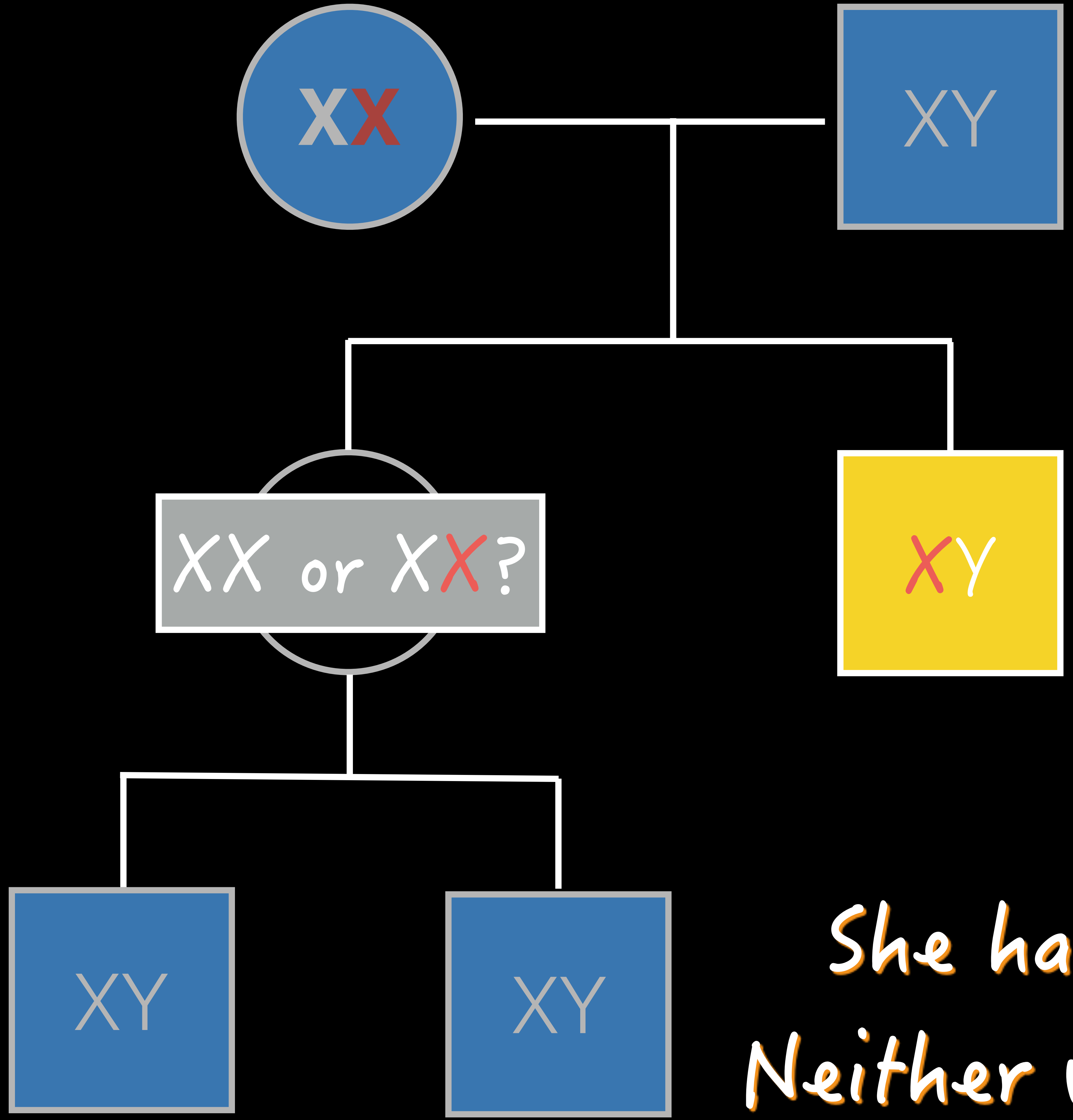
Bayes' formula

$P(A) = P(\text{not } A) = 1/2$

$P(B|A) = 1$

$P(B|\text{not } A) = 1/4$

$P(A|B) = ?$



She has two sons.
Neither was affected

More data

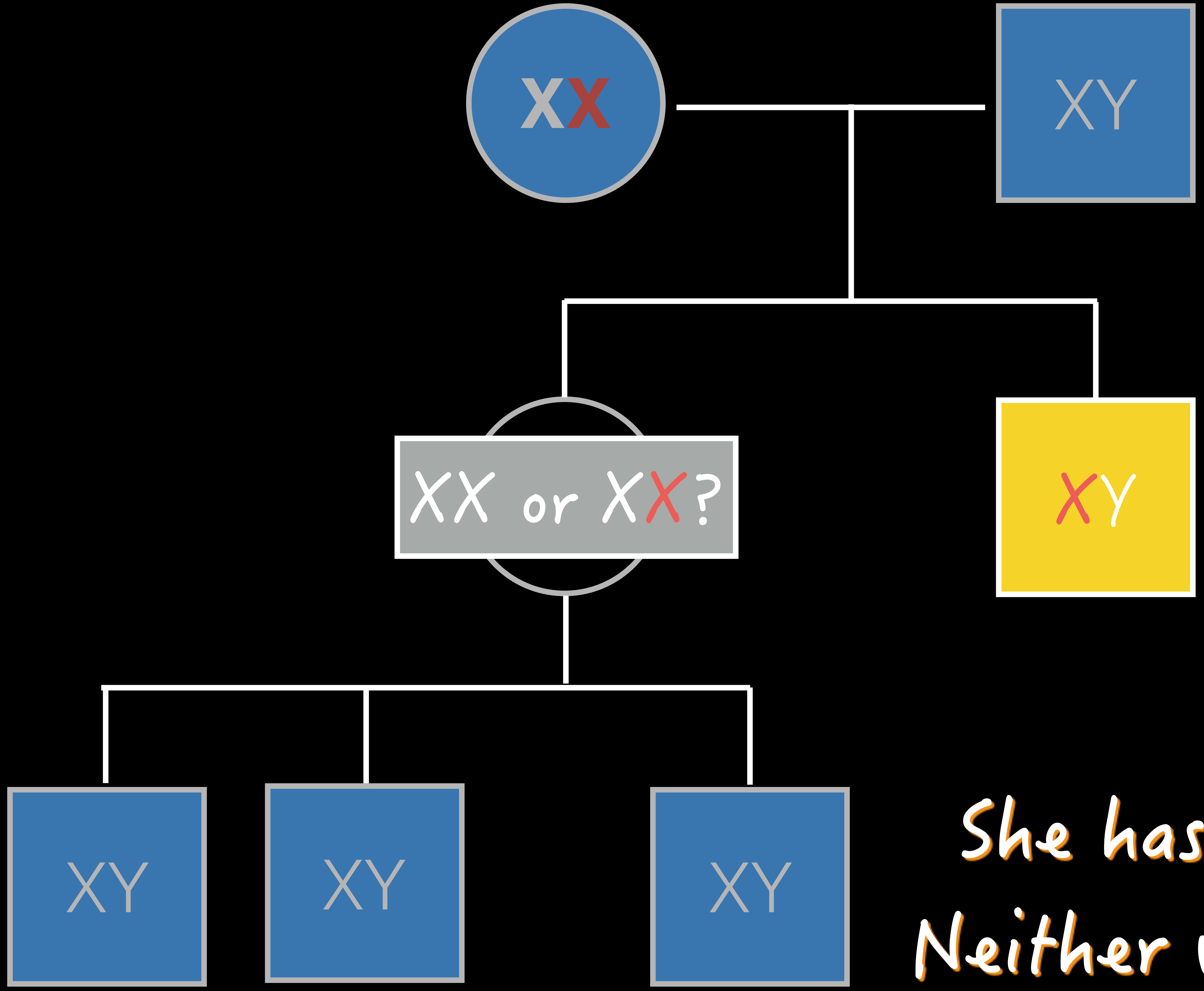
Now B: 3 unaffected sons

$P(A) = P(\text{not } A) = 1/2$

$P(B|A) = 1$

$P(B|\text{not } A) = 1/8$

$P(A|B) = ?$



She has three sons.
Neither was affected

Association: two-way table

Two way table

- Summarize joint occurrences of two categorical variables.

	Use Internet	Do not use	Total
18-29	48	2	50
30-49	93	7	100
50-64	85	15	100
65+	29	21	50
Total	255	45	

Probability under Independence

	Use Internet	Do not use
18-29	P(18-29, Use)	
30-49		
50-64		
65+		
	P(Use)	1-P(Use)

Marginal distribution

sum to 1

From data we can estimate the marginal distribution

- Summarize joint occurrences of two categorical variables.

	Use Internet	Do not use	Total
18-29	48	2	50
30-49	93	7	100
50-64	85	15	100
65+	29	21	50
Total	255	45	

From data we can *estimate* the marginal distribution

- Summarize joint occurrences of two categorical variables.

	Use Internet	Do not use	Total
18-29	????		.17
30-49			.33
50-64			.33
65+			.17
Total	.85	.15	

Relation between two categorical variables

- Say we are looking at **variable X** and **variable Y**, which are both **categorical**.
- Association: *certain values of X occur more frequently with certain values of Y.*
- No association: **independence**.

$$P(A \text{ and } B) = P(A)P(B)$$

Independence test
in a two-way table

- Looking for evidence on association?
  - the null hypothesis should be…**independence!**
  - **statistic measures association as *departure from independence***
- Chi-square test for independence in two-way table
  - the null hypothesis is a **special pattern.**
- What (pattern) does the null hypothesis imply?
  - No association: Given any event A decided by the values of X, and any event B decided by the values of Y, A and B are independent, P(A and B) = P(A)P(B)
  - Model probability of each cell = product of the marginal proportions of the outcome values of X and Y that define this cell.

Chi-square test

- Test statistic

$$\chi^2 = \sum_{\text{cell } i} \frac{(O_i - E_i)^2}{E_i}$$

- Statisticians already identified the distribution of χ^2 as a **chi-square distribution with (r-1)(c-1) d.f.**.

From data we can *estimate* the marginal distribution

- Summarize joint occurrences of two categorical variables.

	Use Internet	Do not use	Total
18-29	????		.17
30-49			.33
50-64			.33
65+			.17
Total	.85	.15	

Expected under independence

- Summarize joint occurrences of two categorical variables.

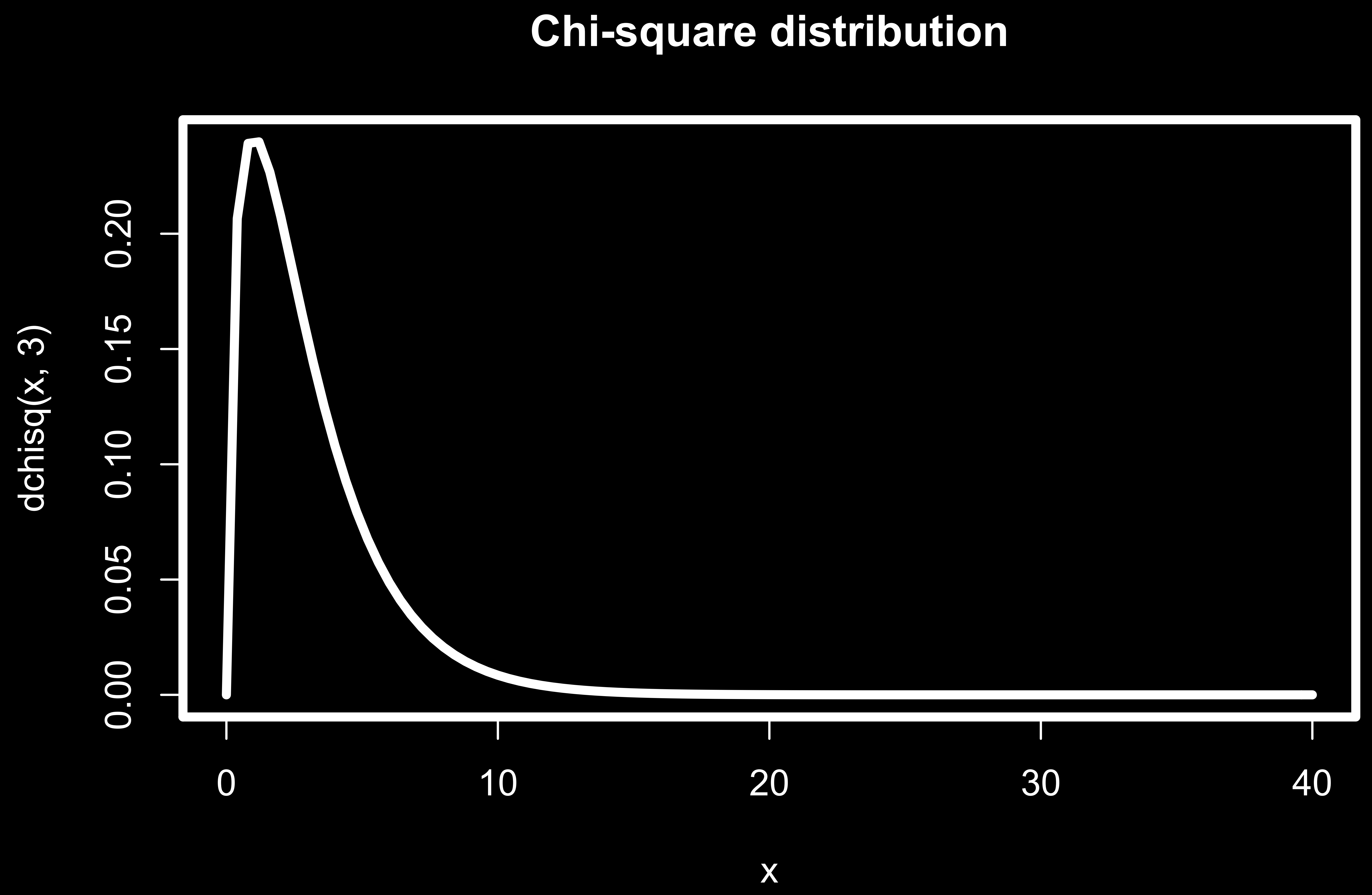
	Use Internet	Do not use	Total
18-29	42.5	7.5	50
30-49	85	15	100
50-64	85	15	100
65+	42.5	7.5	50
Total	255	45	

Expected versus observed

- Summarize joint occurrences of two categorical variables.

	Use Internet	Do not use
18-29	48 5.452.5	2 -5.75.5
30-49	93 8 85	7 -815
50-64	85 0 85	15015
65+	29-1348.5	2113.5.5

chi-square statistic = 38.34



Detecting Association versus measuring association

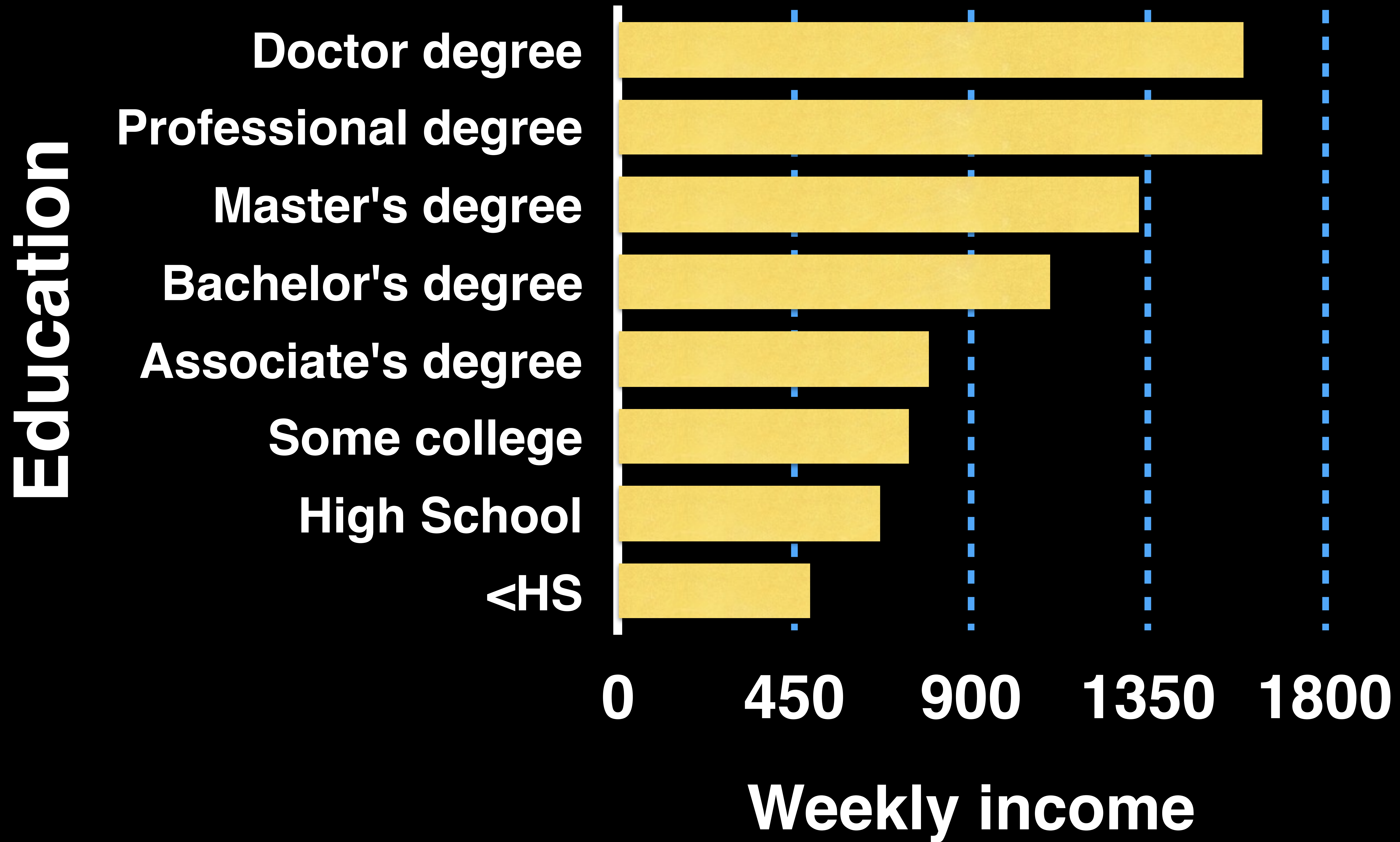
- Sampling distribution of test statistic
  - Given the same joint probability model between X and Y that shows association, the test statistic’s center scales with sample size.
  - The level association is decided by the probability relation between X and Y, not by sample size.
  - Statistics such as odds ratio can be used to show extent of association.

Implication of association

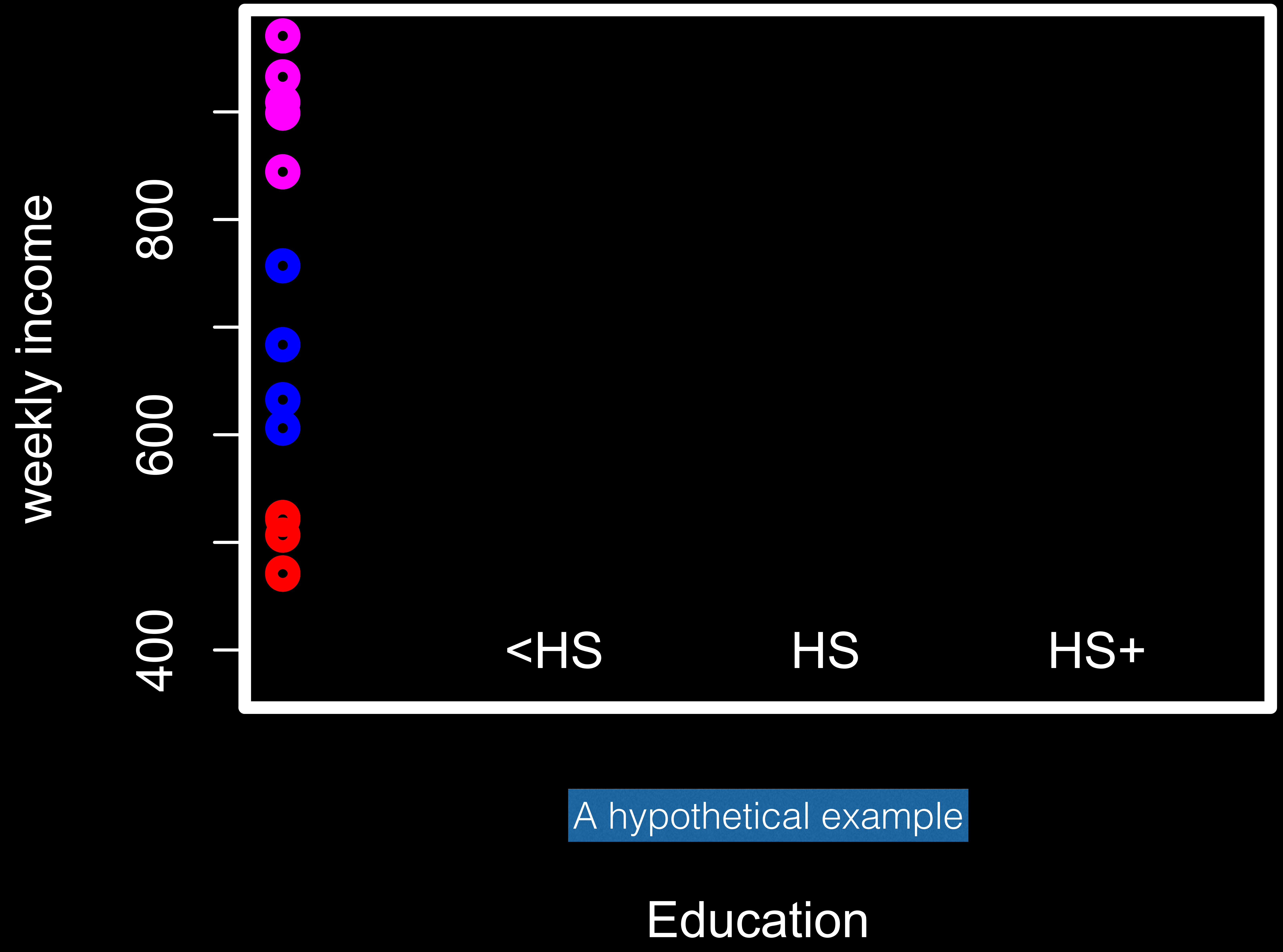
- Or “X and Y are associated. So what?”
- $P(X|Y)$ differs from $P(X)$ to provide more information than the marginal probabilities.
- In other words, we can achieve better prediction of Y given information in X.
- Is it causal relation then?

Association: analysis of variance

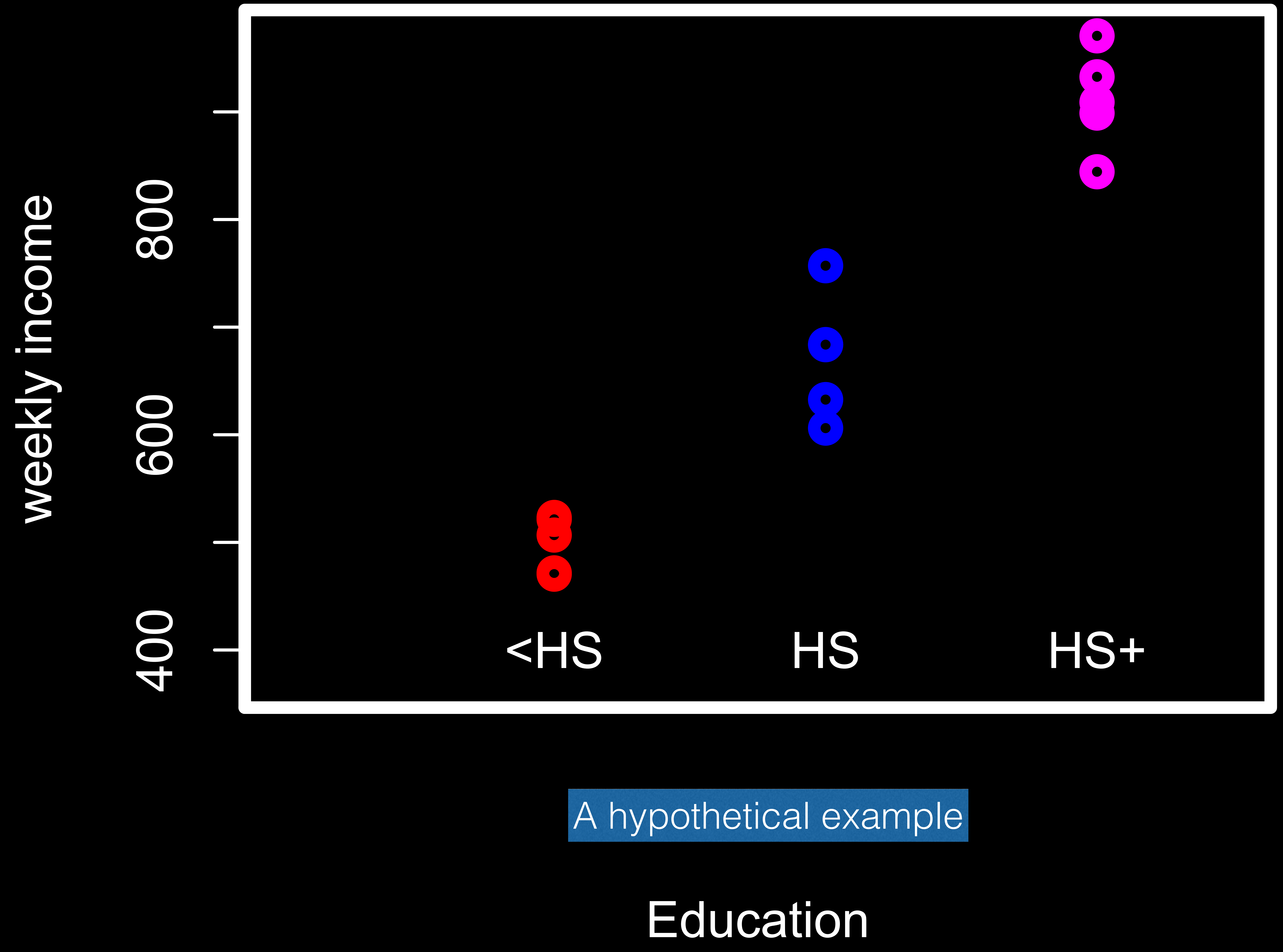
Education versus Income



Data source: Bureau of Labor Statistics (2014)



weekly income
400
600
800
<HS
HS
HS+
A hypothetical example
Education

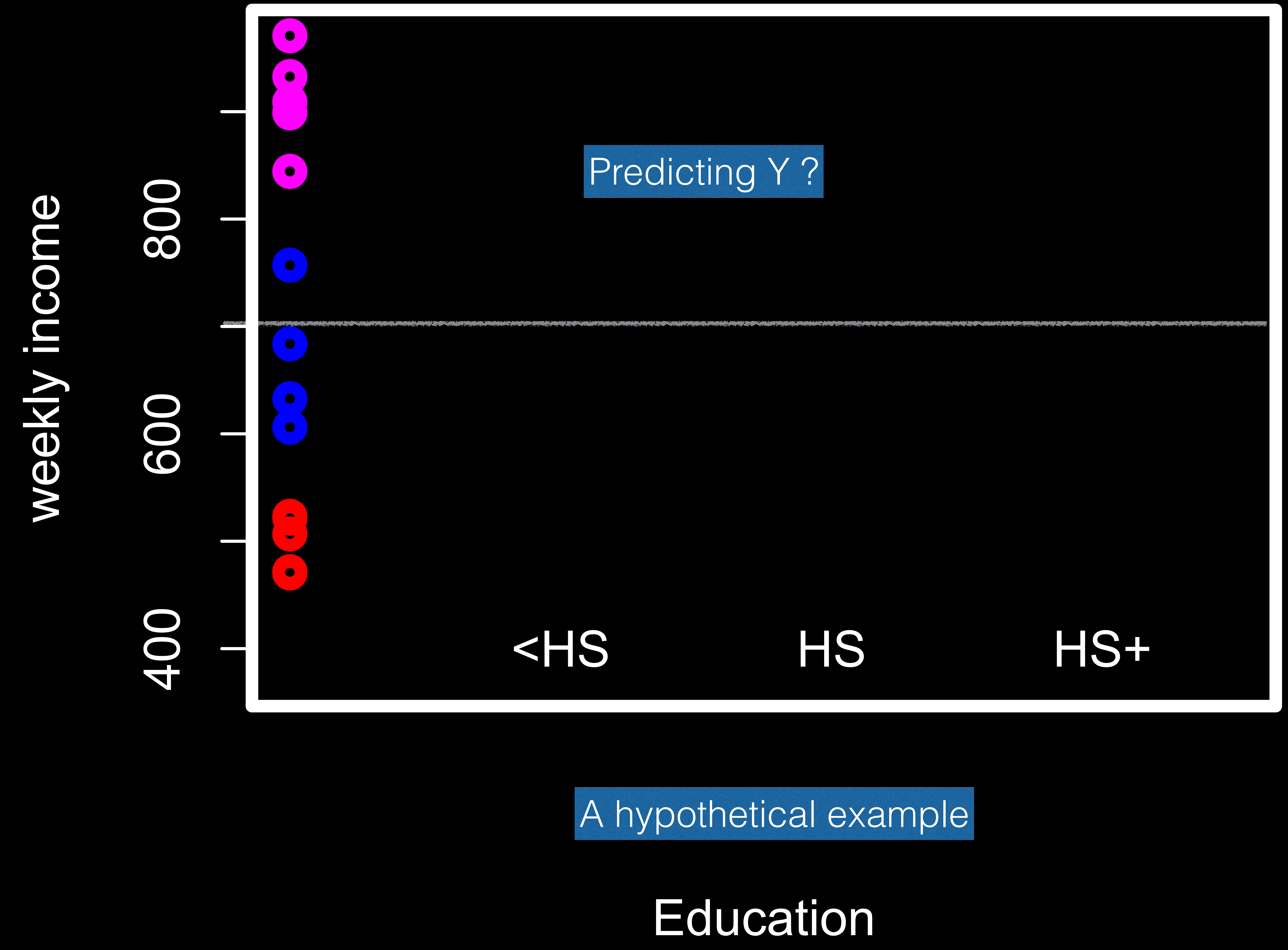

weekly income
400
600
800
<HS
HS
HS+
A hypothetical example
Education

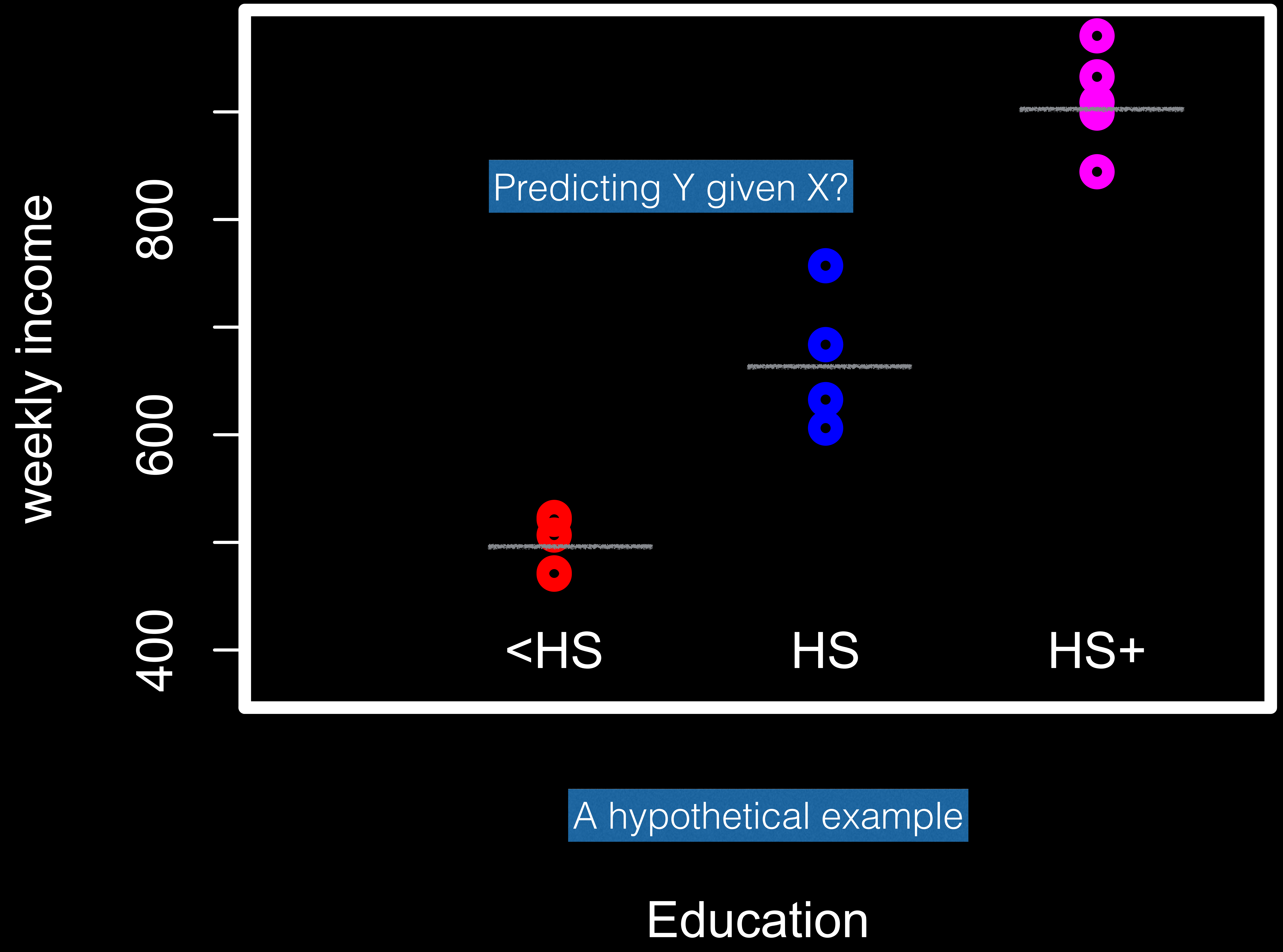
Analysis of variance
(ANOVA)

- It concerns the total variation in Y (the quantitative variable) — Income.

- Considering a variable X, we would like to know how much variation in Y can be explained by X.

- Still not necessarily a causal relation.


Predicting Y ?
weekly income
400
600
800
<HS
HS
HS+
A hypothetical example
Education


Predicting Y given X?
weekly income
400
600
800
<HS
HS
HS+
A hypothetical example
Education

Sources of variation

$$SSG = \sum_{groups} n_i(\bar{x}_i - \bar{x})^2$$

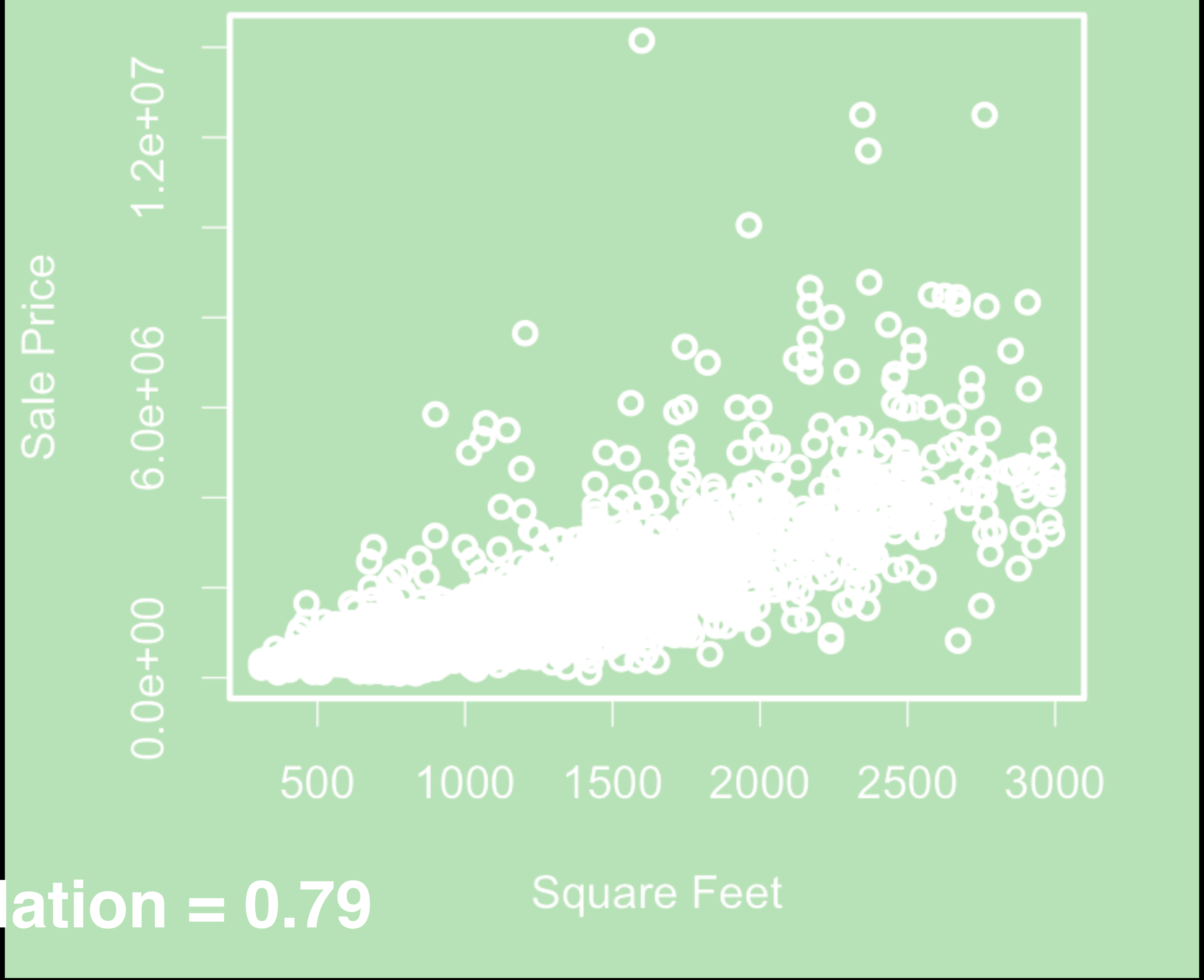
$$SSE = \sum_{all\ obs} (x_{ij} - \bar{x}_i)^2$$

$$SSTO = \sum_{all\ obs} (x_{ij} - \bar{x})^2$$

Manhattan Condo Prices

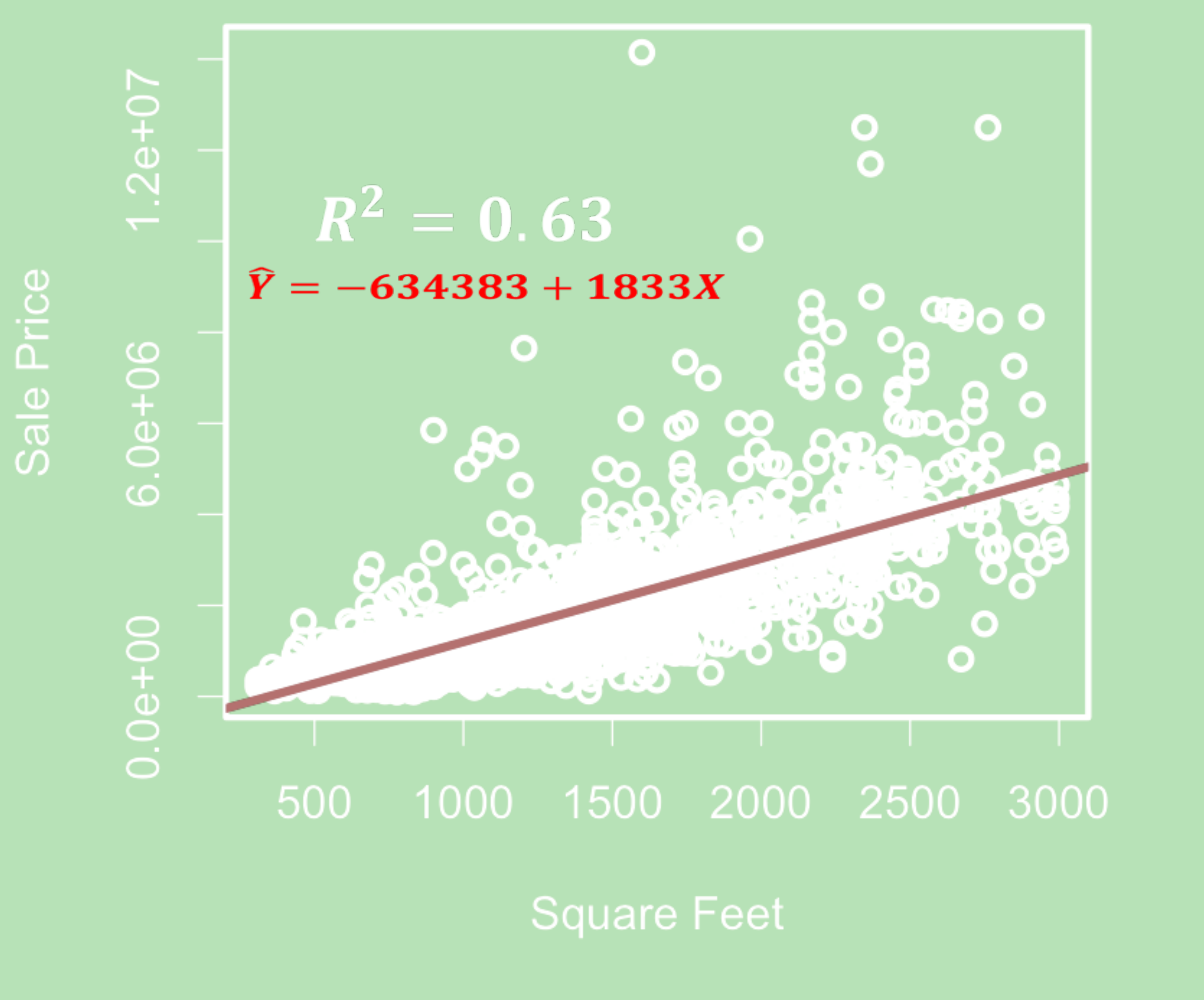
- From NYC Open Data
- Year 2009
- Condo building with elevator
- 3553 apartments

Manhattan Condo Prices

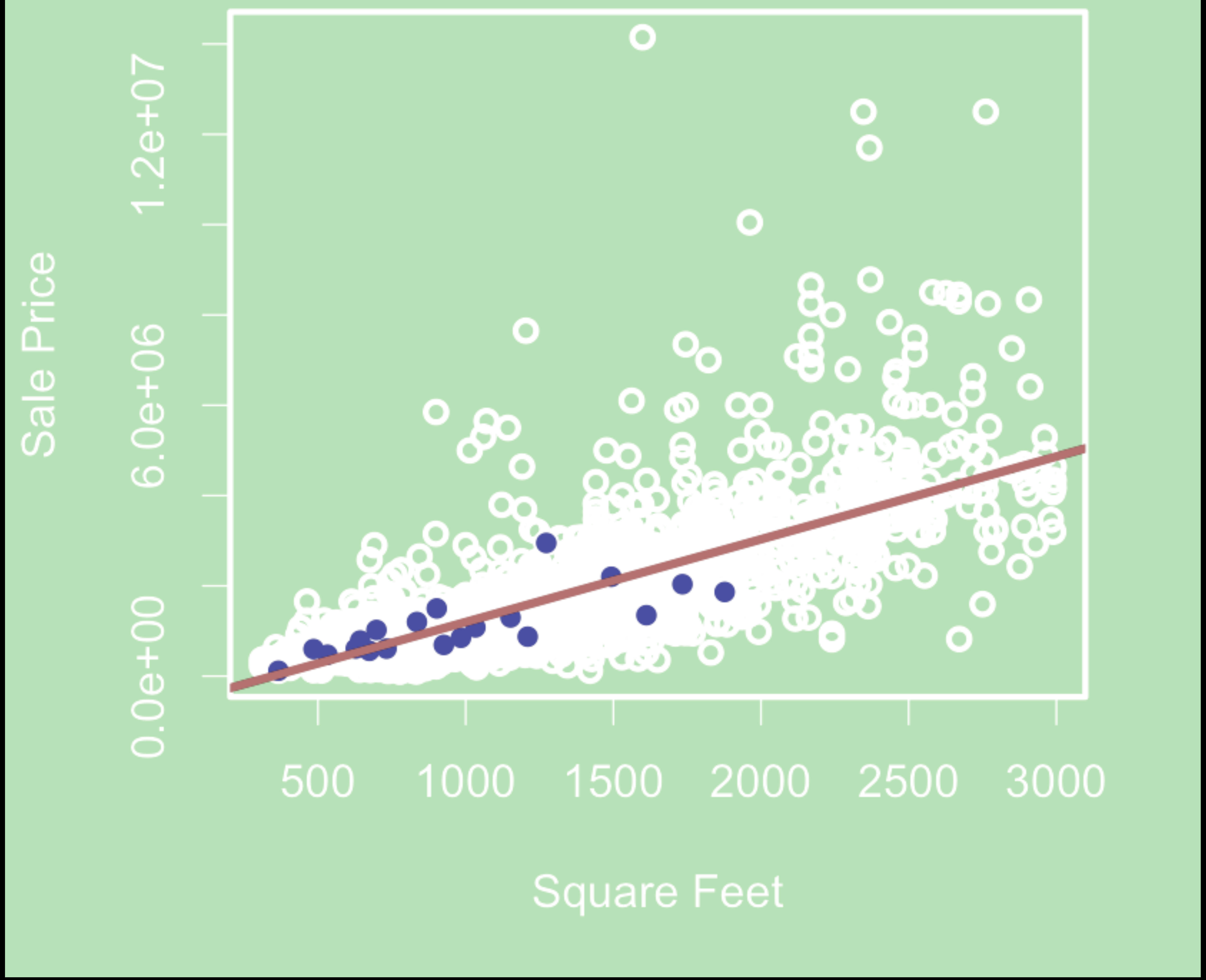


Correlation = 0.79

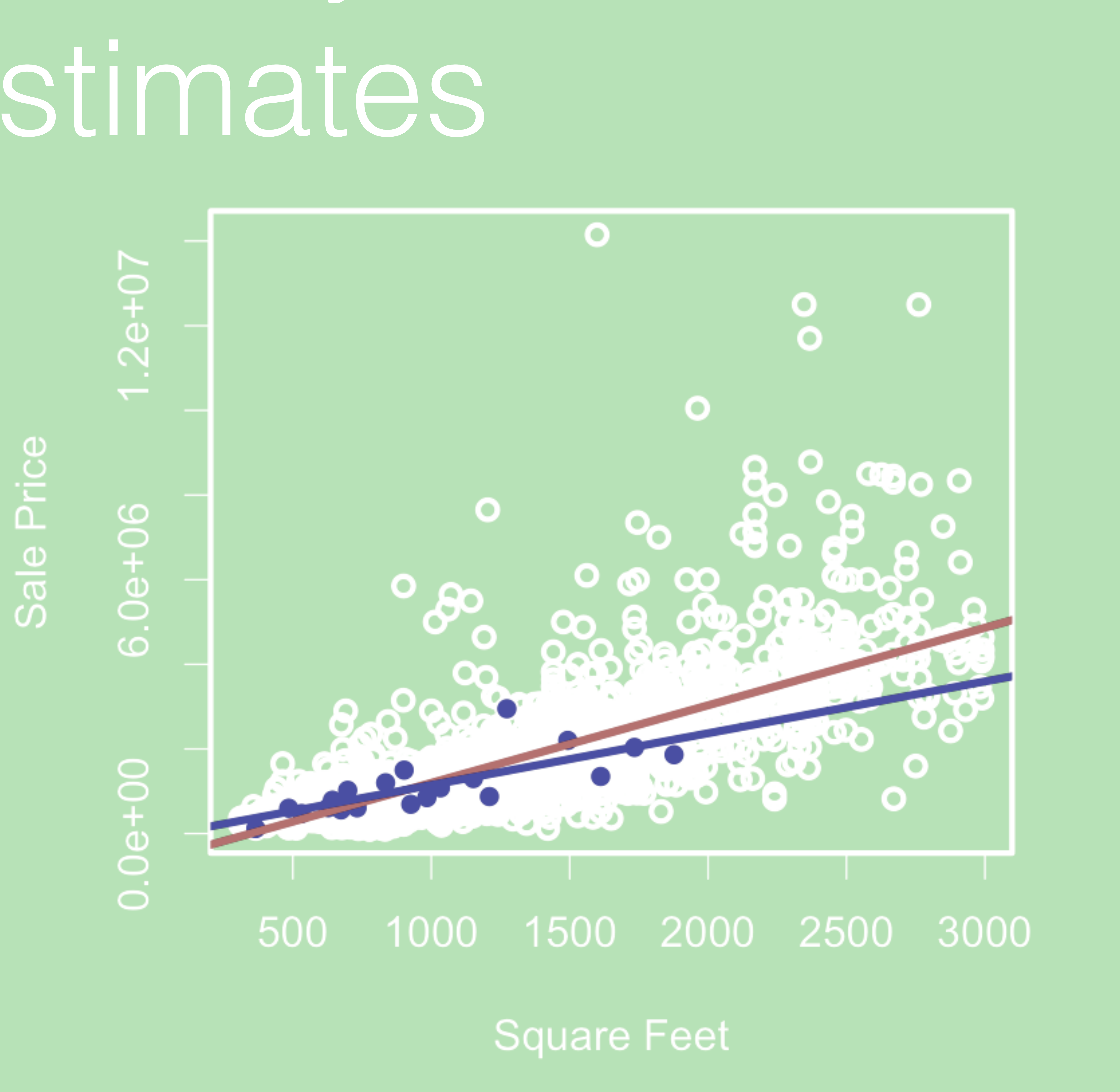
Manhattan Condo Prices



Sampling variability in regression estimates

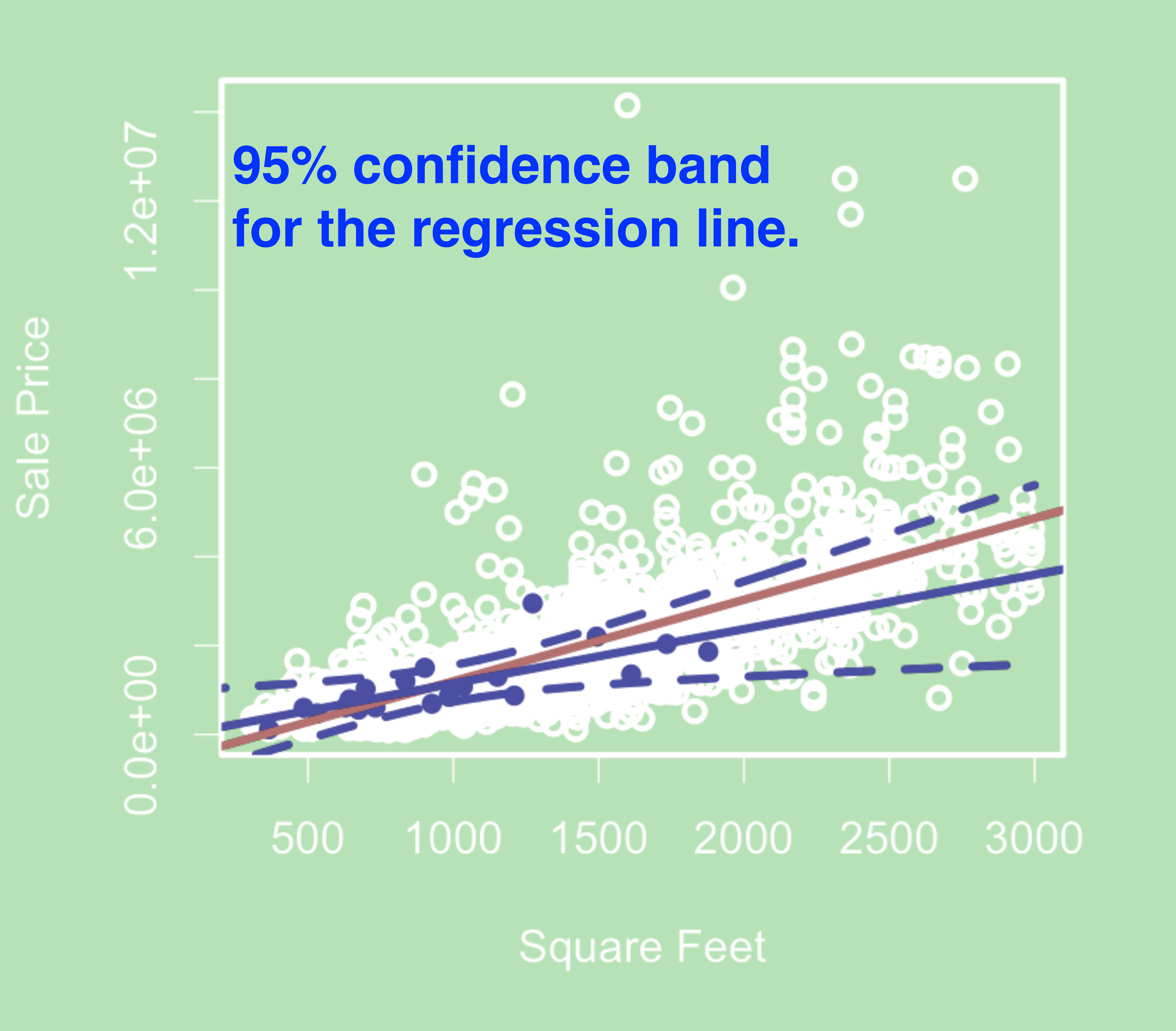


Sampling variability in regression estimates



Sampling variability in regression estimates

- **The confidence band centers at the sample estimate.**
- **It represents interval estimate for the regression line.**
- **Other inference on regression estimates can also be carried out.**



Prediction

- Given a value of X
- The predicted value is $\widehat{Y} = b_0 + b_1 X$
- It is an estimate for the mean (average) value for Y given the X value.
- Most of the time, prediction is **different** from what is actually observed.
  - $\boldsymbol{Y - mean\ of\ Y}$ **(random variation)**
  - $\boldsymbol{mean\ of\ Y - \widehat{Y}}$ **(estimation error)**

Prediction

- Extrapolation happens when one tries to give prediction on values of X outside the data range.

Multiple regression

- **Y: response**
- **Multiple X variables**
- $\hat{Y} = -546944 - 3265\ Age + 1770\ SQFT$
- **Consider interaction**
- $\hat{Y} = -743800 + 3137\ Age + 1996\ SQFT - 5.213\ Age \times SQFT$
- $\hat{Y} = (-743800 + 3137\ Age) + (1996 - 5.213\ Age)\ SQFT$

Multiple regression