# Applied Methods for Ph.D.

Michael E. Kummer

2024

## Causality and Correlation

Slides based on chapters 2 and 3 of the book Mostly Harmless Econometrics and on materials by Miguel Godinho de Matos and Rodrigo Belo

# Framing the problem

# Causality and Correlation

- **Causality** is the relation between two events: a first event, the cause, and a second event, the effect. The second event is a consequence of the first event;
- In **causal** studies are interested in knowing whether X causes Y;
- **Correlation** refers to any of a broad class of statistical relationships involving dependence of events;
- **Correlation** studies are concerned with the association between a variable X an an outcome Y

# Example 1: Do hospitals make people healthier ? (1/3)

- How would you address this question ?
- A natural approach consists in comparing the health status of those who have been to the hospital to the health of those who have not. In the US, the National Health Interview Survey (NHIS) contains the information needed to make this comparison:
  - it includes a question ***During the past 12 months, was the respondent a patient in a hospital overnight?***
  - it includes a question ***Would you say your health in general is excellent (1), very good, good, fair, poor(5)?***

# Example 1: Do hospitals make people healthier ? (2/3)

- The following table displays the mean health status among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

| Group | Sample Size | Mean health | Std. Error |
|---|---|---|---|
| **Hospitalized** | 7,774 | 2.79 | 0.014 |
| **Not Hospitalized** | 90,049 | 2.07 | 0.003 |

- H0: E( Health | Hospitalized ) - E( Health | Not Hospitalized ) = 0
  - The difference in the means is 0.71 with a t-statistic of 55.8 a highly significant contrast in favor favor of the health for the non-hospitalized
- By the way … how do you get the t-stat ?

# Example 1: Do hospitals make people healthier ? (3/3)

- Taken at face value, this result suggests that going to the hospital makes people sicker
- It is not impossible this is the right answer …Hospitals are full of other sick people who might infect us, and dangerous machines and chemicals that might hurt us
- Still, it is easy to see why this comparison should not be taken at face value: people who go to the hospital are probably less healthy to begin with. Moreover, even after hospitalization people who have sought medical care are not as healthy, on average, as those who never get hospitalized in the first place, though they may well be better than they otherwise would have been
- **How do we link this problem to what we have been learning about regression ?**

# Consider the following notation for the hospital example

- Let $D_i = 1$ if individual $i$ went to the hospital and $D_i = 0$ otherwise
- Let $Y_{ji}$ be the health status of patient $i$ indexed by $j$ data determines whether he went to the hospital such that

$$Y_i = \begin{cases} Y_{0i} & \text{if } D_i = 0 \\ Y_{1i} & \text{if } D_i = 1 \end{cases}$$

- Now note that we can write $Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) \times D_i$ and don't forget it because this will be the key …
- $Y_{1i} - Y_{0i}$ is called the causal effect of going to the hospital for individual $i$

# So our goal is to know $Y_{1i} - Y_{0i}$ (1/2)

- **Problem:** we can never see both outcomes $Y_{0i}$ and $Y_{1i}$ at the same time
  - One person either goes to the hospital or it does not (does not do both at the same time)
- **Solution:** we must learn about the effects of hospitalization by comparing the average health of those who were and were not hospitalized
  - $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$
  - This is what we did before with the t-test … so where is the problem !?
- Very slowly lets use the board to replace $Y_i$ by its equivalent $Y_{0i} + (Y_{1i} - Y_{0i}) \times D_i$ in and see what we get

# So our goal is to know $Y_{1i} - Y_{0i}$ (2/2)

- $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] =$
  - Causal effect : $E[Y_{1i} - Y_{0i}|D_i = 1]$ +
  - Selection Bias: $E[Y_{0i}|D_i = 1]$ - $E[Y_{0i}|D_i = 0]$
- Our simple t-test analysis adds to the causal effect the selection bias
- If the sick are more likely than the healthy to seek treatment, those who were hospitalized haveworse $Y_{0i}$ which makes selection bias negative in this example
- The selection bias may be so large (in absolute value) that it completely masks a positive treatment effect
- **The goal of most empirical research is to overcome selection bias, and therefore to say something about the causal effect of a variable like $D_i$**

# The gold standard of social science research are randomized experiments

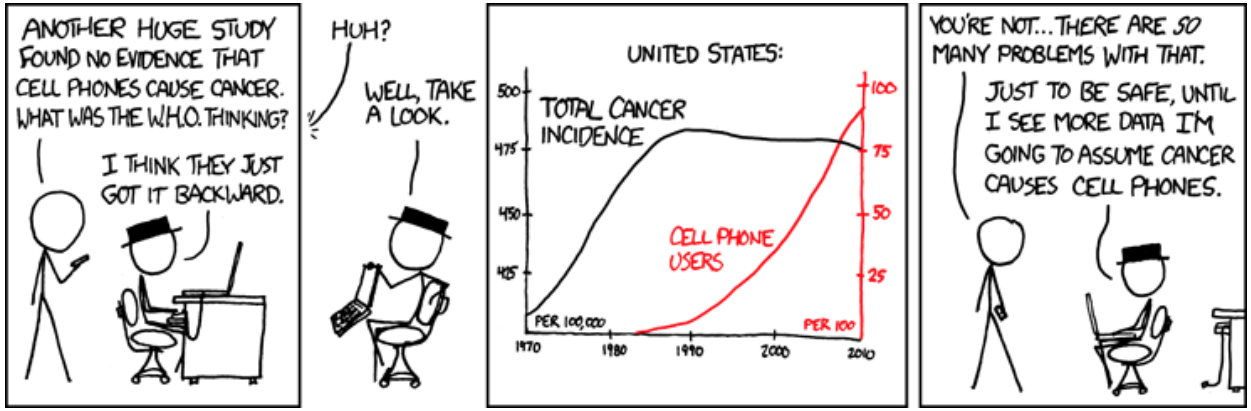
- Remember that if X and Y are independent then $E[X|Y] = E[X]$
- Random assignment of $D_i$ solves the selection problem because random assignment makes $D_i$ independent of potential outcomes:
  - $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] = E[Y_{0i}] - [Y_{0i}] = 0$ (selection bias cancels out)
- **How relevant is our hospitalization allegory?**
  - Experiments often reveal things that are not what they seem on the basis of naive comparisons alone
  - An iconic example from labor economics is the evaluation of government-subsidized training programs (novas oportunidades in Portugal for example)

# How do we link all this talk to regression ?

- Regression is a useful tool for the study of causal questions
- Suppose (for now) that the treatment effect is the same for everyone, say $Y_{1i} - Y_{0i} = \beta_1$, a constant
- Then we can re-write $Y_i$ as:
  - $Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$
  - $\beta_0 = E[Y_{0i}]$
  - $\epsilon_i = Y_{0i} - E[Y_{0i}]$
- Lets return to the board to see what this means …

# Correlation does not imply causation (not everything is what it seems)

- Shoe size is positively correlated with reading skills (why?)
- Icecream production is positively correlated with drowning events (why?)
- Fire destruction is correlated with the number of fire trucks that try to estinguish it (why)



# The Fundamentals of Regression and Causality

# Refresh your memory about the Gauss Markov Assumptions (JWB page 47)

- **SLR.1** in the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as: $y = \beta_0 + \beta_1 x + u$
- **SLR.2** we have a random sample of size n, $(x_i, y_i) : i \in 1, 2, \ldots, n$ following the population model laid out in SLR.1
- **SLR.3** The sample outcomes x ,$x_i, i \in 1, \ldots, n$ are not all the same value
- **SLR.4** The error u has an expected value of zero given any value of the explanatory variable: $E(u|x) = 0$ (conditional independence assumption – CIA)
- **SLR.5** $V(u|x) = \sigma^2$ (homoskedasticity)
- **Which assumption is being violated when we have selection bias ?**

# When does regression analaysis have a causal interpretation ?

- When the path to random assignment is blocked, we look for alternate routes to causal knowledge. The most basic of these tools is regression, which compares treatment and control subjects who have the same observed characteristics
- Regression-based causal inference is predicated on the assumption that when key observed variables have been made equal across treatment and control groups, selection bias from the things we can't see is also mostly eliminated
- Formaly, a regression is causal when the conditional expectation function (**CEF**) — $E[Y|X]$ — that it approximates is causal
- The **CEF** is causal when it describes average potential outcomes for a fixed reference population
    - Example: The effect of hospital on health in a radomized trial (as we seen before)

# When does regression analaysis have a causal interpretation ? (1/2)

- To make concrete assume that we want to find the causal connection of years of schooling (s) to earnigs as the functional relation that determines how much an individual would earn if he obtained different levels of eduction
- Start with model $f_i(s) = \alpha + \beta s + \epsilon_i$ (we assume this to be a causal existing relation)
- s does not have a subscript bcause $f_i()$ tells us what an individual would earn for evey level of $s$ not just observed the $S_i$
- When using regression we substitue s by its oberserved value $S_i$:
  - $Y_i = \alpha + \beta S_i + \epsilon_i$
- Because $S_i$ was not randomly determined – it was the consequence of personal choice – it is possible that it is related to variables in $\epsilon_i$. In that case **SLR.4** will not hold

# When does regression analaysis have a causal interpretation ? (2/2)

- We may be able to solve this if we can decompose $\epsilon_i$ in the variables that make $S_i$ as good as randomly assigned:
  - $\epsilon_i = X_i\gamma + v_i$ ($X_i$ and $\gamma$ are vectors of variables and parameters respectively)
- By construction $E[\epsilon_i|X_i] = X_i\gamma$
- Then $E(f_i(s)|X_i) = \alpha + \beta s + X_i\gamma$
- And $Y_i = \alpha + \beta S_i + X_i\gamma + v_i$ has causal interpretation
- **The key assumption is that $X_i$ are the only reason why $\epsilon_i$ and $S_i$ are correlated (selection on observables)**

# Quantifying the problem of omitted variables

- Option 1:
  - **True** population model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
  - Model estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- Option 2:
  - **Wrong** population model: $y = \beta_0 + \beta_1 x_1 + \epsilon$
  - Model estimate: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$
- In general $\hat{\beta}_1 \neq \tilde{\beta}_1$ ?
- Fortunately turns out there is a simple relationship between $\hat{\beta}_1$ and $\tilde{\beta}_1$ which allows understanding the potential direction of our bias

# Direction of the bias on the single regressor case

- In the single regressor case we are able to know exactly the direction of the bias:
  - $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}$
  - $\tilde{\delta}$ is the slope coefficient from the simple regression of $x_2$ on $x_1$ ($x_2 = \alpha + \delta x_1 + v$)
-

| | $Corr(x_1, x_2) > 0$ | $Corr(x_1, x_2) < 0$ |
|---|---|---|
| $\beta_2 > 0$ | Positive bias | Negative bias |
| $\beta_2 < 0$ | Negative bias | Positive bias |

# When $\delta = 0$ we have $\tilde{\beta}_1 = \hat{\beta}_1$

```r
#-- Simulate x variables
ssize <- 1000
x1 <- rnorm( n = ssize , sd = 3 )
x2 <- rnorm( n = ssize , sd = 5 )
y  <- 2 + 3*x1 + 5 * x2 + rnorm(n = ssize, sd = 5)

out.y.full  <- lm( y  ~ x1 + x2)
out.y.x1.om <- lm( y  ~ x1)
out.y.x2.om <- lm( y  ~ x2 )

cor.test(x = x1, y = x2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x1 and x2
## t = -0.501, df = 998, p-value = 0.6165
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07777428  0.04618083
## sample estimates:
##         cor
## -0.01585765
```

```r
#-- output
stargazer(out.y.full, out.y.x1.om, out.y.x2.om,
  type = 'text', omit.stat = c('f','ser'), no.space=T)
```

```
##
## ==========================================
##                   Dependent variable:
##               ---------------------------
##                           y
##                  (1)       (2)       (3)
## ------------------------------------------
## x1            2.969***  2.843***
##                (0.055)   (0.257)
## x2            4.994***             4.965***
##                (0.034)              (0.068)
## Constant      1.982***  2.790***  2.124***
##                (0.167)   (0.783)   (0.331)
## ------------------------------------------
## Observations   1,000     1,000     1,000
## R2             0.960     0.109     0.841
## Adjusted R2    0.959     0.108     0.841
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

# When $\delta \neq 0$ we have $\tilde{\beta}_1 \neq \hat{\beta}_1$

```
#-- Simulate x variables
ssize <- 1000
x1 <- rnorm( n = ssize , sd = 3 )
x2 <- rnorm( n = ssize , mean = x1, sd = 5 )
y  <- 2 + 3*x1 + 5 * x2 + rnorm(n = ssize, sd = 5)

out.y.full  <- lm( y  ~ x1 + x2)
out.y.x1.om <- lm( y  ~ x1)
out.y.x2.om <- lm( y  ~ x2 )

cor.test(x = x1, y = x2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x1 and x2
## t = 19.1359, df = 998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4712417 0.5620403
## sample estimates:
##       cor
## 0.5180991
```

```
#-- Output
stargazer(out.y.full, out.y.x1.om, out.y.x2.om,
  type = 'text', omit.stat = c('f','ser'), no.space =T)
```

```
##
## ==========================================
##                   Dependent variable:
##               ---------------------------
##                           y
##                  (1)       (2)       (3)
## ------------------------------------------
## x1            3.075***  7.944***
##                (0.060)   (0.260)
## x2            4.957***             5.797***
##                (0.032)              (0.052)
## Constant      1.936***  2.719***  2.042***
##                (0.156)   (0.785)   (0.296)
## ------------------------------------------
## Observations    1,000     1,000     1,000
## R2              0.980     0.484     0.926
## Adjusted R2     0.980     0.483     0.926
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

# We can check that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}$

```
y  <- 2 + 3*x1 + 5 * x2 + rnorm(n = 1000, sd = 5)
out.y.full      <- lm( y  ~ x1 + x2)
out.y.incomp.x1 <- lm( y  ~ x1 )
out.y.incomp.x2 <- lm( y  ~ x2 )
out.x1.parti    <- lm( x1 ~ x2 )
out.x2.parti    <- lm( x2 ~ x1 )
stargazer(
  out.x1.parti,
  out.x2.parti,
  out.y.full,
  out.y.incomp.x1,
  out.y.incomp.x2,
  type = 'text', omit.stat = c('f','ser'))
```

```
##
## ===========================================================
##                            Dependent variable:
##               ---------------------------------------------
##                  x1       x2                 y
##                 (1)      (2)      (3)      (4)      (5)
## -----------------------------------------------------------
## x2           0.273***           5.006***           5.806***
##              (0.014)            (0.033)            (0.050)
##
## x1                     0.982*** 2.929*** 7.846***
##                        (0.051)  (0.062)  (0.262)
##
## Constant      0.035    0.158   1.905*** 2.696*** 2.006***
##              (0.082)  (0.155)  (0.161)  (0.793)  (0.289)
##
## -----------------------------------------------------------
## Observations  1,000    1,000    1,000    1,000    1,000
## R2            0.268    0.268    0.978    0.473    0.930
## Adjusted R2   0.268    0.268    0.978    0.472    0.930
## ===========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```
coef(out.y.full)['x1'] + coef(out.y.full)
['x2']*coef(out.x2.parti)['x1']
```

```
##       x1
## 7.846219
```

```
coef(out.y.full)['x2'] + coef(out.y.full)
['x1']*coef(out.x1.parti)['x2']
```

```
##       x2
## 5.806257
```

# As a side note, lets think about the partialling out interpretation of regression

```
#----------------------
#-- Simulate x variables
ssize <- 1000
x1 <- rnorm( n = ssize , sd = 3 )
x2 <- rnorm( n = ssize , mean = x1 ,sd = 3 )
y  <- 2 + 3*x1 + 5 * x2 + rnorm(n = ssize, sd = 5)

out.y.full      <- lm( y  ~ x1 + x2)
out.parti.x2    <- lm( x1 ~ x2 )
out.y.x1        <- lm( y  ~ residuals(out.parti.x2))
out.y.x1.om     <- lm( y  ~ x1)

#-----------------
#-- Unify Output
stargazer(
   out.y.full,
   out.parti.x2,
   out.y.x1.om,
   out.y.x1,
  type = 'text',
  omit.stat = c('f','ser'), no.space =T)
```

```
##
## ============================================================
##                                  Dependent variable:
##                          ----------------------------------
##                             y         x1             y
##                            (1)       (2)       (3)       (4)
## ------------------------------------------------------------
## x1                       2.992***             8.044***
##                          (0.077)              (0.171)
## x2                       4.971*** 0.485***
##                          (0.053)  (0.016)
## residuals(out.parti.x2)                                2.992***
##                                                        (0.420)
## Constant                 2.283***  0.016    2.784*** 3.811***
##                          (0.164)  (0.068)  (0.514)  (0.897)
## ------------------------------------------------------------
## Observations              1,000    1,000    1,000    1,000
## R2                        0.968    0.493    0.689    0.048
## Adjusted R2               0.968    0.492    0.688    0.047
## ============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

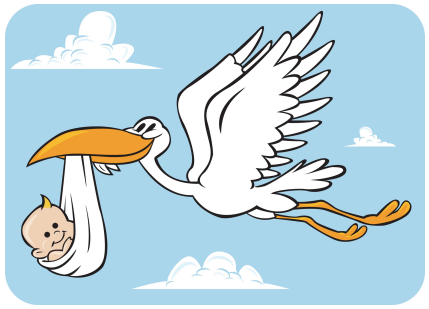
# What about when we have more than one regressors ?

- When there are multiple regressors in the estimated model things are harder
- Correlation between a single explanatory variable and the error generally results in all OLS estimators being biased
- Consider the population model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 u$ such that $x_1$ is correlated with $x_3$ and $x_2$ and $x_3$ are not correlated
- It is tempting to think that omitting $x_3$ will only bias $\beta_1$, but that is only true when $x_1$ and $x_2$ are not correlated. If they are, omitting $x_3$ will also bias the estimate for $\beta_2$.

# Examples of Omitted Variable Bias

# Example 1: Does stork population cause birth rates ? (1/4)

- **We all know how babies are made:** *"The stork can be seen flying over rooftops with a little cloth bundle before landing at the doorstep of a happy couple who then unwrap their precious, smiling newborn—right? This myth was once a common story to tell children who were deemed too young to be told anything different."*



- Storks have been associated with babies and family for centuries. Can we test it with data ?

# Example 1: Does stork population cause birth rates ? (2/4)

```
dataset <-  read.csv(
    file = '../datasets/stork_population.dsv' ,sep  =';')
names(dataset)
```

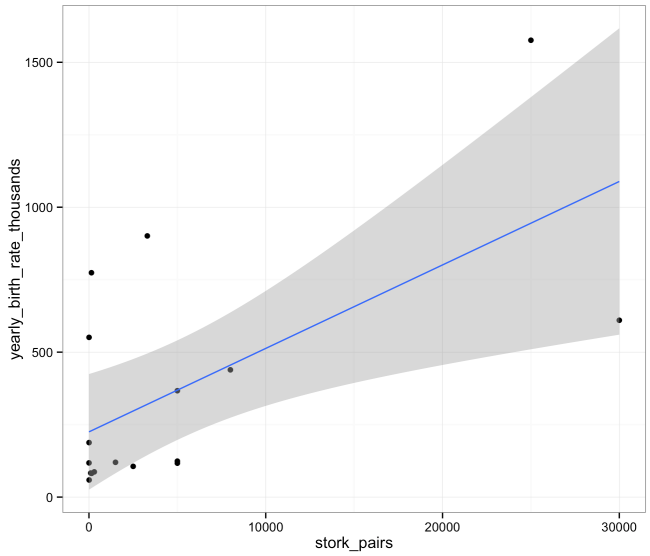
```
## [1] "country"                    "area_km2"
## [3] "stork_pairs"                "human_population_millions"
## [5] "yearly_birth_rate_thousands"
```

```
dataset[1:8,c('country','stork_pairs','yearly_birth_rate_thousands')]
```

```
##    country stork_pairs yearly_birth_rate_thousands
## 1  Albania         100                          83
## 2  Austria         300                          87
## 3  Belgium           1                         118
## 4 Bulgaria        5000                         117
## 5  Denmark           9                          59
## 6   France         140                         774
## 7  Germany        3300                         901
## 8   Greece        2500                         106
```

# Example 1: Does stork population cause birth rates ? (3/4)

```
library(ggplot2)
qplot( data = dataset,x = stork_pairs
      , y = yearly_birth_rate_thousands
      , geom = 'point') +
stat_smooth(method='lm') + theme_bw()
```

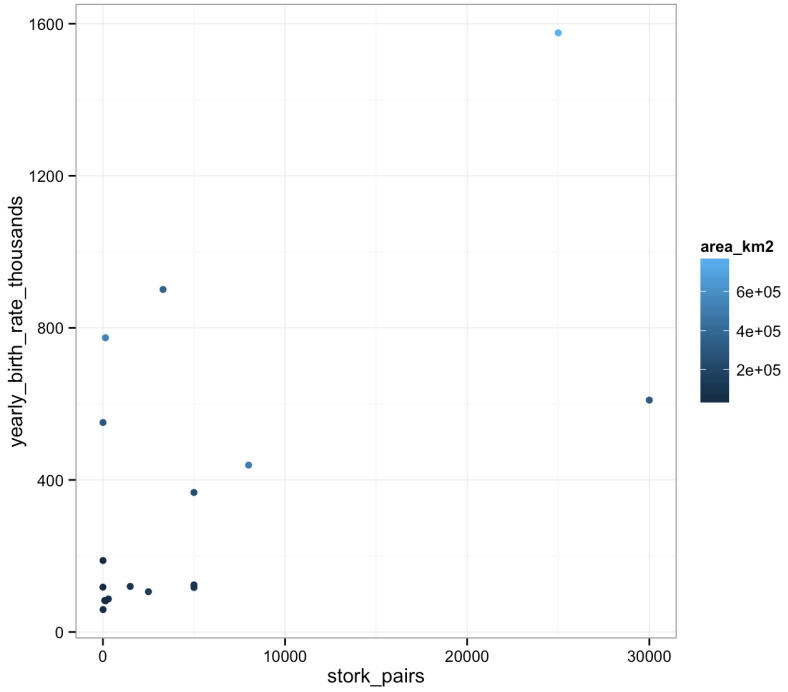



```
out.lm <- lm( yearly_birth_rate_thousands ~ stork_pairs, data
= dataset)
stargazer(out.lm, type = 'text')
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                     yearly_birth_rate_thousands
## -----------------------------------------------
## stork_pairs                   0.029***
##                                (0.009)
##
## Constant                      225.029**
##                                (93.561)
##
## -----------------------------------------------
## Observations                     17
## R2                               0.385
## Adjusted R2                      0.344
## Residual Std. Error       332.185 (df = 15)
## F Statistic             9.380*** (df = 1; 15)
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

# Example 1: Does stork population cause birth rates ? (4/4)

```
qplot( data = dataset,x = stork_pairs
        , y = yearly_birth_rate_thousands
        , geom = 'point', color = area_km2) +
  theme_bw()
```



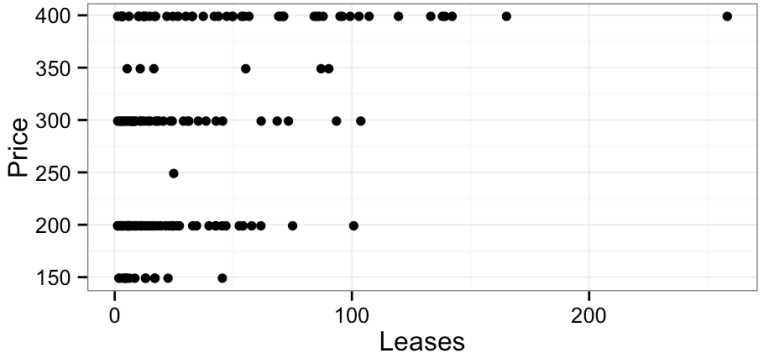

```
out.lm <- lm( yearly_birth_rate_thousands ~ stork_pairs +
area_km2, data = dataset)
stargazer(out.lm, type = 'text')
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                     yearly_birth_rate_thousands
## ------------------------------------------------
## stork_pairs                    0.006
##                               (0.006)
##
## area_km2                     0.002***
##                               (0.0002)
##
## Constant                      -7.412
##                               (56.702)
##
## ------------------------------------------------
## Observations                    17
## R2                             0.862
## Adjusted R2                    0.842
## Residual Std. Error      162.744 (df = 14)
## F Statistic           43.787*** (df = 2; 14)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

# Example 2: Do high prices make movies sell more ?

- You want to estimate: $movie\_sales = \beta_0 + \beta_1 movie\_price + u$
- What do you think will be the relation between sales and price ?

```
library(data.table,verbose=FALSE,quietly=TRUE)
library(stargazer ,verbose=FALSE,quietly=TRUE)
library(ggplot2   ,verbose=FALSE,quietly=TRUE)
dataset <- read.csv(file='../datasets/vod_sales.csv')
dataset <- data.table(dataset)
qplot(data=dataset,y=price,x=leases,geom='point') + labs(x = 'Leases', y = 'Price') + theme_bw()
```



# Example 2: Do high prices make movies sell more ?

```
summary(out     <- lm( leases ~  price,data=dataset))
```

```
##
## Call:
## lm(formula = leases ~ price, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.293 -24.119  -5.695  10.893 204.707
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.46678    8.96496  -2.283   0.0237 *
## price         0.18549    0.03012   6.158 5.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.93 on 169 degrees of freedom
## Multiple R-squared:  0.1832, Adjusted R-squared:  0.1784
## F-statistic: 37.92 on 1 and 169 DF,  p-value: 5.214e-09
```

- Does this make any sense ? (If not what can be happening?)

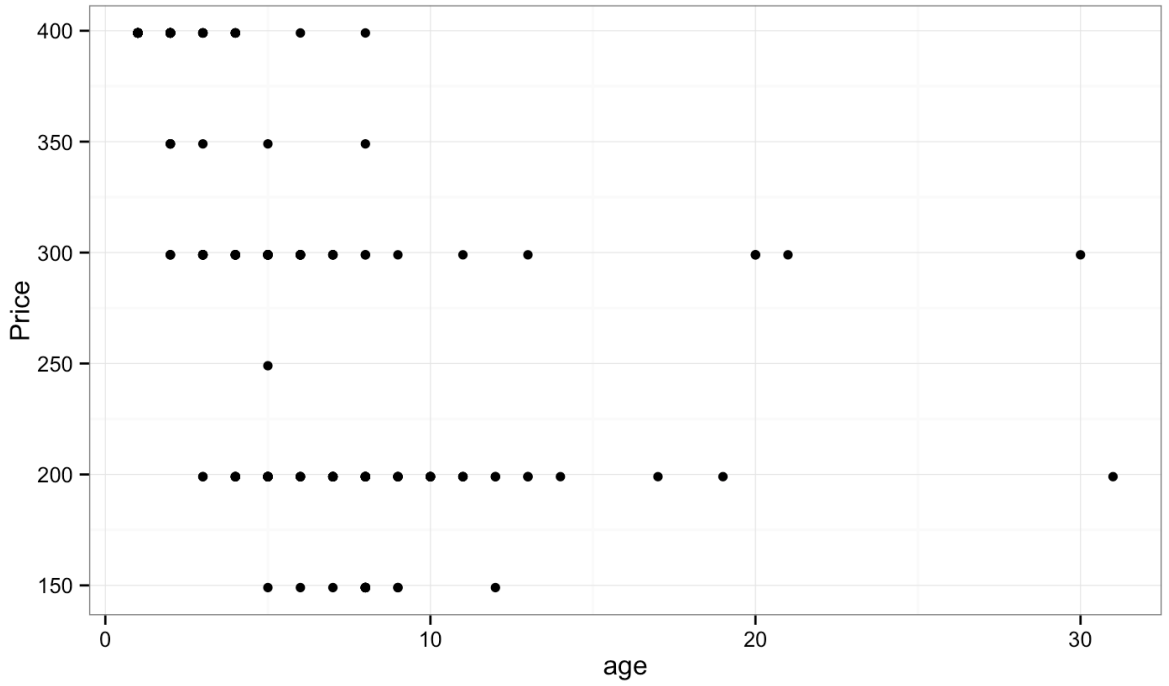
# Example 2: Do high prices make movies sell more ?

```
dataset$age     <- 2013 - dataset$year
summary(out     <- lm( leases ~  price + age,data=dataset))
```

```
##
## Call:
## lm(formula = leases ~ price + age, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.425 -20.664  -6.367  13.102 201.575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.72748   12.47086   0.299 0.765389
## price        0.13712    0.03444   3.982 0.000102 ***
## age         -1.76483    0.64474  -2.737 0.006862 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.28 on 168 degrees of freedom
## Multiple R-squared:  0.2181, Adjusted R-squared:  0.2088
## F-statistic: 23.43 on 2 and 168 DF,  p-value: 1.057e-09
```

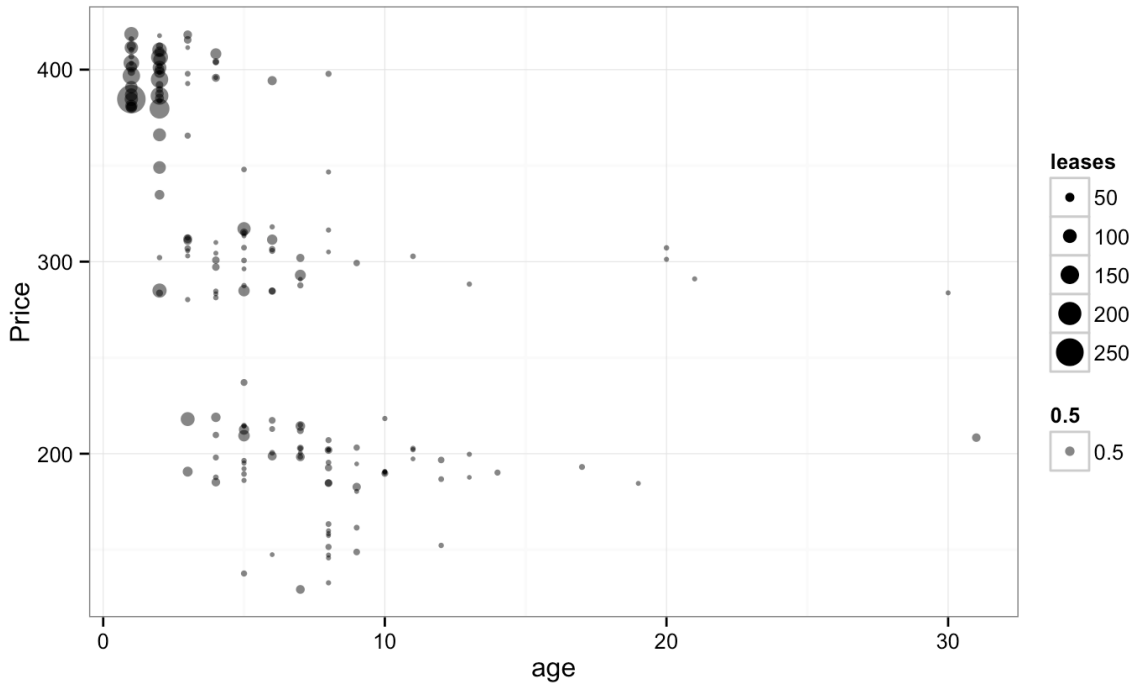
# Example 2: Do high prices make movies sell more ?

```
qplot(data=dataset,y=price,x=age,geom='point') +
  labs(x = 'age', y = 'Price') + theme_bw()
```



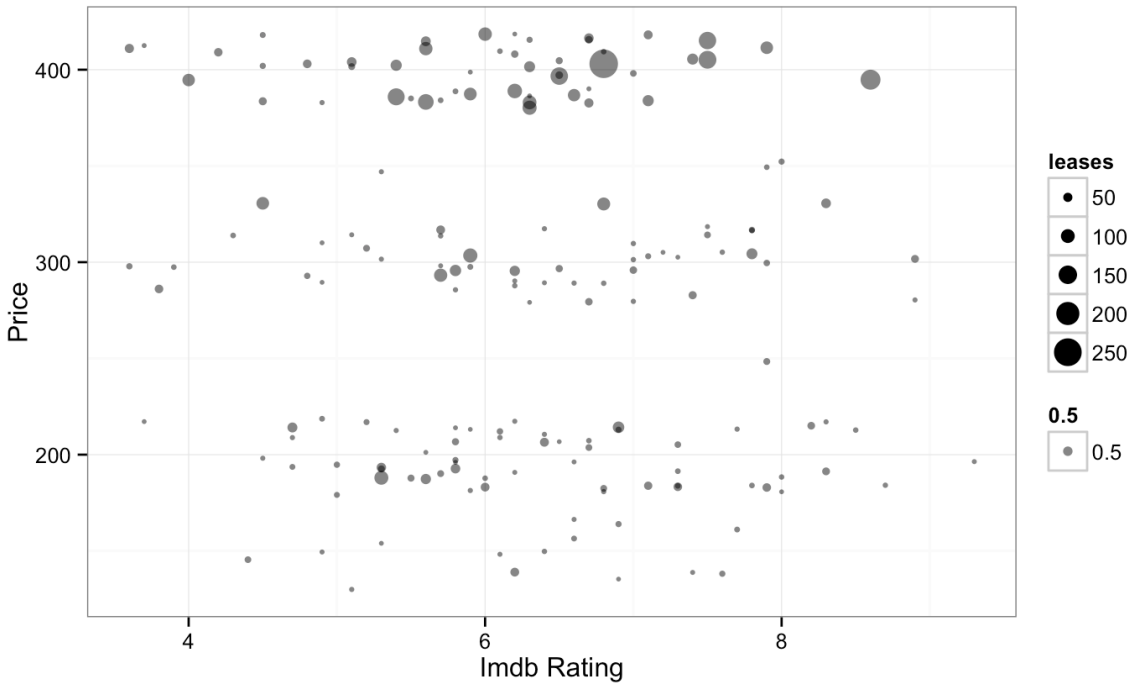
# Example 2: Do high prices make movies sell more ?

```
qplot(data=dataset,y=price,x=age,geom='point', position =  position_jitter(w = 0.0, h = 20)
      , size = leases, alpha = 0.5) +
  labs(x = 'age', y = 'Price') + theme_bw()
```



# Example 2: Do high prices make movies sell more ?

```
qplot(data=dataset,y=price,x=imdb,geom='point', position =  position_jitter(w = 0.0, h = 20)
      , size = leases, alpha = 0.5) +
  labs(x = 'Imdb Rating', y = 'Price') + theme_bw()
```



# Example 2: Do high prices make movies sell more ?

```
summary(out     <- lm( leases ~  price + I(age == 2) + I( age >= 3  & age <= 9) + I(age >=10) ,data=dataset))
```

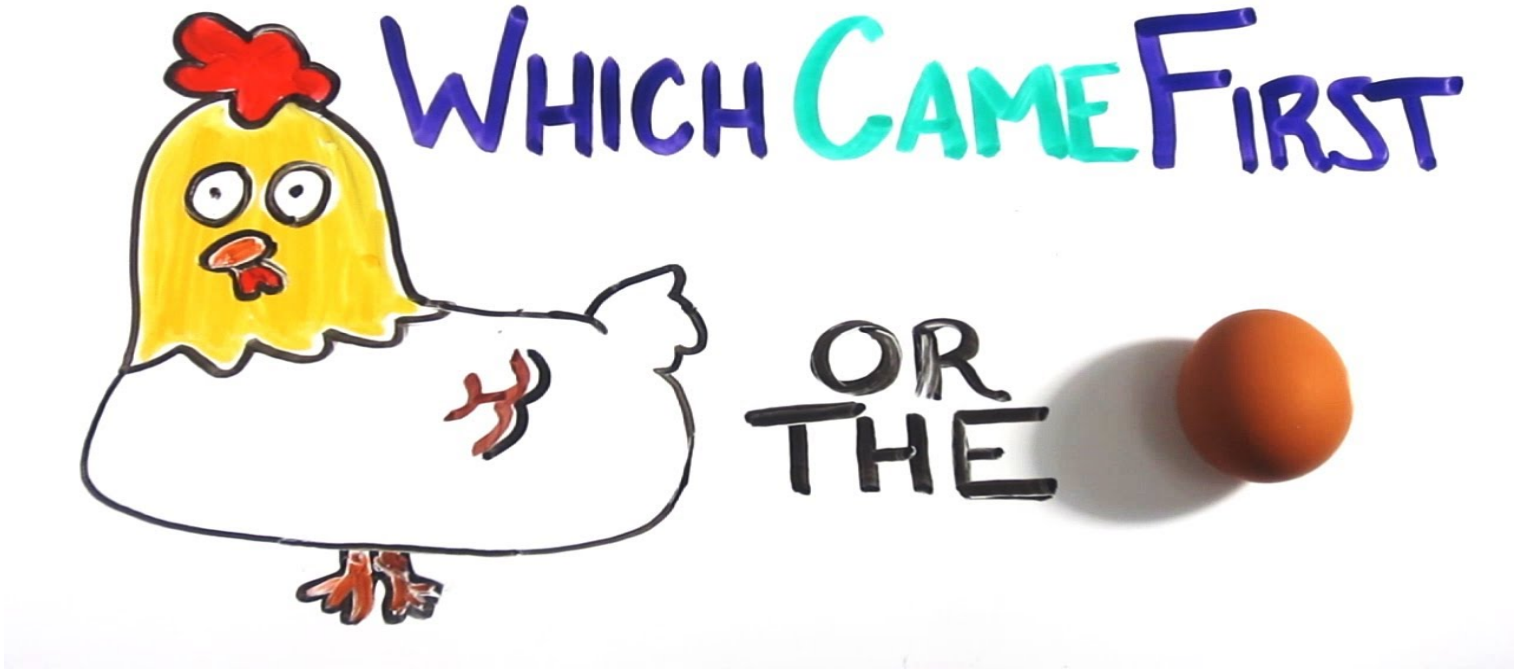
```
##
## Call:
## lm(formula = leases ~ price + I(age == 2) + I(age >= 3 & age <=
##     9) + I(age >= 10), data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.593 -15.199  -5.904  12.194 181.407
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  68.90888   16.88737   4.080 6.97e-05 ***
## price                         0.01988    0.03797   0.524    0.601
## I(age == 2)TRUE             -11.52135    9.81053  -1.174    0.242
## I(age >= 3 & age <= 9)TRUE -52.29505    9.75924  -5.359 2.77e-07 ***
## I(age >= 10)TRUE           -63.44998   11.99849  -5.288 3.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.66 on 166 degrees of freedom
## Multiple R-squared:  0.3409, Adjusted R-squared:  0.325
## F-statistic: 21.46 on 4 and 166 DF,  p-value: 2.758e-14
```

# Reverse Causality

# Simultaneity (1/2)

- So far we showed how OLS can produced biased estimates for the parameters when we have omitted variables
- Conceptually, this problem was straightforward. In the omitted variables case, there is a variable (or more than one) that we would like to hold fixed when estimating the ceteris paribus effect of one or more of the observed explanatory variables
- We could estimate the parameters of interest by OLS if we could collect better data
- **Another important form of bias of explanatory variables is simultaneity**
- **Simultaneity** arises when one or more of the explanatory variables is jointly determined with the dependent variable, typically through an equilibrium mechanism

# Simultaneity (2/2)



# Example: Murder rates and the size of the police force

- Cities often want to determine how much additional law enforcement will decrease their murder rates
- A simple cross-sectional model to address this question is (murdpc is murders per capita, polpc is number of police officers per capita, and incpc is income per capita):

$$murdpc = a_1 polpc + b_{10} + b_{11} incpc + u1$$

- **The question we hope to answer**: If a city **exogenously** increases its police force, will that increase, on average, lower the murder rate?
- But can we ever think of police force size as being exogenously determined?
- **Probably not**, a city's spending on law enforcement is at least partly determined by its expected murder rate. To reflect this, we postulate a second relationship:

$$polpc = a_2 murdpc + b20 + otherfactors$$

# Boar Example: illustrating the problem

- Explanatory variables determined simultaneously with the dependent variable are generally correlated with the error term
- Lets solve for $y_2$ in the board to see what the following equations mean:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$
$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$$

- $(1 - \alpha_1\alpha_2)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2$
- Because $z_1$ and $u_1$ are uncorrelated by assumption, the issue is whether $y_2$ and $u_1$ are correlated. From the reduced form in we see that $y_2$ and $u_1$ are correlated if and only if $\frac{\alpha_2 u_1 + u_2}{1-\alpha_1\alpha_2}$ and u1 are correlated (because z1 and z2 are assumed exogenous). But this is a linear function of $u_1$ and $u_2$, so it is generally correlated with $u_1$

# Conclusion

- Regression-based causal inference, the work horse of the data (policy) analyst, is predicated on the assumption that when key observed variables have been made equal across treatment and control groups, selection bias from the things we can't see is also mostly eliminated

- The greatest threat to causal modeling in regression analysis are violations of SLR.4:
  - **Omitted variables**: we omit a variable that actually belongs in the true (or population) model that is correlated with a variable in the model (which we have seen in detail)
  - **Reverse causality**: the chicken and the egg problem. One of the explanatory variables determines the outcome which in turn determines the explanatory variable (which we have only touched)
  - **Measurement error in the explanatory variables**: we use an imprecise measure of an economic variable in a regression model (which i am just now mentioning)