# POLI210: Political Science Research Methods

Lecture 13.2: Linear regression II

Olivier Bergeron-Boutin

November 25th, 2021

## Boring admin stuff

- Due dates:
    - Problem set: December 2nd
        - Try to knit right now!
    - Team project: December 6th
    - Quiz: December 2nd to December 6th – no late penalty until the 17th
        - Will be posted soon
- Please complete course evals on Minerva

## Adding more covariates

As seen repeatedly in the class, correlation $\neq$ causation

- Why? Because of **confounders**
- But if we can “adjust” for all relevant confounders (“control” for them)
- We have a stronger claim to causality
- In addition, from a predictive inference framework, we can make better predictions of the value $Y$ will take on

In our regressions, we will include additional covariates

- Covariates = independent variables = explanatory variables
- Just to be clear, we keep the same dependent variable
- But now seek to explain it using multiple variables

# Our new regression equation

With two independent variables, we now have:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

We still have our intercept, $\beta_0$

- But now we have one coefficient for each independent variable: $\beta_1$ and $\beta_2$
- Our interpretation of each coefficient is now a bit different
- $\beta_1$ representes the expected change in $Y$ occuring as a result of a one-unit change in $X_1$…**holding other covariates constant**
- In this case, holding $X_2$ constant
- What we can now say:
    - The association between $X_1$ and $Y$ that $\beta_1$ identifies is not due to confounding by $X_2$

# New regression model with incumbent data

```r
reg2 <- lm(formula = partyincshr ~ gdpchangeyr3 + age,data = economy)
summary(reg2)
```

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3 + age, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7772  -3.6559  -0.1206   3.6909  10.6179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.97700    7.56543   7.267 4.65e-09 ***
## gdpchangeyr3  0.63960    0.21801   2.934   0.0053 **
## age          -0.08605    0.12965  -0.664   0.5104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.581 on 44 degrees of freedom
##   (184 observations deleted due to missingness)
## Multiple R-squared:  0.1664, Adjusted R-squared:  0.1285
```

# Comparing our two models

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 50.254*** | 54.977*** |
| | (0.999) | (7.565) |
| GDP change (year 3) | 0.605** | 0.640** |
| | (0.220) | (0.218) |
| Age | | -0.086 |
| | | (0.130) |
| Num.Obs. | 48 | 47 |
| R2 | 0.142 | 0.166 |
| R2 Adj. | 0.123 | 0.129 |

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

# Non-linear relationships are not well captured

Linear regression models are good at estimating **linear** relationships

- When the relationship between $X$ and $Y$ is non-linear, things get more complicated
- (There are ways to account for this, but that's for 311)
- In short, our $\beta$'s will not capture the relationship well

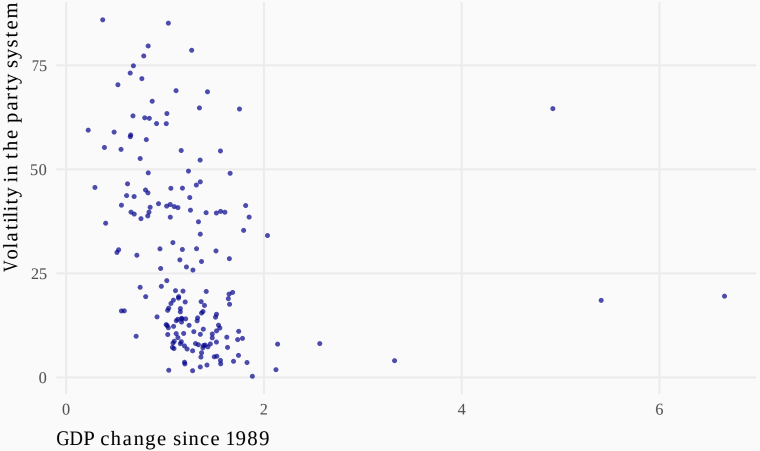
# Airbnb's in London


**Figure 1:** Longitude and price of London (UK) Airbnb listings on March 4th, 2017

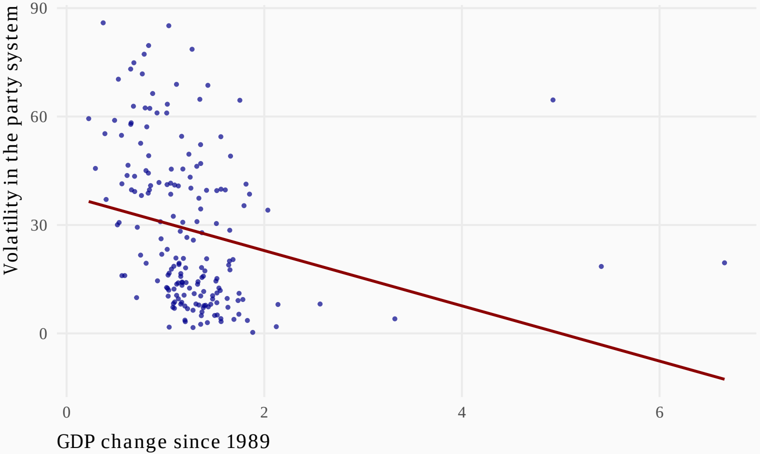
# Airbnb's in London

```
##
## Call:
## lm(formula = price ~ longitude, data = london)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117.52  -49.04  -21.41   22.07  893.80
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.2799     0.6018  131.75   <2e-16 ***
## longitude   -114.8015     3.8881  -29.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.69 on 53815 degrees of freedom
##   (60 observations deleted due to missingness)
## Multiple R-squared:  0.01594,    Adjusted R-squared:  0.01592
## F-statistic: 871.8 on 1 and 53815 DF,  p-value: < 2.2e-16
```

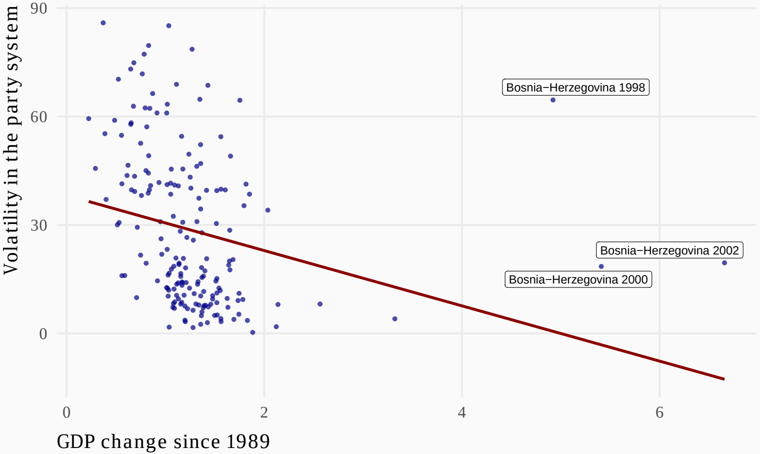
# Outliers can mess with your results



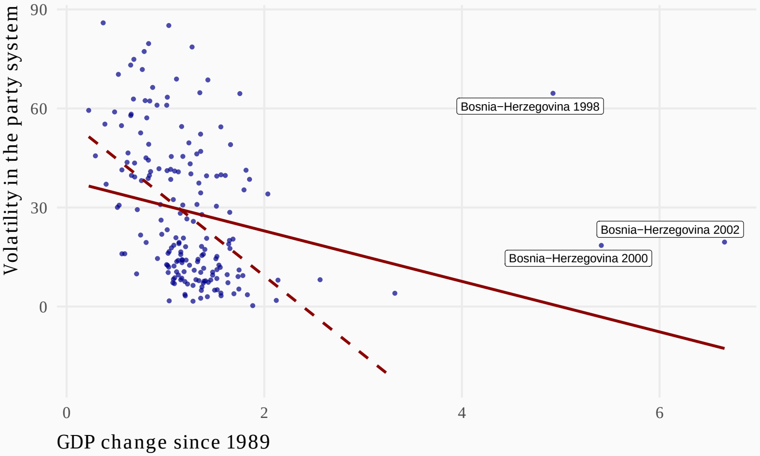
# Outliers can mess with your results



# Outliers can mess with your results



# Outliers can mess with your results

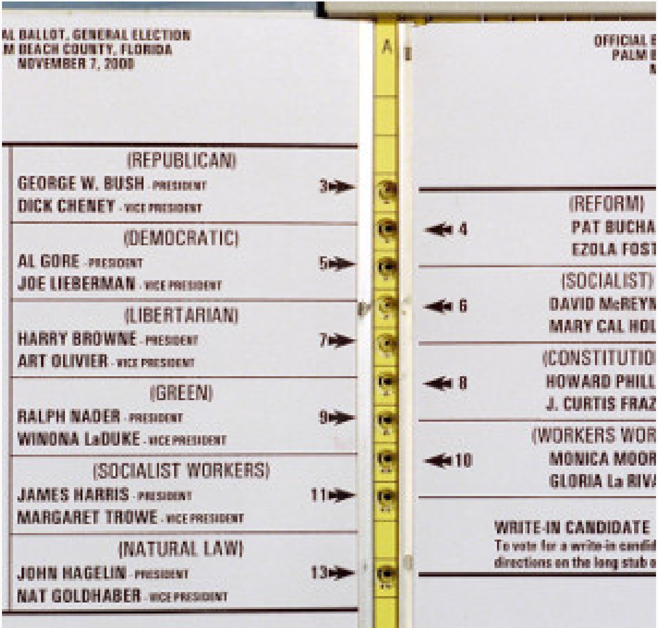


# Outliers can mess with your results

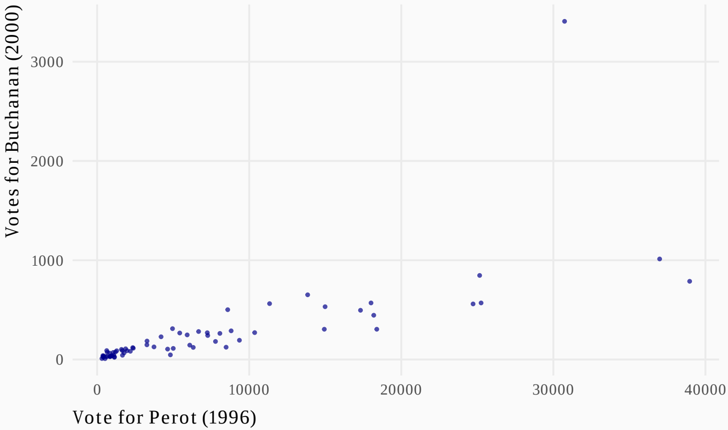
| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 38.201*** | 56.773*** |
| | (3.157) | (4.315) |
| GDP change since 1989 | -7.645*** | -23.763*** |
| | (2.178) | (3.415) |
| Num.Obs. | 184 | 181 |
| R2 | 0.063 | 0.213 |
| R2 Adj. | 0.058 | 0.208 |

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

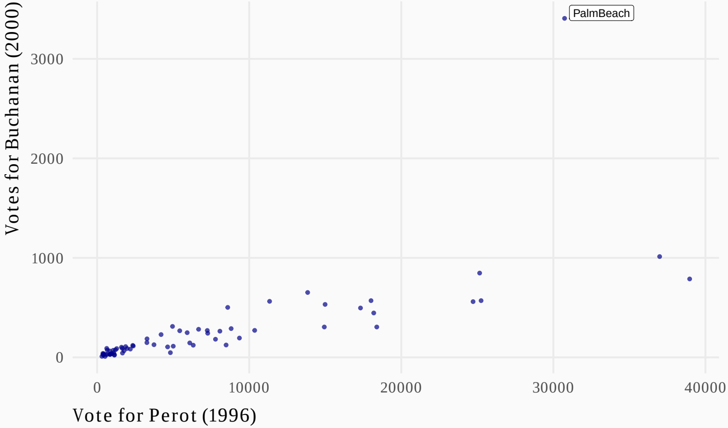
# Outliers are not just a nuisance



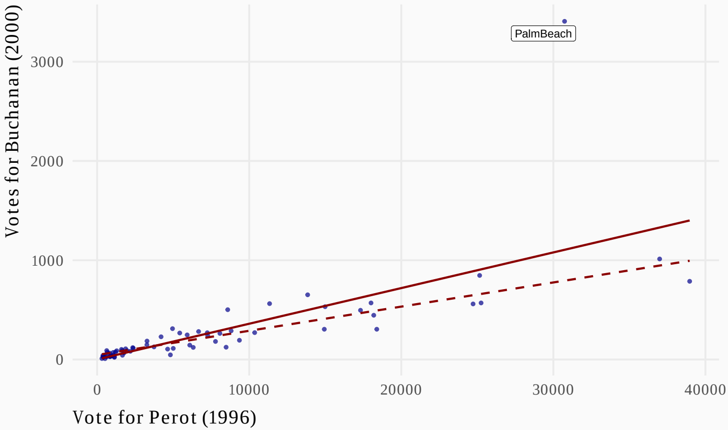
# Outliers are not just a nuisance



# Outliers are not just a nuisance



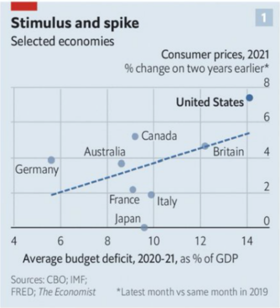
# Outliers are not just a nuisance



# But generally we should be wary of their influence

**Steven Rattner**
@SteveRattner

Proponents of “transitory” inflation cite rising prices as a global phenomenon. True, but 1) our inflation is much higher than peers and 2) there is, unsurprisingly, a relationship between pandemic fiscal response and prices. cc: @TheEconomist




The Economist

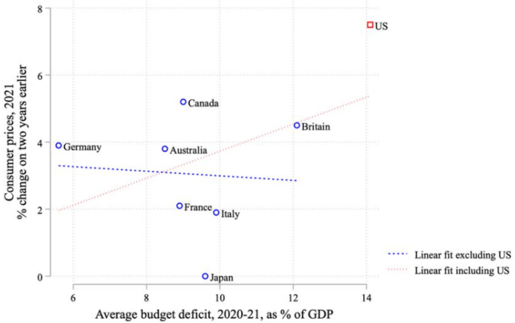
2:06 PM · Nov 22, 2021 · Sprout Social

# But generally we should be wary of their influence

Arindrajit Dube
@arindube

I usually wouldn't comment on a N=8 scatterplot, but this particular example turns out to be very useful pedagogically, illustrating the impact of a single, *high leverage*, observation can have on the fitted regression slope.

Inclusion of US (just 1 data point) flips the sign.




Source: approximated from Economist.com

# Seems like an important guy...

## A note on listwise deletion

What happens to your regression when the dataset has missing data?

- Listwise deletion: any observation that is missing at least one value for any independent variable or the dependent variable will be thrown out
- i.e. the model will not use that observation

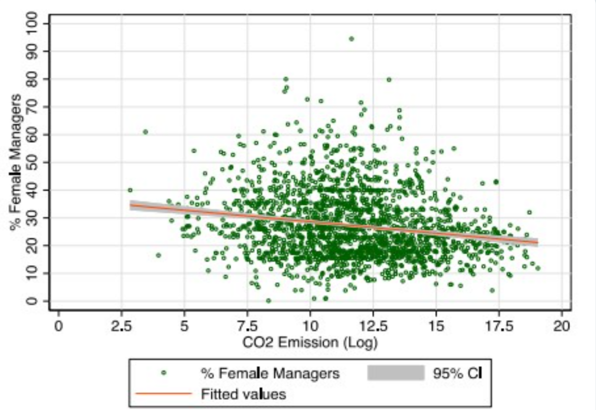
| Grade (DV) | Happiness | Hours of sleep |
|---|---|---|
| 87 | NA | 7 |
| 81 | 8 | NA |
| NA | 6 | 3 |

Running a model: $\text{Sleep}_i = \beta_0 + \beta_1 \text{Happiness}_i + \beta_2 \text{Sleep}_i + \epsilon_i$

# What’s wrong here?

**Bank for International Settlements**
@BIS_org

Appointing more women to managerial positions improves firms’ environmental performance 🌎
#GenderDiversity #ClimateChange #COP26🌍
bis.org/publ/work977.h…



# COVID and democracy

```r
load("lectures/lecture_13.1/survey.RData")
# Model only with anxiety
m1_anxiety <- lm(
  ea_3item ~ anxiety_scale,
  data = survey
)

# formula for the fully-specified model
reg_formula <- ea_3item ~ anxiety_scale + birth_decade + educ_4cat +
  deprivation_scale + authority_scale_alt + partyid

# Fully-specified model
m2_anxiety <- lm(
  formula = reg_formula,
  data = survey
)
```

# Covid and democracy

| | Model 1 | Model 2 |
|---|---|---|
| COVID-related anxiety | 0.243 (0.014)*** | 0.166 (0.017)*** |
| Born in the 1950s | | -0.017 (0.015) |
| Born in the 1960s | | 0.004 (0.016) |
| Born in the 1970s | | 0.050 (0.016)** |
| Born in the 1980s | | 0.085 (0.016)*** |
| Born since 1990 | | 0.071 (0.015)*** |
| Deprivation | | 0.090 (0.016)*** |
| Authoritarianism | | 0.055 (0.009)*** |
| Completed high school | | -0.010 (0.015) |
| Some postsecondary | | -0.034 (0.016)* |
| College graduate | | -0.035 (0.016)* |
| Conservative partisan | | -0.079 (0.010)*** |
| NDP partisan | | -0.043 (0.013)*** |
| Green partisan | | -0.009 (0.019) |
| Non-partisan | | -0.035 (0.015)* |
| Constant | 0.255 (0.006)*** | 0.242 (0.019)*** |
| Num.Obs. | 2417 | 2177 |
| R2 | 0.114 | 0.206 |

# Some lessons from the class

# Big takeaways

- Empirical research is hard!
  - People who spend their lives doing this get it wrong all the time
  - The first step: recognize how hard this is
- Match the strength of your claims to the strength of your evidence
  - Recognize uncertainty
  - When reading about politics in popular media, notice how people don't do that
- Think about the sort of evidence that would make you change your mind
  - If the answer is none.............

# How can I use this?

To learn more...

- POLI311: Quantitative methods
- POLI313: Qualitative methods
- Other than course work: find data that you like!

To apply what we've learned...

- In popular media:
  - How strong are the claims being made
  - How strong is the evidence that is being presented?
    - Sometimes, there is no empirical evidence; there are entire news articles based on the intuition of "some dude"
- In academics:
  - When reading empirical research
  - When reading non-empirical research: what would a good empirical test look like?

# References