REGRESSION AND CAUSALITY

PROF. XINXIN LI

CAUSAL VS. PREDICTIVE

Causality

- Discovering whether or not x truly causes y.
- For example, did a particular action (say, a Facebook promotion) *cause* a particular outcome (increased sales)? Or was the change due to some other unseen factor?

Predictive modeling

- What's likely to happen next? With enough data, researchers can observe patterns and develop models to predict likely outcomes under different circumstances.
- For example, which customers are likely to cancel our service in the next 30 days?

CAUSAL VS. PREDICTIVE

When should we do prediction vs. causal analysis?

- Suppose an insurance company wants to know whether people who do exercise live longer?

Which method should they use?

- Suppose a policy maker wants to know whether people who do exercise live longer?

What method should they use?

REGRESSION METHOD

- Regression method can be used to answer both causal and predictive questions, but there are differences in how it is used.

- The basic regression formulation for the previous question is

$$\text{Life expectancy} = b_0 + b_1 * \text{Amount of exercise}$$

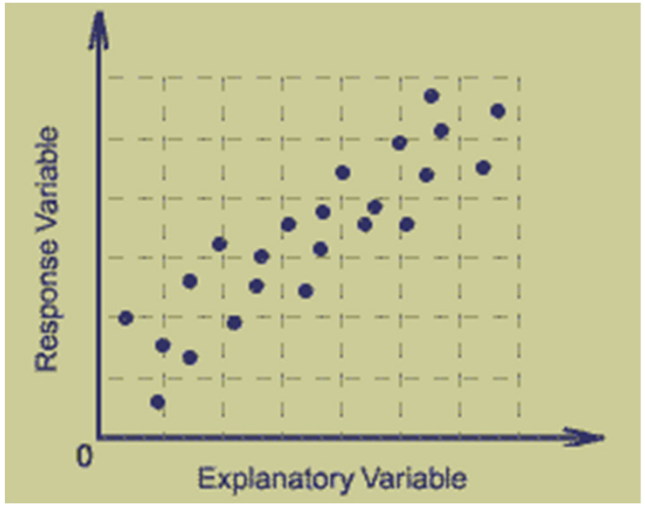
- Regression analysis is used to discover values of b_0 and b_1 from data, i.e., the relationship between "Life expectancy" and "Amount of exercise"

- In case of prediction, we are interested in finding value of "Life expectancy" given the "Amount of exercise" for a particular person

- In case of casual analysis, we are interested in finding what is the true value of "b_1"

REGRESSION ANALYSIS

- At its simplest, it is fitting a line to a scatterplot of data

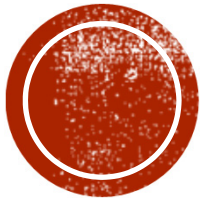


Equation of a line: $Y = b_0 + b_1 X$

Suppose $b_0=1$ and $b_1=2$

We can say that at X=5, the value of Y is **predicted** to be: $Y = 1 + 2*5 = 11$

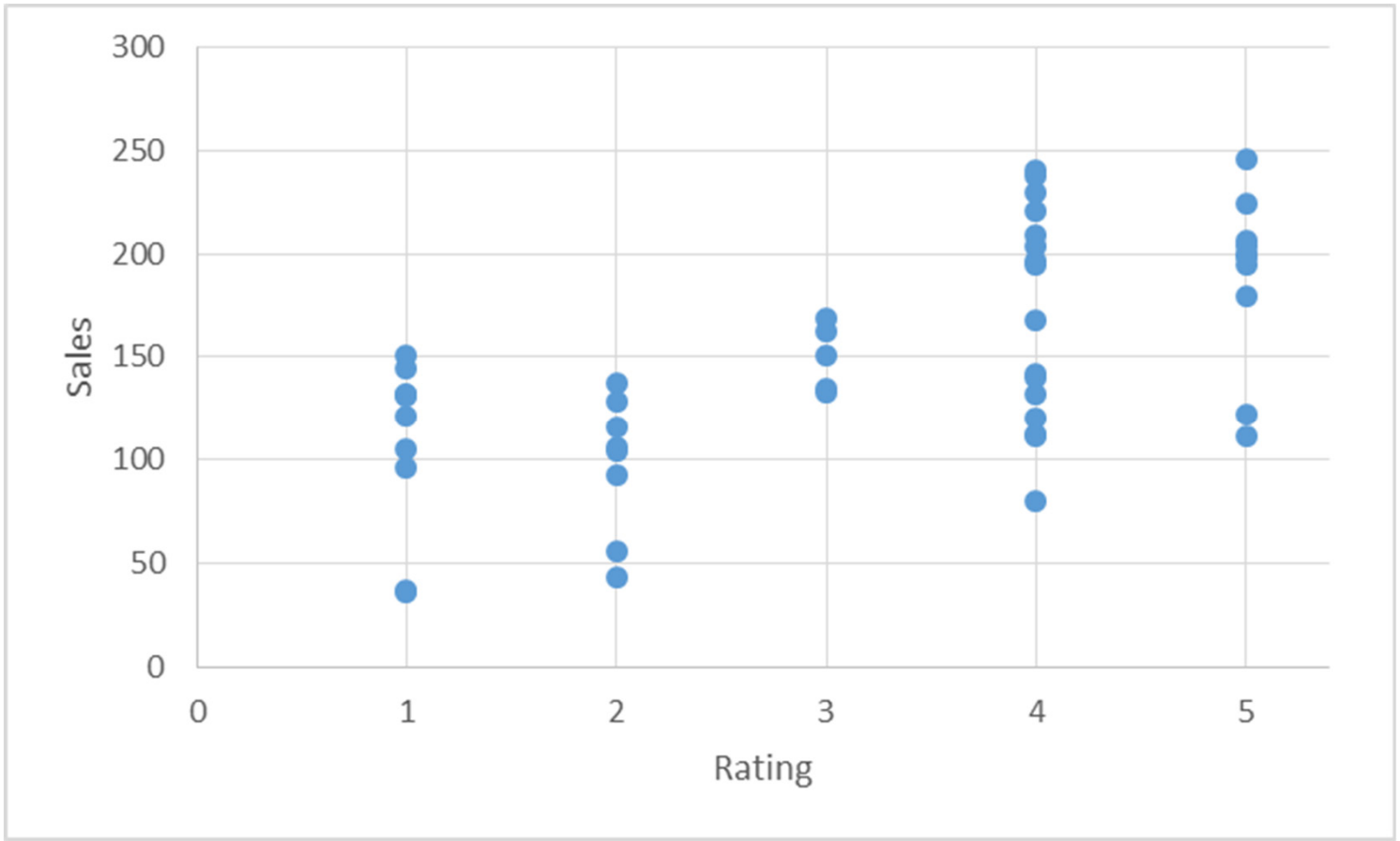
We can also say that that increasing X by 1 unit increases Y by 2 units (as $b_1=2$), which is a **causal** statement



I. SIMPLE REGRESSION

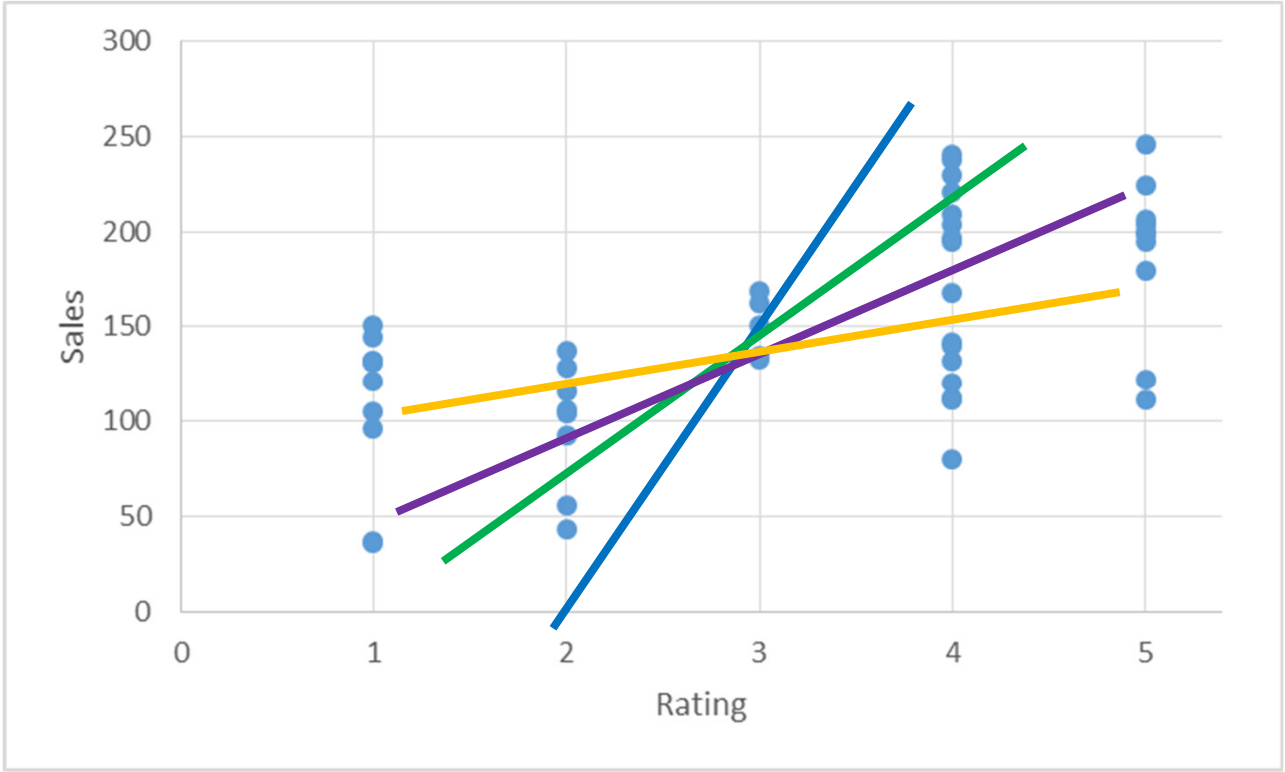
EXAMPLE: BOOK SALES

- What is the relationship between sales of books and their quality?
- Data with sales and ratings of 49 books



FITTING A LINE

- The estimated equation is $\hat{y} = b_0 + b_1 x$
  - b_0 is the intercept of the line
  - b_1 is the slope of the line
  - $\hat{y}$ is the ***estimated*** (or ***fitted***) value of y for a given x value
- Then, which line most closely matches the observed relationship between y and x?



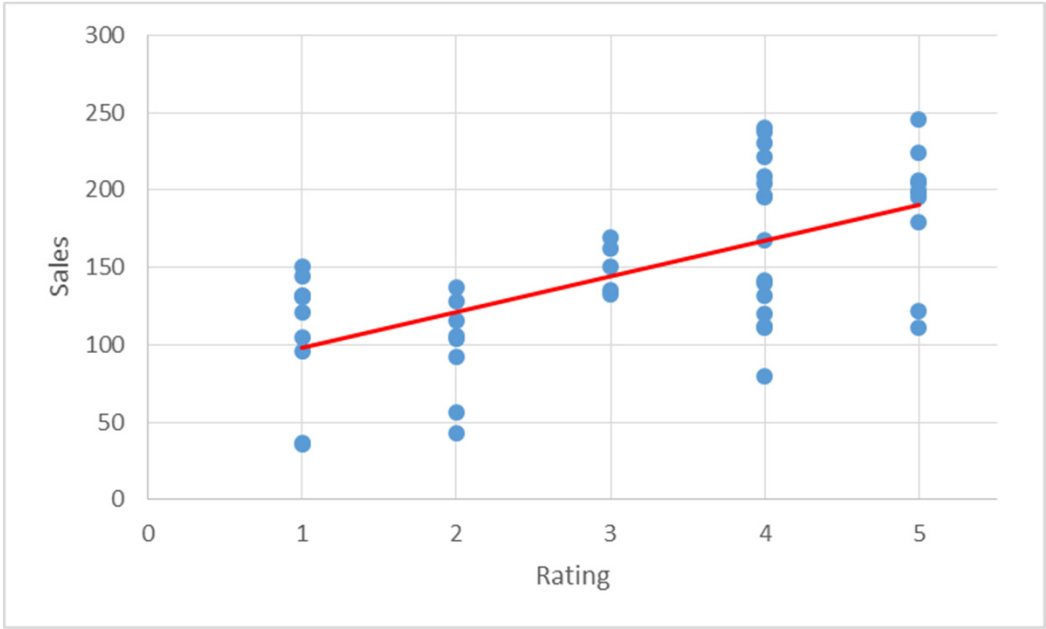
CHOOSING THE RIGHT LINE – ORDINARY LEAST SQUARE (OLS) METHOD

- The error in estimation is given by: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$
- The e_i's are called the residuals
- Choose b_0 and b_1 such that they minimize sum of squared residuals $\sum_i (e_i{}^2)$
- Why square the residuals?

- OLS Estimation:

  Sales = 74.63 + 23.12 * Rating

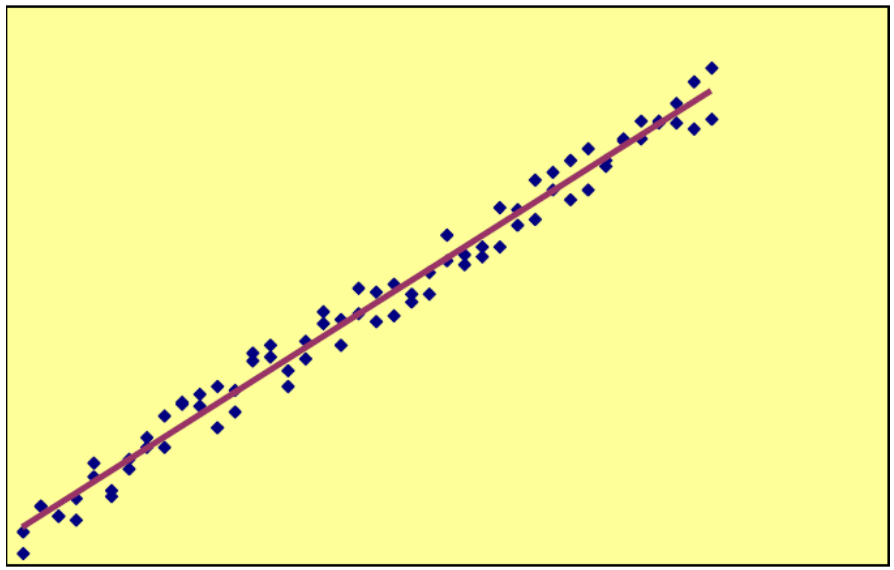
  Does it make sense to have a book with a negative rating or a rating greater than 5?

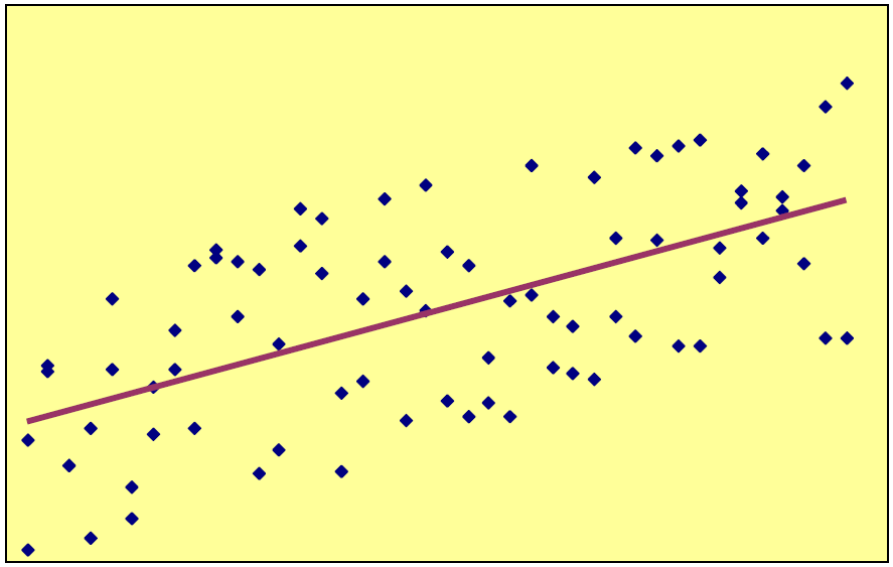


USING THE FITTED LINE

- For predicting sales of a book with a certain rating
  - What is the sales of a book with a rating of 4?
  - Different books with a rating of 4 will have different prices. But, the average sales of such books will be 74.63 + 23.12 * 4 = 167.11
- Ascribing causal meaning
  - The sales of books that are rated higher by 1 star increase by 23.12, on average
  - Again, this meaning is applicable within the range of the data

HOW GOOD IS THE MODEL FIT?





- Two different summary statistics to look at
  - R^2, the coefficient of determination, measures the proportion of the variance in the dependent variable that can be explained by the independent variable(s).
  - Root Mean Square Error (RMSE) measures the difference between the values observed and the values estimated by the regression line

HOW GOOD IS THE MODEL FIT?

Remember the estimated equation is: $\hat{y} = b_0 + b_1 x$

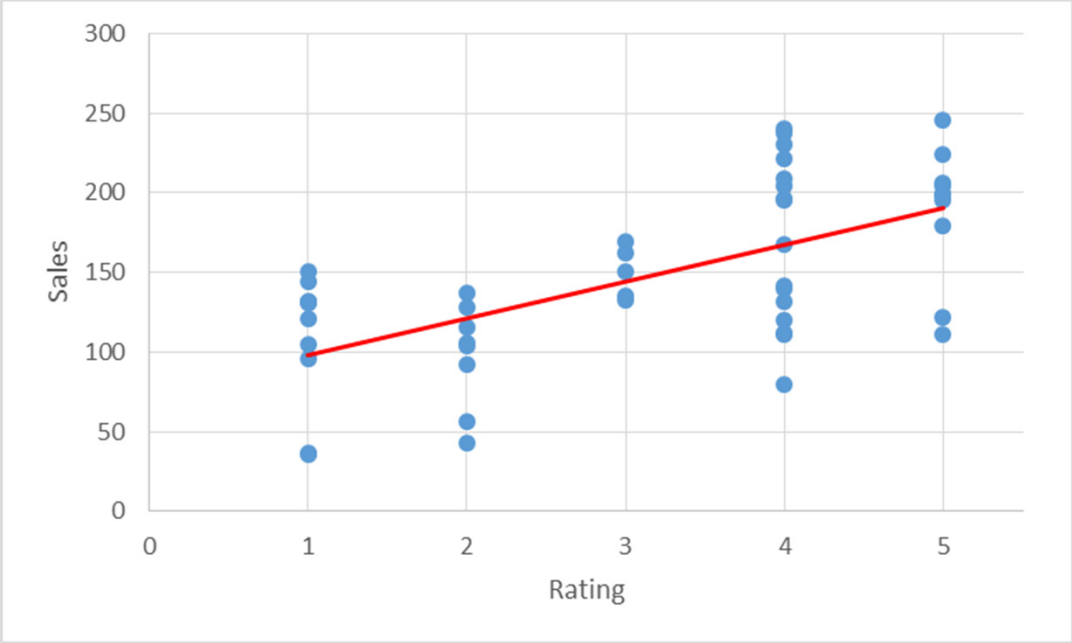
- Root Mean Square Error (RMSE) = $\sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}}$

- $R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2 / n}{\sum_i (y_i - \bar{y})^2 / n} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$
  - R^2 always increases with the number of independent variables

- Adjusted $R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2 / (n-k-1)}{\sum_i (y_i - \bar{y})^2 / (n-1)} = 1 - (1 - R^2) \frac{n-1}{n-k-1}$
  - k is the number of independent variables (excluding the constant term)
  - Penalizes more complicated models, since adding additional variables may not add a lot more explanation, but runs the risk of multicollinearity

GOODNESS OF FIT: BOOK SALES

Sales = 74.63 + 23.12 * Rating

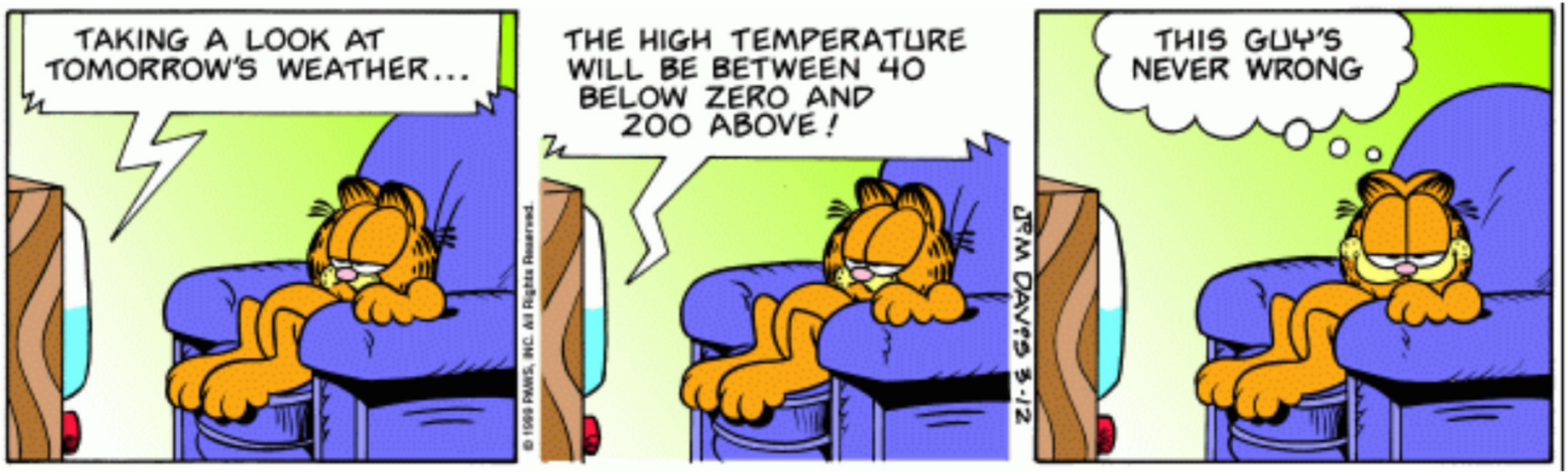


- $R^2 = 0.3762$: the estimated linear function of ratings explains 37.62% of the variation in sales
- Root Mean Squared Error (RMSE) = 42.8

INFERENCE IN REGRESSION

- Until now, we studied how to find the best fit line to the sample data
  - Using OLS, we found b_0 and b_1
  - If b_1 is non-zero then change of x is related to change of y
- We want to infer whether the relationship found in the **sample** extends to the **population**
  - **i.e., we want to know the value of b_0 and b_1 for the population – we call these β_0 and β_1 (called population parameters)**
- Since β_0 and β_1 may not be exactly the same as b_0 and b_1, we are not sure about their actual value, so we provide a range of values rather than a specific number

CONFIDENCE INTERVALS: A RANGE RATHER THAN A POINT VALUE WHEN YOU ARE NOT CERTAIN

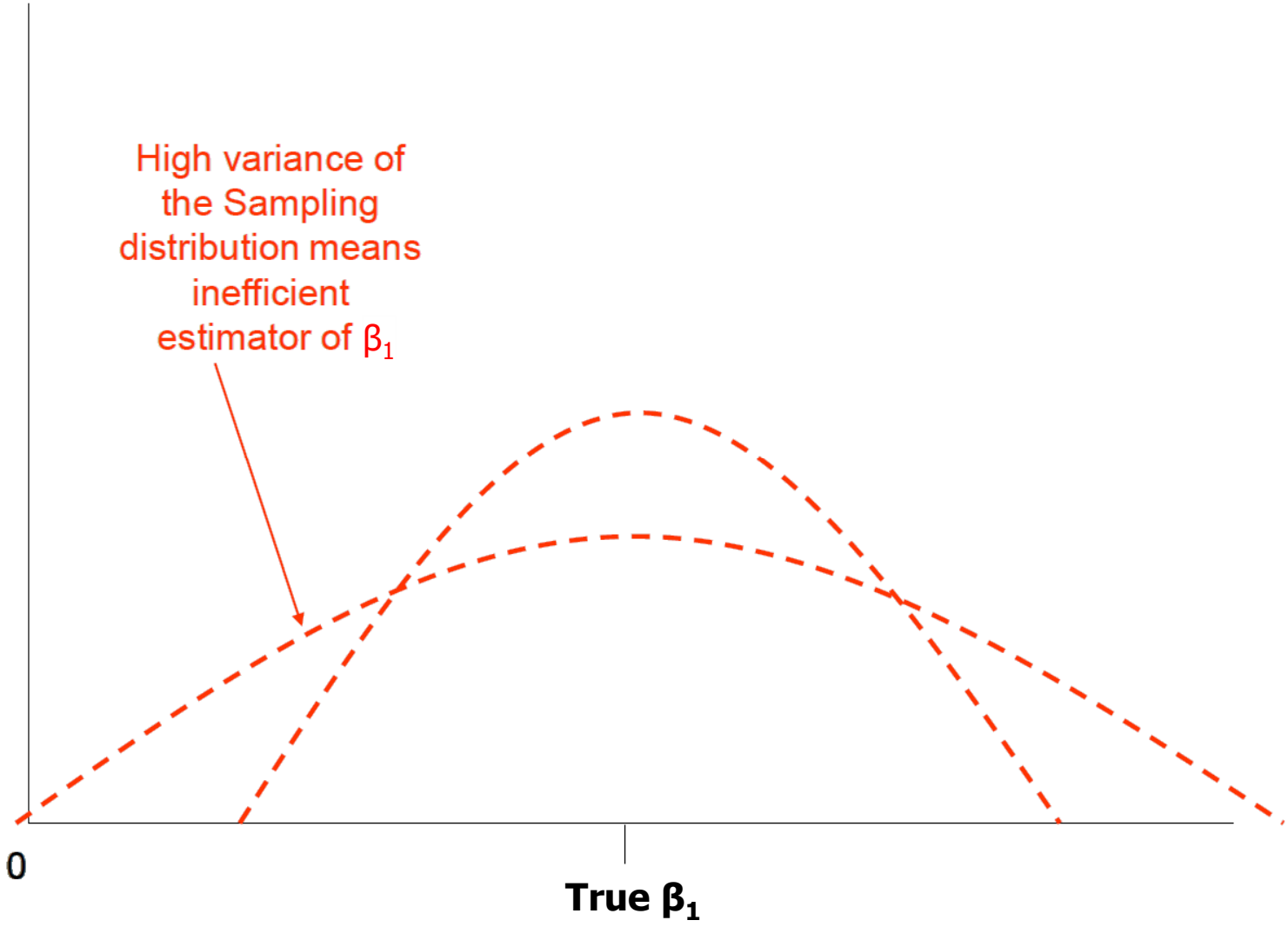


The smaller the range, the more informative it is

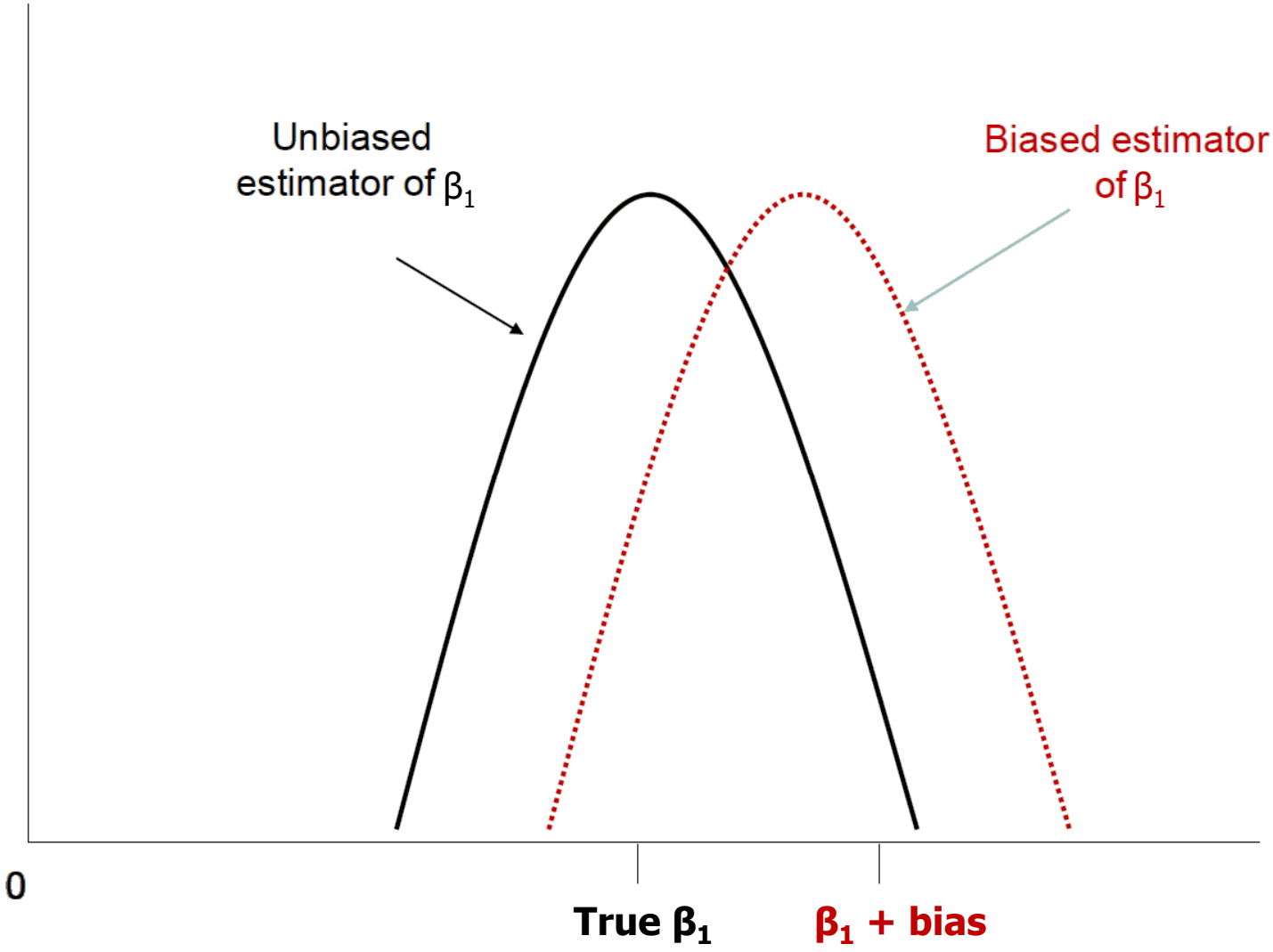
CONFIDENCE INTERVAL FOR POPULATION PARAMETERS

- To get a better idea on the value of the population parameters, β_0 and β_1, we can use **confidence intervals for the population parameters, or do hypothesis testing**, and to do that we need to know the distribution of b_0 (intercept) and b_1 (slope)
  - CLT (Central Limit Theorem) implies that these distributions follow t-distribution with n-2 degrees of freedom, where n is the number of data points in the sample
  - Distributions of b_0 and b_1 are called sampling distributions of intercept and slope, respectively
- The distributions must have the important properties of unbiasedness and efficiency

EFFICIENCY OF THE SAMPLING DISTRIBUTION



UNBIASEDNESS OF THE SAMPLING DISTRIBUTION



SIMPLE LINEAR REGRESSION MODEL

- Use a linear equation to model the population relationship between the variables

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - Y: Outcome or dependent variable – the variable we are interested in explaining.
  - X: Explanatory or independent variable – the variable that is useful in explaining.
  - β_0 and β_1: parameters of the model
  - ε : error term (disturbance or noise)

- We need to make assumptions about the error term to ensure that distributions of β_0 and β_1 have good properties (high efficiency and unbiasedness)
  - A very important assumption requires that the errors are not correlated (systematically related) to the value of X, i.e., $Corr(X, \varepsilon) = 0$

OLS REGRESSION OUTPUT: BOOK SALES

Estimate

Inference

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    74.630     14.694   5.079 6.44e-06 ***
quality        23.117      4.224   5.473 1.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.8 on 47 degrees of freedom
Multiple R-squared:  0.3892,   Adjusted R-squared:  0.3762
F-statistic: 29.95 on 1 and 47 DF,  p-value: 1.674e-06
```

Goodness of Fit

HYPOTHESIS TESTING: BOOK SALES

- The standard errors, t-statistics and p-values are provided by the software. The two tests here are:

$$H_0: \beta_0 = 0 \qquad H_0: \beta_1 = 0$$

$$H_A: \beta_0 \neq 0 \qquad H_A: \beta_1 \neq 0$$

- Can we say that sales increase with ratings? **Yes, because the p-value for the slope is small, so the slope is different from 0**
  - p-value (Type–I error) represents the **chance of being incorrect** in saying that the estimated coefficient β_1 is different from 0

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.630     14.694   5.079 6.44e-06 ***
quality       23.117      4.224   5.473 1.67e-06 ***
```

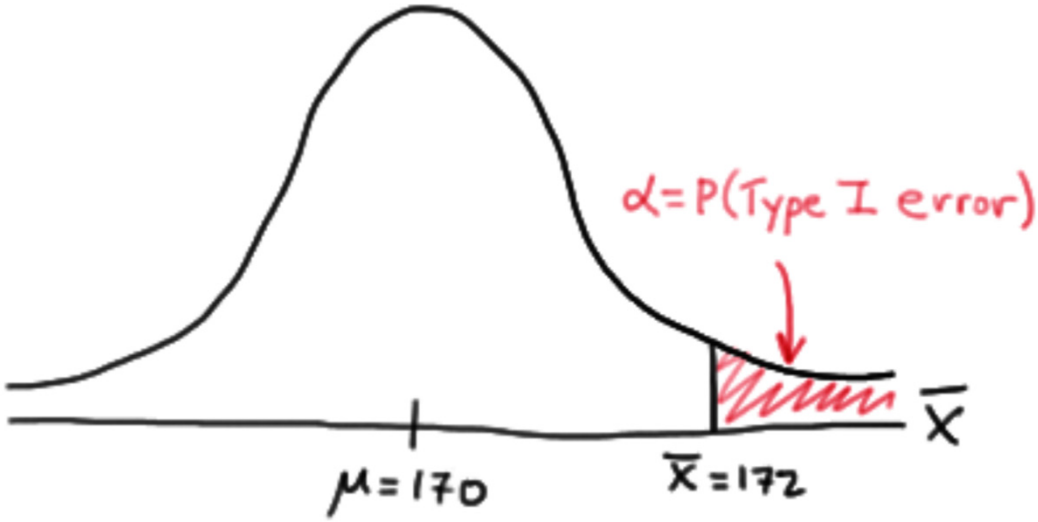
TYPE-I AND TYPE-II ERRORS

- A **Type I error** occurs if we reject the null hypothesis H_0 (in favor of the alternative hypothesis H_A) when the null hypothesis H_0 is true. We denote $\alpha = P(\text{Type I Error})$.
- A **Type II error** occurs if we fail to reject the null hypothesis H_0 when the alternative hypothesis H_A is true. We denote $\beta = P(\text{Type II Error})$.

Table 2.1 The Possibilities of Error in Statistical Significance Testing of Treatment (T) Versus Control (C) Group Differences

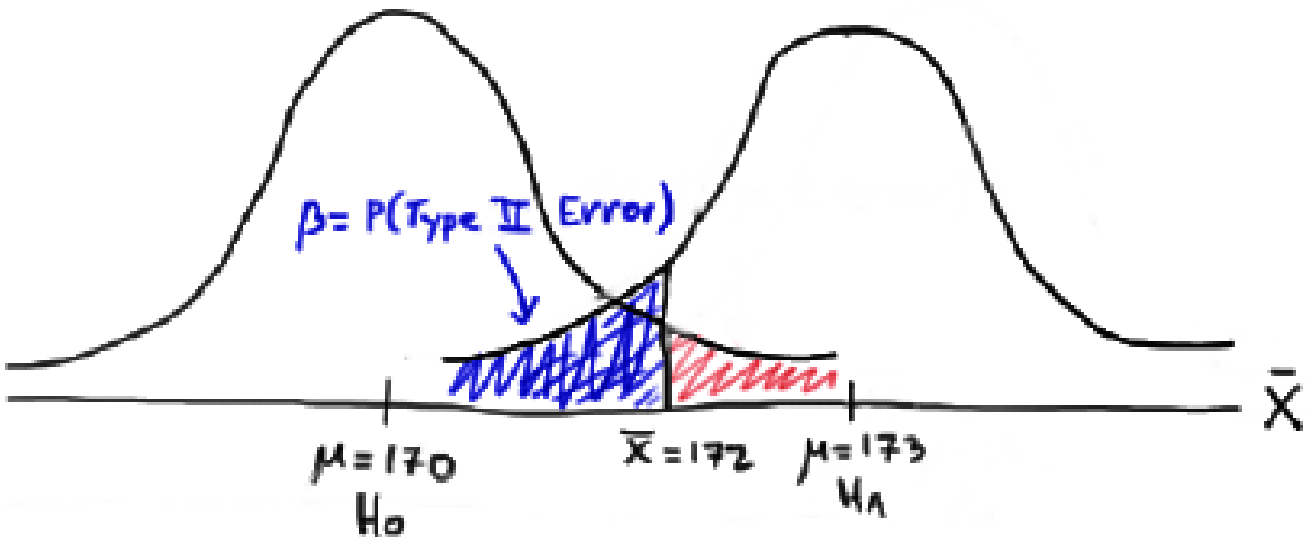
Conclusion From Statistical Test on Sample Data	*Population Circumstances*	
	T and C Differ (H_A)	*T and C Do Not Differ (H_0)*
Significant difference (reject null hypothesis)	Correct conclusion Probability = 1 – β (power)	Type I error Probability = α
No significant difference (fail to reject null hypothesis)	Type II error Probability = β	Correct conclusion Probability = 1 – α

TYPE-I ERROR (REJECT NULL HYPOTHESIS WHEN NULL IS TRUE)



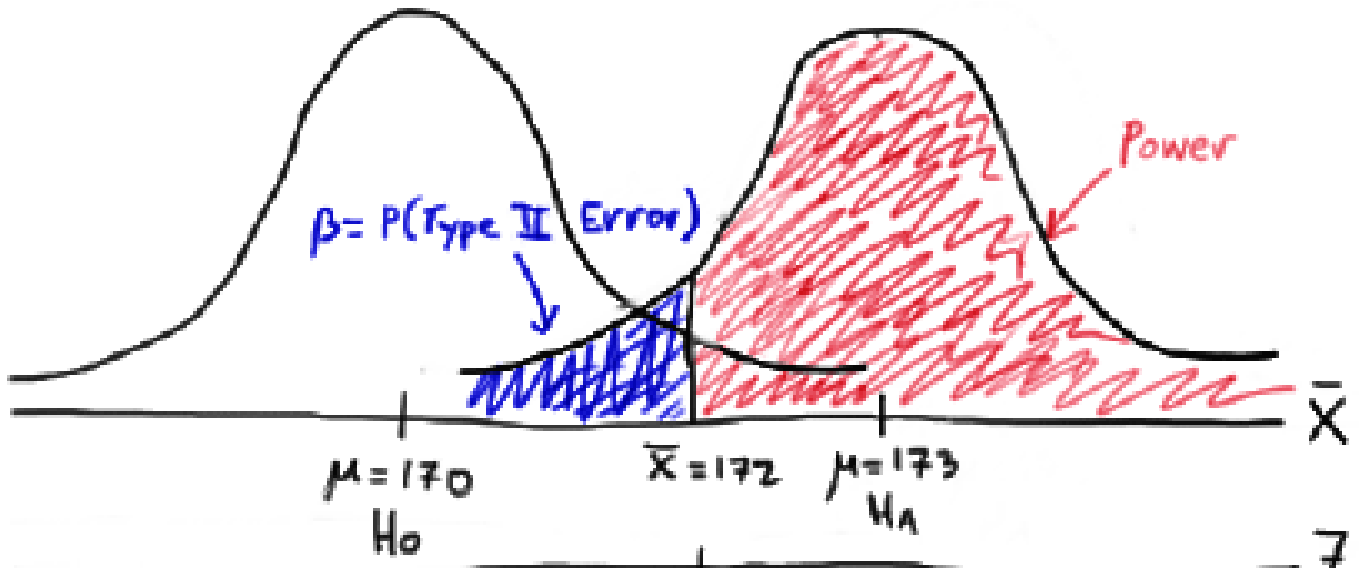
The null hypothesis H_0: $\mu = 170$
The alternative hypothesis H_A: $\mu > 170$

TYPE-II ERROR (FAIL TO REJECT NULL HYPOTHESIS WHEN ALTERNATIVE IS TRUE)



Suppose the true value of $\mu = 173$

POWER OF TEST (REJECT NULL HYPOTHESIS WHEN ALTERNATIVE IS TRUE)



The **power** of a hypothesis test = 1 - β. It measures the probability of rejecting the null hypothesis when the alternative hypothesis is true.

IN ONLINE ADVERTISING CONTEXT

- What is type I error?
  - The chance that advertising is ineffective, but the test suggests it is effective
- What's type II error?
  - The chance that advertising is effective, but the test fails to catch it
- What's power of test?
  - The chance that advertising is effective, and the test is able to catch it (i.e. you reject H_0 when H_0 is false)

OUR GOAL

- Minimize the probability of committing a Type I error.
  - i.e., minimize α = P(Type I Error).
  - Typically, a significance level of $\alpha \leq 0.05$ is desired.

- Maximize the power (at a value of the parameter under the alternative hypothesis that is scientifically meaningful).
  - Typically, we desire power to be 0.80 or greater (equivalent to a type II error rate of 0.20 or less).
  - As the actual mean μ moves further away from the value of the mean under the null hypothesis, the power of the hypothesis test increases.

- Given sample size n, a decrease in α causes an increase in β. To decrease α and β simultaneously, increase the sample size n.

II. OTHER VARIATIONS OF REGRESSIONS

OTHER VARIATIONS OF REGRESSIONS

- Regressions can have more than one explanatory variables (X's)
- Explanatory variables can be categorical variables rather than numerical variables

- Example: salary discrimination
  - Data on a random sample of 208 employees at a bank – data includes info on gender and salary
  - Is there gender based salary discrimination at the bank, i.e., are male and female employees paid differently?
  - How can we test for discrimination?

CATEGORICAL EXPLANATORY VARIABLES

- Gender is a categorical variable, not a numerical variable
- We can use dummy (or binary) variables
  - Can take one of only two values, coded as 0 and 1
  - For two categories, we create one dummy variable
  - If there are more than two categories, we use multiple dummies
    - n categories need n-1 dummies

- Consider the model: $\text{Salary} = \beta_0 + \beta_1 \text{GenderDum} + \varepsilon$
  - GenderDum = 1 for women employees
  - GenderDum = 0 for men employees

- Examples in online advertising context
  - New vs. existing customers
  - Customers exposed to ads vs. not exposed to ads

REGRESSION WITH GENDER DUMMY

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	45.505441	1.28353	35.45	<.0001*
GenderDum	-8.295513	1.564493	-5.30	<.0001*

Estimated Salary = 45.5 – 8.3 GenderDum

Is there evidence for discrimination?

ADDING EXPERIENCE VARIABLE

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	35.82379	1.259106	28.45	<.0001*
GenderDum	-8.011886	1.193089	-6.72	<.0001*
Experience	0.9811509	0.080285	12.22	<.0001*

Estimated Salary = 35.8 – 8.0 GenderDum + 0.98 Experience

How do we interpret the result now?
What can we conclude?

INTERACTION BETWEEN EXPERIENCE AND GENDER

- Does salary increase at different rates with experience for males and females?
- Create a new variable GenderDum*Experience as the product of the GenderDum and Experience variables
- This variable, which is a product of two predictor variables, is called an interaction term
  - $\text{GenderDum*Experience} = \begin{cases} \text{Experience} & \text{for female employees} \\ 0 & \text{for male employees} \end{cases}$
- Consider the model:

$$\text{Salary} = \beta_0 + \beta_1 \text{ GenderDum} + \beta_2 \text{ Experience} + \beta_3 \text{ GenderDum*Experience} + \varepsilon$$

REGRESSION WITH INTERACTION TERM ADDED

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	30.430028	1.216574	25.01	<.0001*
GenderDum	4.0982519	1.665842	2.46	0.0147*
Experience	1.5277617	0.09046	16.89	<.0001*
GenderDum*Exp	-1.247798	0.136676	-9.13	<.0001*

Estimated Salary = 30.4 + 4.1 GenderDum + 1.5 Experience – 1.2 GenderDum * Exp

How do we interpret the result now?
What can we conclude?

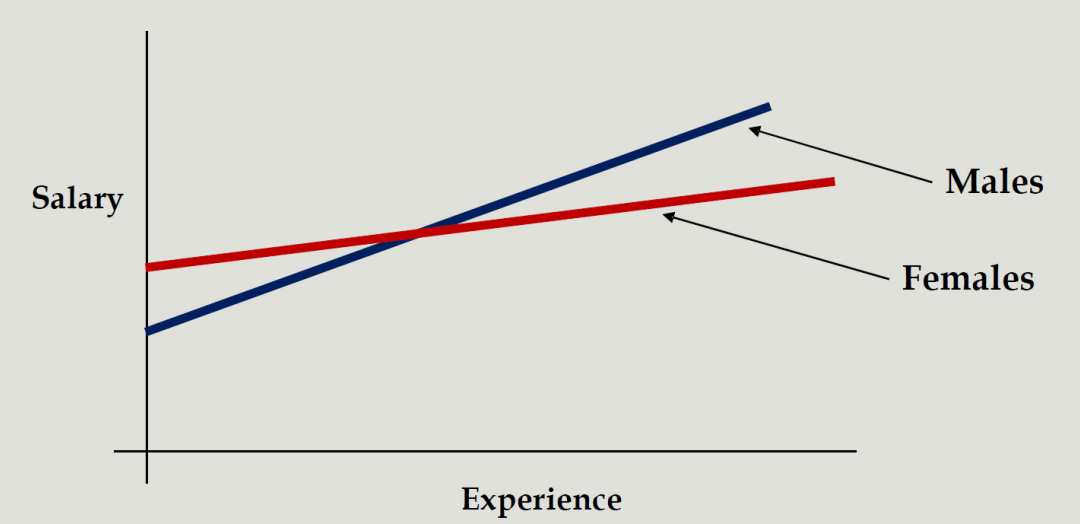
HETEROGENEOUS EFFECT

$$\text{Salary} = \beta_0 + \beta_1\ \text{GenderDum} + \beta_2\ \text{Experience} + \beta_3\ \text{GenderDum*Experience} + \varepsilon$$

- $\text{Salary} = \begin{cases} \beta_0 + \beta_1 + (\beta_2 + \beta_3)\text{Experience} & \text{for female employees} \\ \beta_0 + \beta_2\ \text{Experience} & \text{for male employees} \end{cases}$
- β_1 measures the difference in based salary between female and male employees
- β_3 measures the difference in experience effect between female and male employees

Estimated Salary = 30.4 + 4.1 GenderDum + 1.5 Experience – 1.2 GenderDum * Experience

- $\text{Estimated Salary} = \begin{cases} 34.5 + 0.3\ \text{Experience} & \text{for female employees} \\ 30.4 + 1.5\ \text{Experience} & \text{for male employees} \end{cases}$



WHAT ELSE SHOULD WE CONTROL FOR?

- The current job grade of the employee is also likely to have a significant effect on salary
- JobGrade is a categorical variable that can take values from 1 to 6 (6 being the highest)
- To include JobGrade in the regression, we need to create five dummy variables: JobGrade2, JobGrade3, JobGrade4, JobGrade5, JobGrade6.
- In general, if there are m categories for the variable, we need to create (m-1) dummy variables.
- The category for which dummy is not created (in this case, JobGrade=1) serves as the reference category

CREATING DUMMIES FOR JOBGRADE

Actual JobGrade	Dummy Variables				
	JobGrade2	JobGrade3	JobGrade4	JobGrade5	JobGrade6
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1

$$Salary = \alpha + \beta_1 GenderDum + \beta_2 Experience + \beta_3 GenderDum*Experience + \beta_4 JobGrade2 + \beta_5 JobGrade3 + \beta_6 JobGrade4 + \beta_7 JobGrade5 + \beta_8 JobGrade6 + \varepsilon$$

REGRESSION WITH JOBGRADE DUMMIES ADDED

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.104223	1.105443	23.61	<.0001*
GenderDum	6.0633281	1.266322	4.79	<.0001*
Experience	1.070883	0.102013	10.50	<.0001*
GenderDum*Exp	-1.021051	0.118726	-8.60	<.0001*
JobGrade2	2.5964926	1.010122	2.57	0.0109*
JobGrade3	6.2213937	0.998177	6.23	<.0001*
JobGrade4	11.071954	1.172588	9.44	<.0001*
JobGrade5	14.946576	1.340249	11.15	<.0001*
JobGrade6	17.097372	2.390671	7.15	<.0001*

MORE ON INTERACTION TERMS

- Can we have a regression with all dummy variables?
  - Yes
- Interaction terms
  - Can be a product of two dummy variables
  - A dummy and a continuous variable
  - Two continuous variables
- Is it possible that the interaction is significant but one or both of the main variables it depends on are not?
  - Yes
  - The interaction terms and the main terms capture different effects

III. CORRELATION VS. CAUSALITY

III.1 OMITTED VARIABLE BIAS

DO YOU LIKE ICE CREAM?

When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?

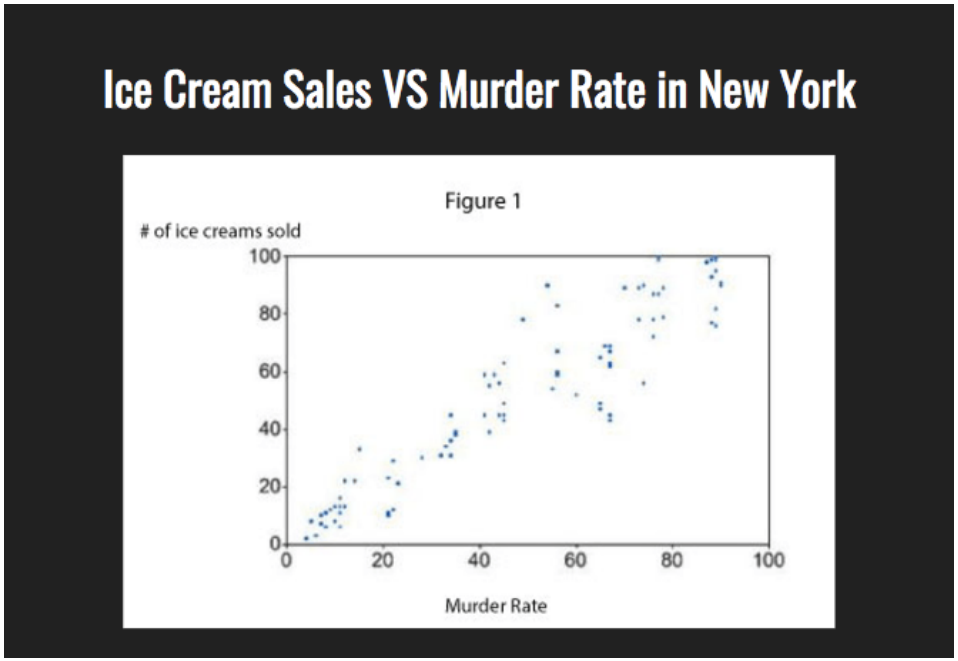
By JUSTIN PETERS JULY 09, 2013 • 2:59 PM

Selling a boy an ice cream cone, or a murder magnet?

Photo by Andrew Burton/Getty Images

*Crime is **Slate**'s crime blog. Like us on Facebook, and follow us on Twitter @slatecrime.*

The New Orleans *Times-Picayune* ran a piece last Friday attempting to answer a question the entire world has been asking: Should ice cream be blamed for murders? "The correlation between homicides and ice cream sales—when ice cream sales increase, the rate of homicides also increases—has long been a topic in statistics and science classrooms," writes John Harper, citing several recent cases of ice cream-related crime.



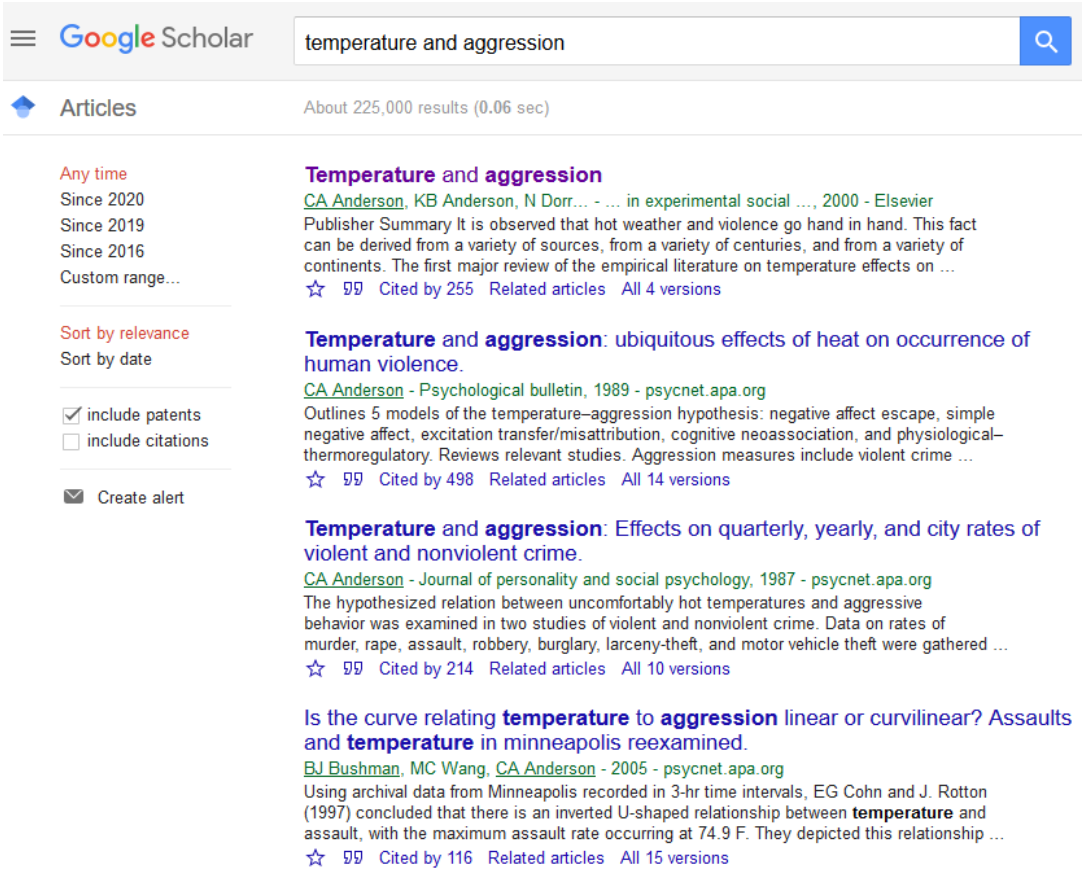
Ice Cream Sales Lead to Higher Homicide Rates: How Correlation Doesn't Always Equal Causation

A topic discussed in classrooms for years has been the strong positive correlation between ice cream sales and homicide rates. When ice cream sales rise, so does homicide. So does this mean ice cream causes us to commit violent crime? Of course not.

There is plenty of evidence to show that rising temperature causes an increase in crime,* most likely due to the fact that we're more likely to lose our temper when trapped in hot weather. We're also more also likely to gather and drink in the summer after work, filling the streets with more opportunity for crime in the evenings. Ice cream is always more appealing when it's warm outside, so it stands to reason that sales are understandably higher in summer months. Both ice cream sales and crime rates soar in the summer, but the two variables are completely unrelated to one another.

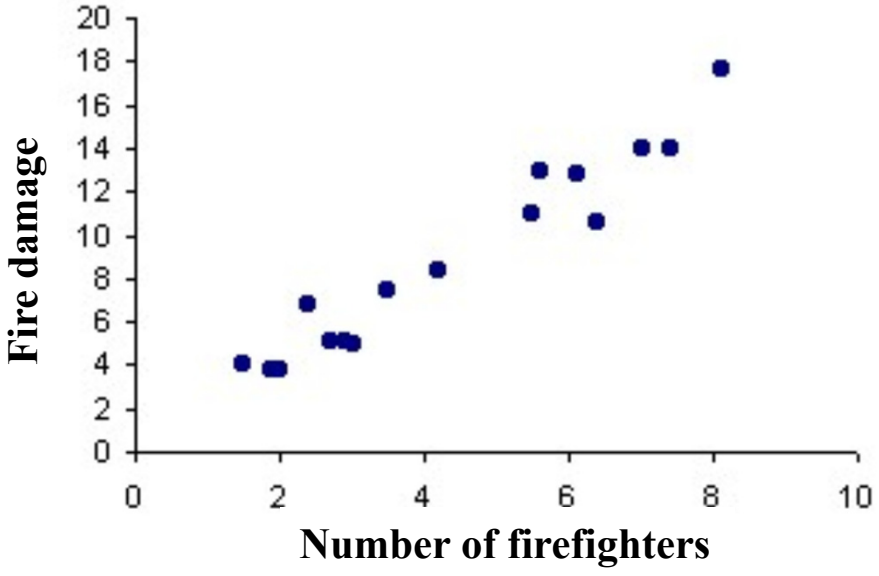
What does this data mean for your firm?

The ice cream/homicide fallacy is discussed in classrooms for a good reason: It proves correlation does not necessarily equal causation. When two or more variables correlate, you may be tempted to conclude that the correlation is due to a relationship between the variables. In this instance, there's no connection between ice cream and homicide at all: the causation is actually between temperature and ice cream, and temperature and homicide.



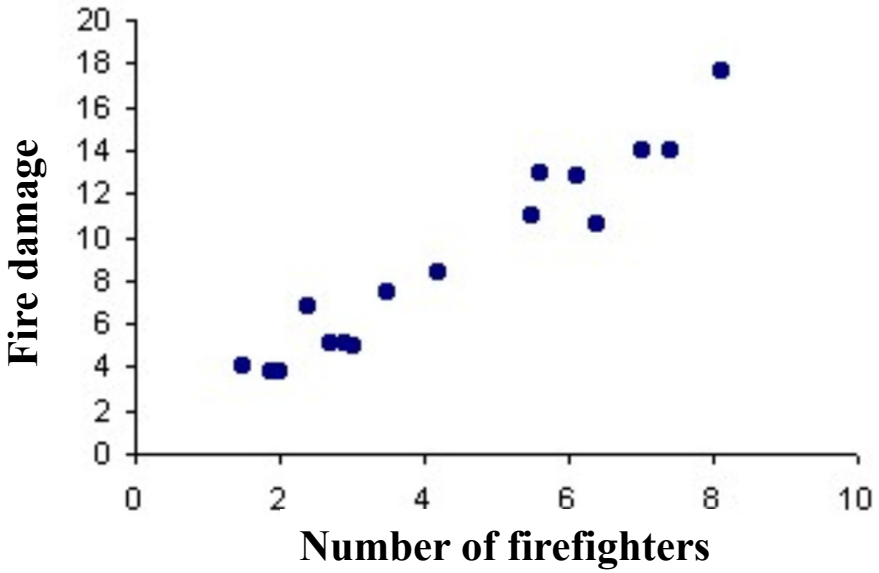
Omitted factor: temperature/heat

WHAT ABOUT THIS PATTERN?



Fire damage = 5 + 2 * Number of firefighters

CORRELATION DOES NOT EQUAL CAUSATION

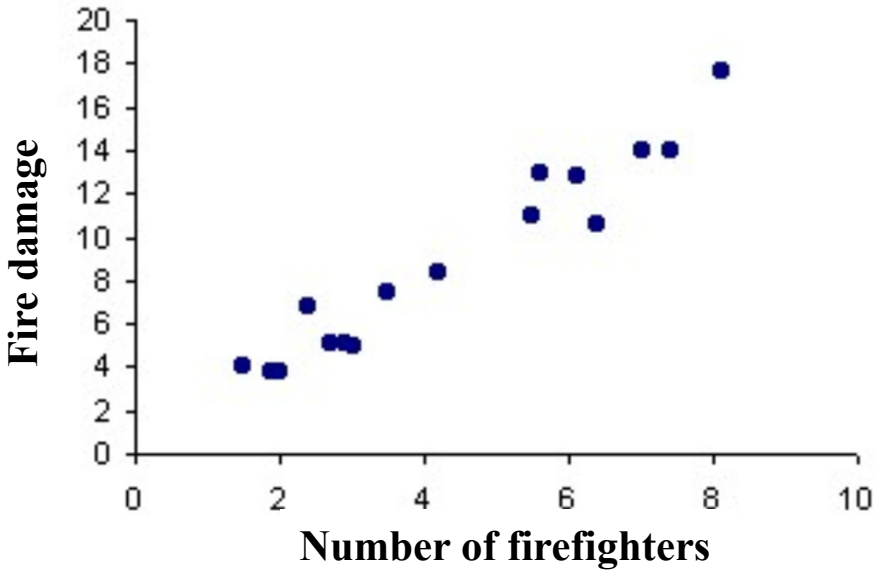


Fire damage = 5 + 2 * Number of firefighters

- Significance of regression estimates shows correlation, but correlation does not equal causation
- The significant coefficient above can be driven by a missing variable that is associated (correlated) with both fire damage and # of firefighters

(Problem will not arise if missing variable is not correlated with # of firefighters)

ADD THE MISSING VARIABLE



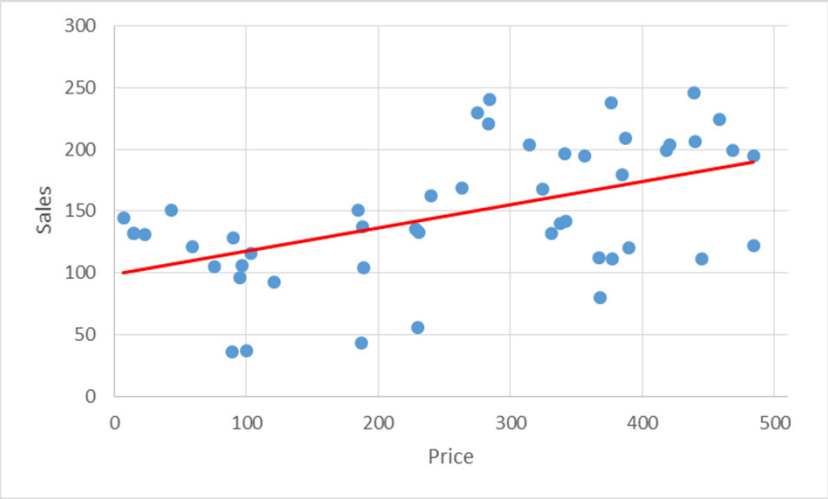
Fire damage = 5 + 2 * Number of firefighters

Fire damage = 5 - 3 * Number of firefighters + 6 * Size of fire

IN CLASS EXERCISE: BOOK SALES

- A regression analysis was conducted to understand how price impacts sales for books
- What do you think of the results?

Sales = 98.70 + 0.19 * Price




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.69775   13.77521   7.165  4.6e-09 ***
price        0.18825    0.04614   4.080 0.000173 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.07 on 47 degrees of freedom
Multiple R-squared:  0.2616,   Adjusted R-squared:  0.2459
F-statistic: 16.65 on 1 and 47 DF,  p-value: 0.0001729
```

WHAT IF WE ADD ANOTHER FACTOR: QUALITY

Sales = 49.28 – 0.53 * Price + 75.06 * Quality

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.2805    14.9069   3.306 0.001841 **
price        -0.5333     0.1486  -3.590 0.000801 ***
quality      75.0611    14.9551   5.019 8.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.24 on 46 degrees of freedom
Multiple R-squared:  0.5229,  Adjusted R-squared:  0.5021
F-statistic:  25.2 on 2 and 46 DF,  p-value: 4.061e-08
```

HOW TO INTERPRET THE COEFFICIENTS?

Sales = 49.28 – 0.53 * Price + 75.06 * Quality

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.2805    14.9069   3.306 0.001841 **
price        -0.5333     0.1486  -3.590 0.000801 ***
quality      75.0611    14.9551   5.019 8.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.24 on 46 degrees of freedom
Multiple R-squared:  0.5229,  Adjusted R-squared:  0.5021
F-statistic:  25.2 on 2 and 46 DF,  p-value: 4.061e-08
```

- Interpretation of the coefficients in multiple regression (regression with more than one independent variables):
  - Price: For books with the same quality, a price that is higher by $100, results in sales that are higher by 53, on average.
  - Rating: For books with the same price, a higher quality (specifically, a rating that is higher by 1 star) results in an increase of sales by 75, on average
- Can we say quality has more impact than price?

COMPARISON OF THE SIMPLE AND MULTIPLE REGRESSION ESTIMATES

Sales = 98.70 + 0.19 * Price

Sales = 49.28 – 0.53 * Price + 75.06 * Quality

- In this example, quality (measured by Rating) drives both Price and Sales, specifically:

  Price = – 66 + 100 Rating

- Accordingly, we can see where the positive coefficient of price comes from:

OMITTED VARIABLE BIAS

- Omitting "Quality" in the regression of "Sales" and "Price" introduces a bias
- This is because the omitted variable is correlated with the included independent variable and *also* with the dependent variable
- In other words, in the regression of "Sales" and "Price", the error term is correlated with the independent variable
- Thus, this omission violates the OLS assumption $E[\varepsilon|X] = 0$, or $Cov(\varepsilon,X) = 0$

SOME OMISSION IS OK

- Add Channels to the regression: convert channel to two dummies
  - $\text{Dummy1} = \begin{cases} 1 & \text{if channel} = 2 \\ 0 & \text{otherwise} \end{cases}$
  - $\text{Dummy2} = \begin{cases} 1 & \text{if channel} = 3 \\ 0 & \text{otherwise} \end{cases}$

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          69.9662    13.4207   5.213 4.75e-06 ***
price                -0.5694     0.1243  -4.580 3.81e-05 ***
quality              79.0015    12.5617   6.289 1.27e-07 ***
as.factor(channel)2 -44.3656    10.2172  -4.342 8.16e-05 ***
as.factor(channel)3 -40.5213    12.7136  -3.187  0.00264 **
```

  - Channels 2 and 3 have lower sales than Channel 1 (the reference channel)
  - Although channels do show an effect on sales, they do not influence the estimation of the price effect or quality effect as long as they are not correlated with price or quality

LINEAR HYPOTHESIS TESTING

– Channel 2 and 3 are no different from each other

```
Linear hypothesis test

Hypothesis:
as.factor(channel)2 - as.factor(channel)3 = 0

Model 1: restricted model
Model 2: sales ~ price + quality + as.factor(channel)

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     45 44871
2     44 44786  1    85.296 0.0838 0.7736
```

HETEROGENEOUS EFFECT

- Although omitting channels does not affect estimation of the price effect and the quality effect, it may lead to overlook of interesting heterogeneous effects

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 59.9815    17.1444   3.499 0.001162 **
price                       -0.3966     0.1817  -2.183 0.034972 *
quality                     67.8438    18.0837   3.752 0.000557 ***
as.factor(channel)2        -10.6738    28.3786  -0.376 0.708815
as.factor(channel)3        -53.7848    32.6396  -1.648 0.107220
price:as.factor(channel)2   -0.1500     0.2578  -0.582 0.563978
price:as.factor(channel)3   -0.7654     0.3782  -2.024 0.049685 *
quality:as.factor(channel)2  2.3958    26.0093   0.092 0.927068
quality:as.factor(channel)3 67.4142    38.6090   1.746 0.088475 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HETEROGENEOUS EFFECT

- The regression specification:

$$\text{Sales} = a + b\,\text{Price} + c\,\text{Quality} + d_1\,\text{Channel2} + d_2\,\text{Channel3} + e_1\,\text{Price} * \text{Channel2} + e_2\,\text{Price} * \text{Channel3} + f_1\,\text{Quality} * \text{Channel2} + f_2\,\text{Quality} * \text{Channel3}$$

  - $\text{Sales} = \begin{cases} a + b\,\text{Price} & \text{Channel 1} \\ a + d_1 + (b + e_1)\,\text{Price} & \text{Channel 2} \\ a + d_2 + (b + e_2)\,\text{Price} & \text{Channel 3} \end{cases}$
  - $\text{Price effect} \begin{cases} b & \text{Channel 1} \\ b + e_1 & \text{Channel 2} \\ b + e_2 & \text{Channel 3} \end{cases}$
    - e1 measures the difference in price effects between Channels 1 and 2
    - e2 measures the difference in price effects between Channels 1 and 3

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                59.9815    17.1444   3.499 0.001162 **
price                      -0.3966     0.1817  -2.183 0.034972 *
quality                    67.8438    18.0837   3.752 0.000557 ***
as.factor(channel)2       -10.6738    28.3786  -0.376 0.708815
as.factor(channel)3       -53.7848    32.6396  -1.648 0.107220
price:as.factor(channel)2  -0.1500     0.2578  -0.582 0.563978
price:as.factor(channel)3  -0.7654     0.3782  -2.024 0.049685 *
quality:as.factor(channel)2  2.3958    26.0093   0.092 0.927068
quality:as.factor(channel)3 67.4142    38.6090   1.746 0.088475 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OMITTED VARIABLE BIAS

- When this bias is present, the causal interpretation of the regression coefficients can be incorrect!
- For example, what can be wrong with this statement?
  - Sellers who advertise on this platform receive twice as many sales as those who do not advertise
- But, we want to make causal interpretation, e.g., how much does an extra display ad impression improve sales. What should we do?

HOW TO DEAL WITH OMITTED VARIABLE BIAS

- Include the missing variable in the regression if we know what is missing and have data for it
- Can rely on domain knowledge, or use statistical methods like instrument variables.

- But both approaches have their limitations.

- **Randomized field experiments** can remove the effects of missing variables through randomization

III.2 REVERSE CAUSALITY

Feeling the Pressure to Drink for Work

By DOUGLAS QUENQUA

Béatrice de Géa for The New York Times

COLD SHOULDER: Terry Lavin sipping seltzer water at the Pig 'n' Whistle in Manhattan.

As an ad-sales executive with Forbes magazine, Terry Lavin worked hard to earn his reputation as a dependable drinking buddy.

"I just basically rented space at P. J. Clarke's," he said, referring to the Midtown Manhattan watering hole. "I was always the last to leave, always had a cocktail in my hand."

In a business built on likability, the role helped him succeed. Until 2010, when he decided to give his body a break and quit drinking for six months. His health got better; his business did not.

WHICH ONE IS TRUE?

- Drinkers can build up more social capital, making them more successful at work?

or

- People who earn more money drink more either because they have a greater disposable income or due to stress?

Reverse causality or simultaneity effect

Journal of Labor Research
December 2006, Volume 27, Issue 3, pp 411–421 | Cite as

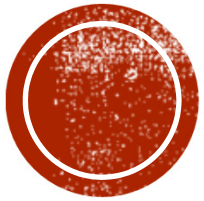
No booze? You may lose: Why drinkers earn more money than nondrinkers

Authors | Authors and affiliations

Bethany L. Peters, Edward Stringham

Abstract

A number of theorists assume that drinking has harmful economic effects, but data show that drinking and earnings are positively correlated. We hypothesize that drinking leads to higher earnings by increasing social capital. If drinkers have larger social networks, their earnings should increase. Examining the Gen-eral Social Survey, we find that selfreported drinkers earn 1014 percent more than abstainers, which replicates results from other data sets. We then attempt to differentiate between social and nonsocial drinking by comparing the earn-ings of those who frequent bars at least once per month and those who do not. We find that males who frequent bars at least once per month earn an addi-tional 7 percent on top of the 10 percent drinkers' premium. These results sug-gest that social drinking leads to increased social capital.



III.3 SELECTION BIAS

Hormone replacement therapy and your heart

Are you taking — or considering — hormone therapy to treat bothersome menopausal symptoms? Understand potential risks to your heart and whether hormone therapy is right for you.

By Mayo Clinic staff

Long-term hormone replacement therapy used to be routinely prescribed for postmenopausal women to relieve hot flashes and other menopause symptoms. Hormone replacement therapy was also thought to reduce the risk of heart disease.

Before menopause, women have a lower risk of heart disease than men do. But as women age, and their estrogen levels decline after menopause, their risk of heart disease increases. In the 1980s and 1990s, experts advised older women to take estrogen and other hormones to keep their hearts healthy.

However, hormone replacement therapy — or hormone therapy, as it's now called — has had mixed results. Many of the hoped-for benefits failed to materialize for large numbers of women. The largest randomized, controlled trial to date actually found an insignificant increase in heart disease in postmenopausal women using hormone therapy.

Still, some data suggest that estrogen may decrease the risk of heart disease when taken early in postmenopausal years:

- In a recent Danish study, after 10 years of treatment, women receiving hormone replacement therapy early after menopause had a significantly reduced risk of mortality, heart failure or heart attack, without any apparent increase in risk of cancer or stroke.
- A randomized, controlled clinical trial — the Kronos Early Estrogen Prevention Study (KEEPS) — exploring estrogen use and heart disease in younger postmenopausal women was recently completed, with results expected soon.

ARE THEY REPRESENTATIVE?

- Re-analysis of the data from the epidemiological studies showed that women undertaking Hormone Replacement Therapy were more likely to be from higher-socio-economic groups, with better-than-average diet and exercise regimens

Sample Selection Bias

International Journal of Epidemiology

IEA International Epidemiological Association

Article Navigation

Commentary: The hormone replacement–coronary heart disease conundrum: is this the death of observational epidemiology?

Debbie A Lawlor, George Davey Smith, Shah Ebrahim

International Journal of Epidemiology, Volume 33, Issue 3, June 2004, Pages 464–467, https://doi.org/10.1093/ije/dyh124

Published: 27 May 2004

PDF Split View Cite Permissions Share ▾

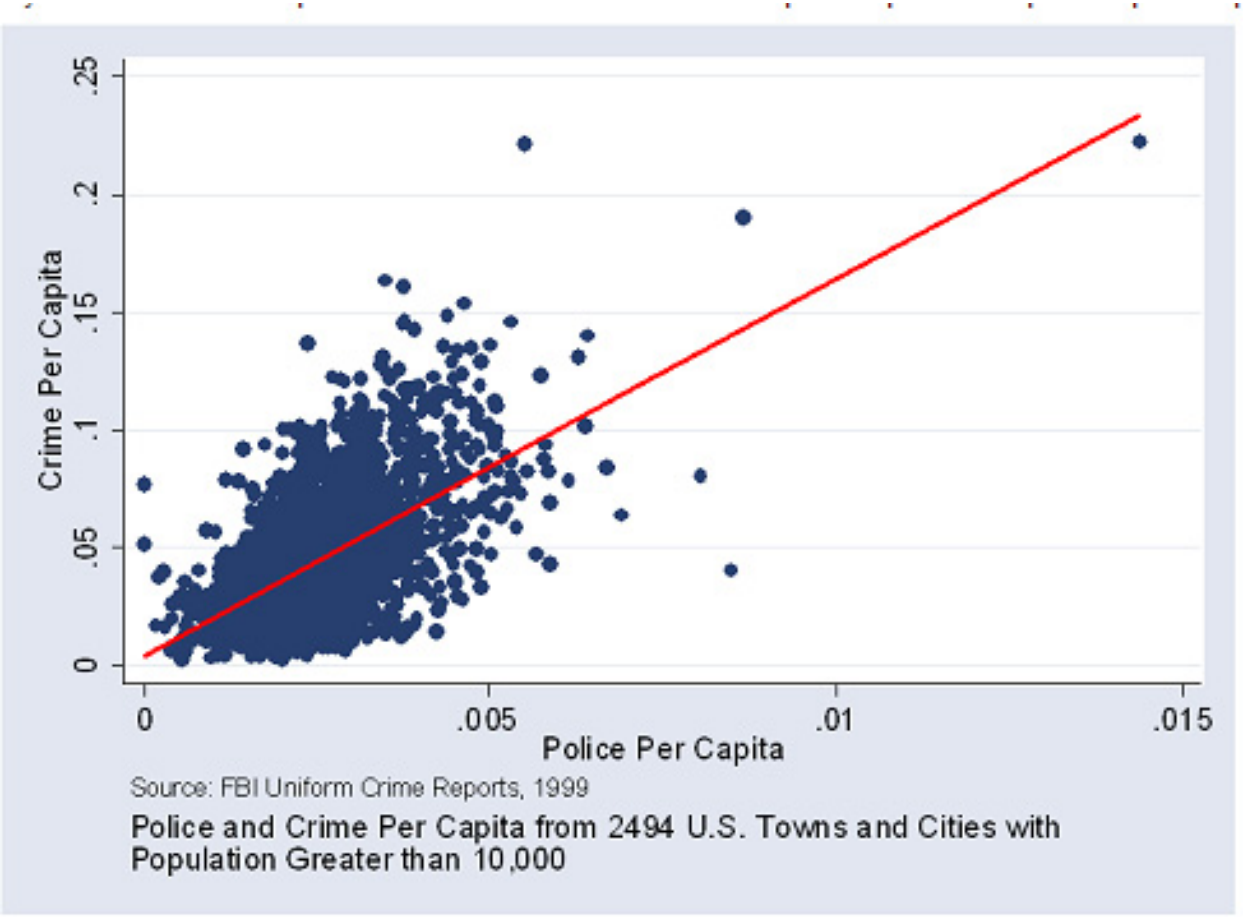
Topic: hormone replacement therapy, epidemiology, coronary heart disease

Issue Section: Reprints and Reflections

Under its definition for the word 'hindsight' the *Oxford English Dictionary* includes the following statement 'hindsight is always better than foresight' (http://dictionary.oed.com/), and the slogan of a private survey and evaluation company, ingeniously called Hindsight, is 'remember hindsight is always 20/20!' (http://www.hndsight.com/). We have the benefit of the 'hindsight' from randomized controlled trials (RCT) when we comment on this meta-analysis of observational studies, but whether the conflicting results between the trial and observational evidence on the association between hormone replacement therapy (HRT) use and coronary heart disease (CHD) will lead to 20/20 vision remains to be seen.

The disparity between findings from observational studies and RCT of the effects of HRT on CHD,[1–4] has created considerable debate among researchers, practitioners and postmenopausal women. The authors of the meta-analysis reprinted in this issue of the *International Journal of Epidemiology* concluded that the pooled estimate of effect from the best quality observational studies (internally controlled prospective and angiographic studies) inferred a relative reduction of 50% with ever use of HRT and stated that 'overall, the bulk of the evidence strongly supports a protective effect of estrogens that is unlikely to be explained by confounding factors'.[4] By contrast, recent randomized trials among both women with established CHD and healthy women have found HRT to be associated with slightly increased risk of CHD or null effects.[1,2] For

WHAT DO YOU THINK OF THIS PATTERN?



Source: FBI Uniform Crime Reports, 1999

Police and Crime Per Capita from 2494 U.S. Towns and Cities with Population Greater than 10,000

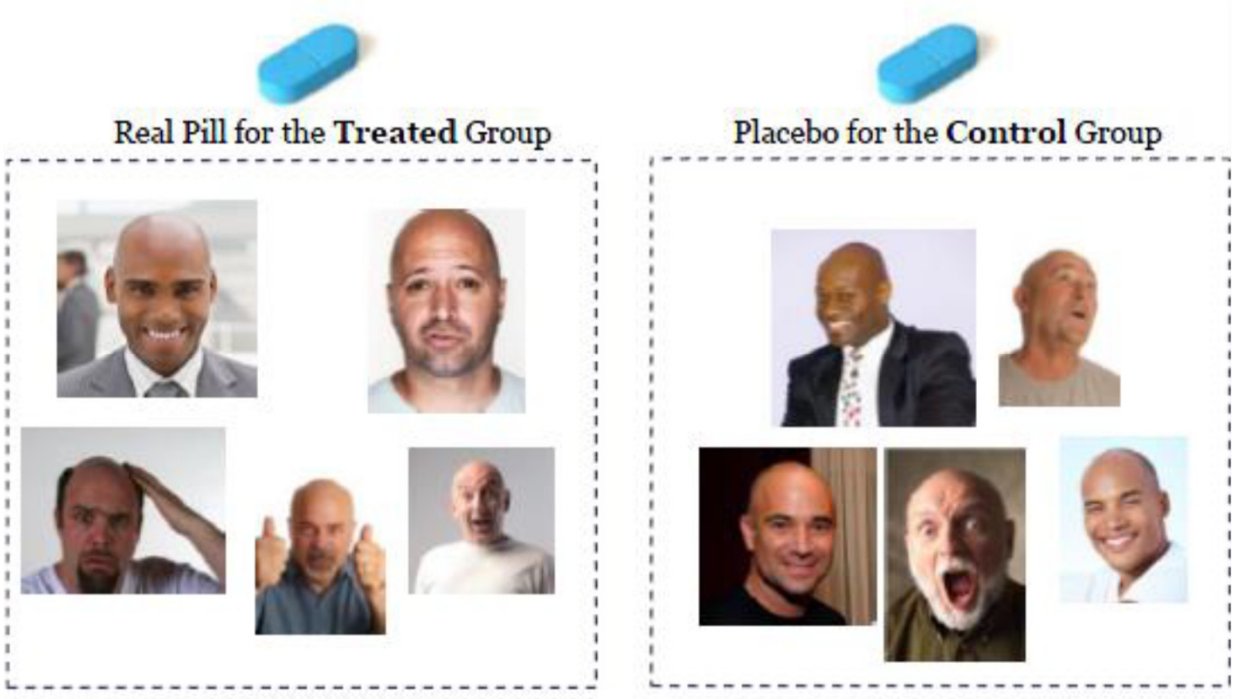
CORRELATION DOES NOT EQUAL CAUSALITY

- Omitted variable Bias
- Reverse causality
- Sample selection bias

HOW DO WE DRAW CAUSAL INFERENCE?

The Randomized Controlled Experiment

- Randomly allocating the experimental units across different treatment groups, where randomization equalizes other factors that have not been accounted for in the experiment design
  - e.g. to test if a new drug is useful in treating baldness, we randomly assign different people to either the new drug and a placebo drug



CAUSAL INFERENCE IS MADE BY COMPARING THE TREATED GROUP AND THE CONTROL GROUP

- No omitted variable bias: randomization of subjects removes effect from other factors
- No reverse causality: the effect *comes after* the treatment
- No sample selection bias: randomization takes care of this too

- When it is impossible to conduct randomized controlled experiments, economists rely on "natural experiments", which typically require an unexpected shock/change to the system, where the change is introduced randomly (exogenously) with respect to the outcome.