Bush 631-607: Quantitative Methods

Lecture 14 (11.30.2021): Review & Summary

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Fall 2021

What is today's plan?

- Review course topics.
- Research designs.
- Predictions, probability and estimations.
- Social science and real-world politics.
- Data science in the real-world.

Course review: Causality

- Establish Cause $\rightarrow$ effect.
- Using (hypothetical) *Counterfactual*.

Establish causality

Experimental research design

- Treatment and control groups.
- Same outcome measures.
- Gold standard → randomization.
- Calculate ATE over group of respondents.

$$SATE = \frac{1}{n} * \sum_{i=1}^{n} * Y_i(1) - Y_i(0)$$

Experiments: Example

- President's type and support for FP (2019)

```
# Diff-in-means = ATE of type
mean(hawks$approve_b[hawks$hawk_t == 1], na.rm = T) -
  mean(hawks$approve_b[hawks$hawk_t == 2], na.rm = T)
```

```
## [1] -0.1202408
```

```
# Also with tapply()
tapply(hawks$approve_b, hawks$hawk_t, mean, na.rm = TRUE)
```

```
##         1         2
## 0.5774336 0.6976744
```

```
# Can also use subsets and diff-in-means
```

Establish causality

Observational research design

- Using data to assess causality.
- Good for generalizing results.
- Not as good for randomization.
- Problem of pre-treatment variables (confounders).

Causality in observational data

- Survey: political polarization and views of China (2020)
- Party *Thermometer* $\rightarrow$ *Dems*/*Reps* > 50 support for party.

```
levels(as.factor(threat$china_frenemy))
```

```
## [1] ""           "Ally"       "Enemy"      "Friendly"   "Unfriendly"
```

Causality in observational data

- China as the enemy?

```r
# Diff-in-means (Support Dems thermometer)
mean(threat$china[threat$affective_Dem < 50], na.rm = T) -
  mean(threat$china[threat$affective_Dem > 50], na.rm = T)
```

```
## [1] 0.2243064
```

```r
# Using tapply() by political interest
app <- tapply(threat$china, threat$pol_interest, mean, na.rm = TRUE)
sort(app)
```

```
## Not interested at all Moderately interested   Slightly interested
##              2.526570              2.580093              2.643333
##       Very interested                           Extremely interested
##              2.650142              2.700000              2.713311
```

Assessing research designs

Strengths and Weaknesses

- Internal validity:
  - How does the design helps answering the research Q?
  - Experiments $\rightarrow$ strong (randomization).
  - Observational $\rightarrow$ weak (confounders).
- External validity:
  - Can we generalize the results from sample?
  - Experiments $\rightarrow$ weak (hypothetical).
  - Observational $\rightarrow$ strong (real-world, cross-national).

Course review: Measurement

- Apply quant methods for social science.
- Measures $\rightarrow$ the context of concepts.
- Challenge: latent factors
  - What is ideology? How do we measure it?
  - Terrorism?
  - Democracy - polity vs. freedom house scores.

Measurement challenge

- Challenge of missing values (non-responses)
- No data collected / respondents refuse to answer (DKs).
- NAs in our data

```
# Solving with na.rm = TRUE
mean(bushdata$Pol_survMusl)
```

```
## [1] NA
```

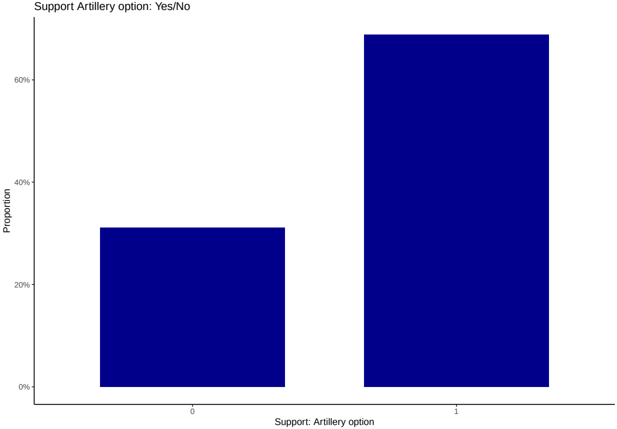
```
mean(bushdata$Pol_survMusl, na.rm = TRUE)
```

```
## [1] 2.067584
```

Visuals

- Barplot: counts/proportions for categories

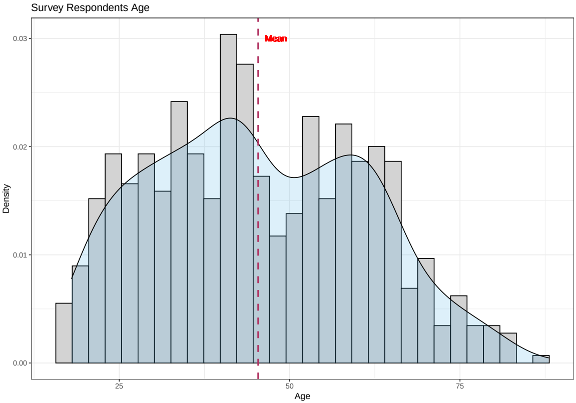
```
ggplot(wardata, aes(x=factor(prefer_artillery_dummy))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), width = 0.7, fill = "darkblue")
  xlab("Support: Artillery option") + ylab("Proportion") +
  scale_y_continuous(labels=scales::percent) + ggtitle("Support Artillery optio
  theme_classic()
```



Visuals

- Histogram: distribution of numerical variable

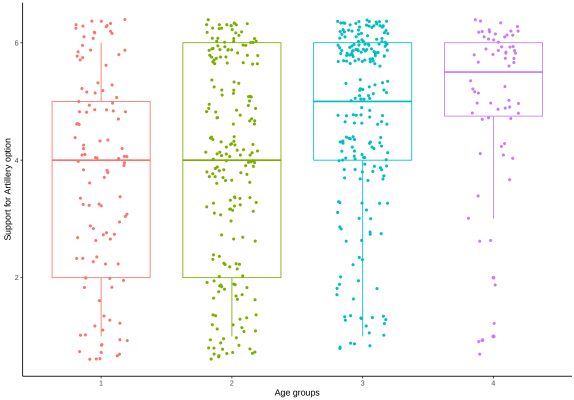
```
ggplot(wardata, aes(x=age)) +
  geom_histogram(aes(y=..density..), colour="black", fill="lightgrey")+
  geom_density(alpha=.2, fill="#56B4E9") +
  xlab("Age") + ylab("Density") + theme_bw() + ggtitle("Survey Respondents Age") +
  geom_vline(aes(xintercept=mean(age)),
             color="maroon", linetype="dashed", size=1) +
  geom_text(x = 48, y = 0.03, label = "Mean", col = "red")
```



Visuals

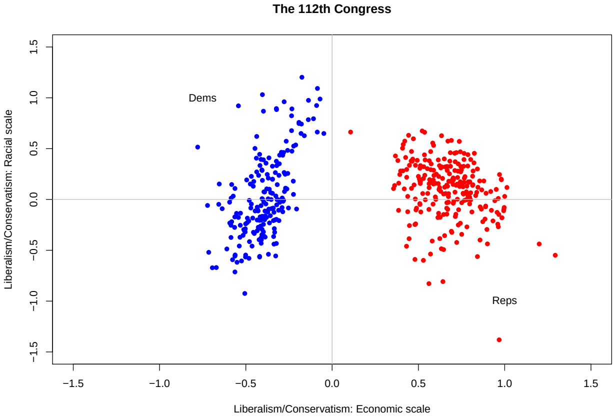
- Boxplot: Compare single variable distribution

```
ggplot(wardata, aes(x=factor(agegroup), y = artillery_approve,
                    color = factor(agegroup))) +
  geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2)) +
  xlab("Age groups") + ylab("Support for Artillery option") +
  theme_classic() + theme(legend.position = "none")
```



Visuals

- Scatterplot: Visualize bivariate relationship



Visuals

- Plot time trends (UN Voting data 1946-2012)

Predictions

- Predict with sample mean: using loops.
- Prediction error = actual outcome - predicted outcome.
- RMSE: average magnitude of prediction error.

- Correlations:
  - Summary of bivariate relationship.
  - How two factors 'move together' on average.
  - Always relative to mean value.

```
# Voting with US
cor(unvoting$idealpoint, unvoting$PctAgreeUS, use = "pairwise")
```

```
## [1] 0.7498446
```

Predictions

- The linear model: $Y = \alpha + \beta * X_i + \epsilon$
- Model elements:
  - Intercept (α): the average value of Y when X is zero.
  - Slope (β): the average increase in Y when X increases by 1 unit.
  - Error/disturbance term (ϵ): the deviation of an observation from a perfect linear relationship.
- Least squared:
  - How to estimate the regression line.
  - 'Select' $\hat{\alpha}, \hat{\beta}$ to minimize SSR.
  - R syntax: lm(y ~ x, data = mydata)

Regression to the mean

- High (low) observations are followed by low (high) observations.
- Observations 'regress' towards the average value of the data.

Fighter pilots performance?

Cursed??

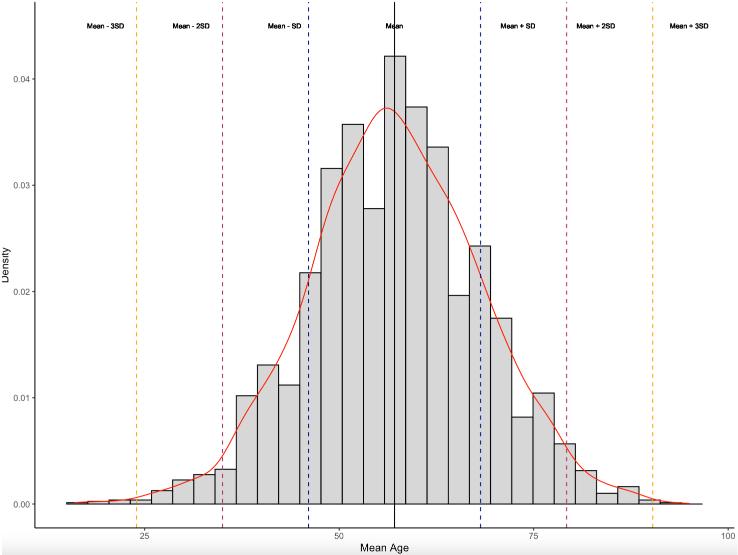
Probability

- Quantify uncertainty: step 1.
- Sample space, events (cards, coin toss)
- Probability: $P(A) = \frac{Elements(A)}{Elements(\Omega)}$
- Conditional probability $= P(A|B) = \frac{P(A\&B)}{P(B)}$
- *Monty Hall problem*.

Probability

- Random variables: from events to numbers.
- Uncertainty of sample means or sums.
- Probability distributions: Bernoulli (binary), Binomial (discrete).
- Expectations of r.v. $\rightarrow$ population value.
- Variance of r.v. $\rightarrow$ 'spread' of distribution.
- CLT and large samples.

Normal distribution

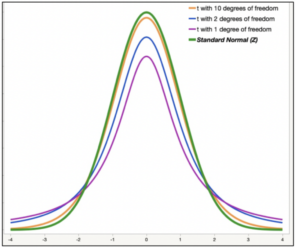


Uncertainty and estimation

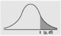
- Estimate based on r.v.s.
- Quantity of interest: *point estimate* (mean / diff-in-means)
- How to learn of estimator distribution? simulations...
- Calculate SD, or SE in single sample.
- Construct 95% CIs - how to interpret?

t-distribution

- Small samples.
- Account for DOF.
- More conservative, why? ‘fatter tails’.



Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	——	——	80%	90%	95%	98%	99%	99.9%

Hypothesis tests

- Estimators: sample means / diff-in-means
- *Proof by contradiction*.
- Steps for testing:
  1. Define null and alternative hyps (H_0; H_1).
  2. Select *test statistic* and level of test (α).
  3. Derive reference distribution.
  4. Calculate p-values.
  5. Make a decision: reject/retain.
- **Decision rule**:
  - **Reject null** if p-value is *below* $\alpha = 0.05$
  - Otherwise **retain the null** or **fail to reject**.

Hypothesis test

- Run *Two-sample t-test* with t.test()

```
t.test(exp.dat$cont_cor1[exp.dat$trt1 == 0],
       exp.dat$cont_cor1[exp.dat$trt1 == 1])
```

```
##
##  Welch Two Sample t-test
##
## data:  exp.dat$cont_cor1[exp.dat$trt1 == 0] and exp.dat$cont_cor1[exp.dat$tr
## t = -13.697, df = 993.53, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.59653 -17.68267
## sample estimates:
## mean of x mean of y
##  1489.333  1509.973
```

Least squared estimator

- Uncertainty in *least squared* estimator:
  - Generate reference distribution.
  - Calculate SEs.
  - Construct 95% CIs.
  - Run hypotheses tests.
  - Results are 'statistically significant', or not.

- Assumptions for regression estimates:

(1) Exogeneity: mean of ϵ_i does not depend on X_i

$$E(\epsilon_i|X_i) = E(\epsilon_i) = 0$$

(2) Homoskedasticity: variance of ϵ_i does not depend on X_i

$$V(\epsilon_i|X_i) = V(\epsilon_i) = \sigma^2$$

Putting everything together

- Hypotheses:
  - $H_0 : \beta_1 = 0$
  - $H_a : \beta_1 \neq 0$
- Our estimators: $\hat{\beta}_0, \hat{\beta}_1$
- SE and CIs:
  - $\hat{\beta}_0 \pm 1.96 * \hat{SE}(\hat{\beta}_0)$
  - $\hat{\beta}_1 \pm 1.96 * \hat{SE}(\hat{\beta}_1)$
- Hypotheses test:
  - Test statistic: $\frac{\hat{\beta}_1 - \hat{\beta}_1^*}{\hat{SE}(\hat{\beta}_1)} \sim$ N(0,1)
  - $\hat{\beta}_1$ is **statistically significant** if $p < 0.05$.

Rebels and Nukes (2015)

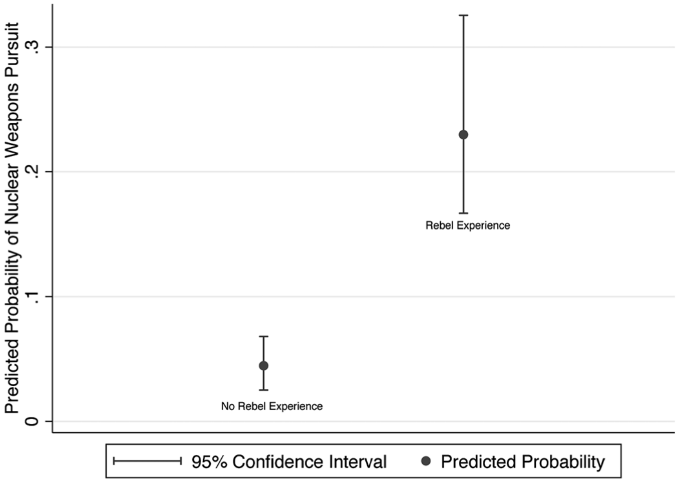
- Multivariate regression: account for confounders

```
summary(lm(pursuit ~ rebel + milservice + polity2, data = nukes))
```

```
##
## Call:
## lm(formula = pursuit ~ rebel + milservice + polity2, data = nukes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06587 -0.04408 -0.02544 -0.01020  0.99682
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0073899  0.0027782   2.660  0.00783 **
## rebel       0.0320096  0.0044238   7.236 5.08e-13 ***
## milservice  0.0217914  0.0045106   4.831 1.38e-06 ***
## polity2     0.0004679  0.0002801   1.670  0.09489 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1672 on 7684 degrees of freedom
##   (1164 observations deleted due to missingness)
## Multiple R-squared:  0.01596,    Adjusted R-squared:  0.01558
## F-statistic: 41.54 on 3 and 7684 DF,  p-value: < 2.2e-16
```

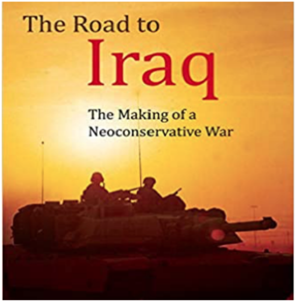
OLS coefficient interpretation

- Rebel experience and nuclear technology (2015)



How to use our research?

Applying theories and IR research



Applying IR research in global affairs

- The motivation:
  - We study IR or social science dynamics.
  - Do policymakers use? or even care about this knowledge?

The New York Times http://nyti.ms/1f4Huyy

SundayReview | OP-ED COLUMNIST

Professors, We Need You!

Nicholas Kristof FEB. 15, 2014

SOME of the smartest thinkers on problems at home and around the world are university professors, but most of them just don't matter in today's great debates.

The most stinging dismissal of a point is to say: "That's academic." In other words, to be a scholar is, often, to be irrelevant.

Applying IR research

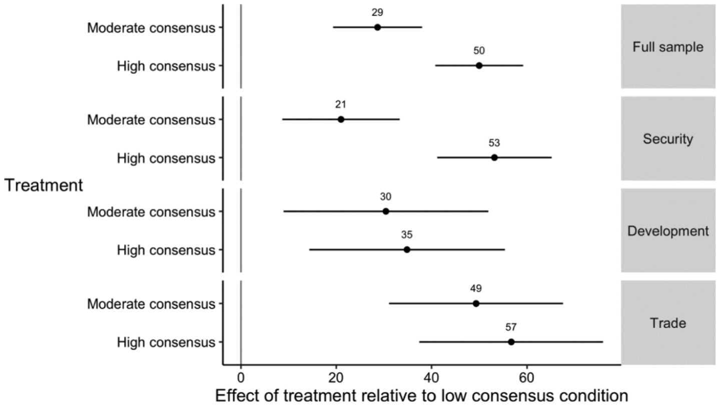
- Main challenge → **Time** and reading research article.
- Scholars adapt:
  - Joint forums: *bridging the gap*.
  - Policy-focused writing (Lawfare blog, War on the rocks, Monkey cage).
- Is it working?
- Ask policymakers. . .
- Previous work (2014):
  - Not really.
  - Academic work not aimed to 'close the gap'.

Applying IR research

- Recent evidence (2021) $\rightarrow$ replicate 2014 survey.
- Innovations for better insights:
  - Diverse sample - more areas FP.
  - Both high and low-ranking officials.
  - Embedded experiment for direct effects.
  - Broader conception for engagement (social media).
- Sample: 616 officials (Clinton, Bush, Obama administration).

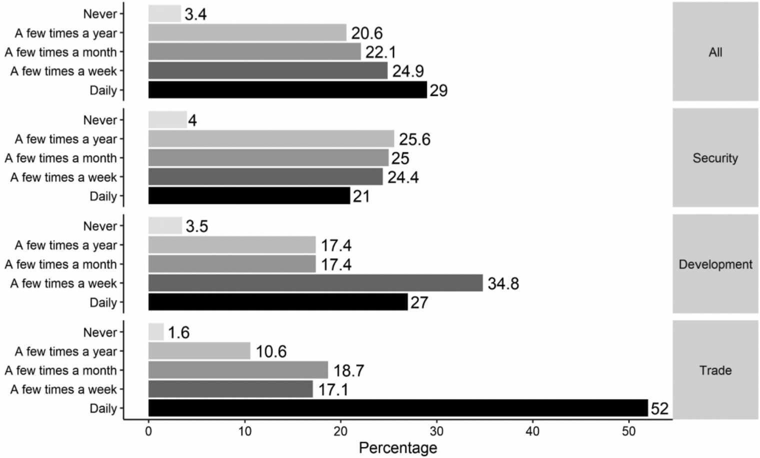
Useful political research?

- Research consensus and updating policy views.



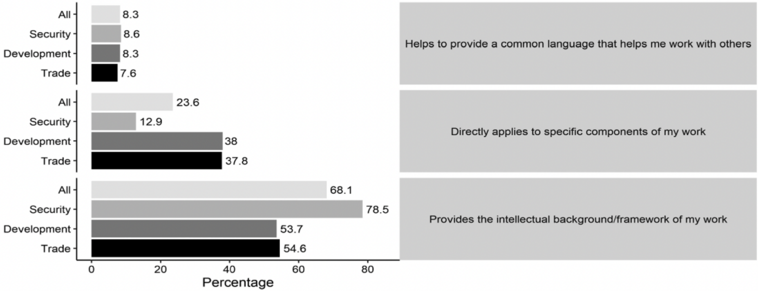
Useful political research?

- Frequency of using research into government work.



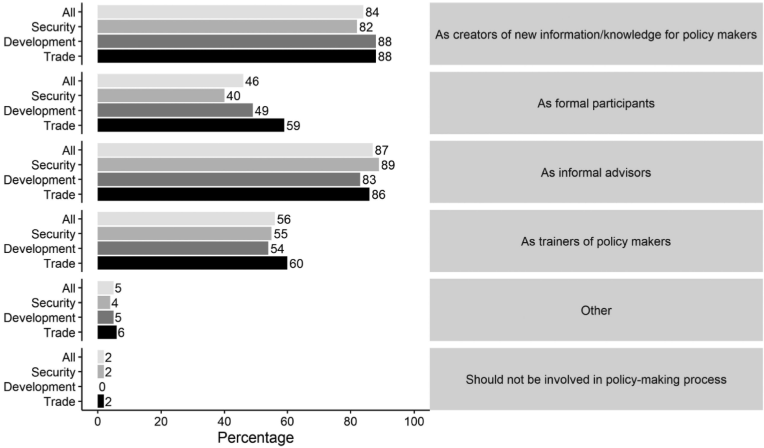
Useful political research?

- How do you use academic research?

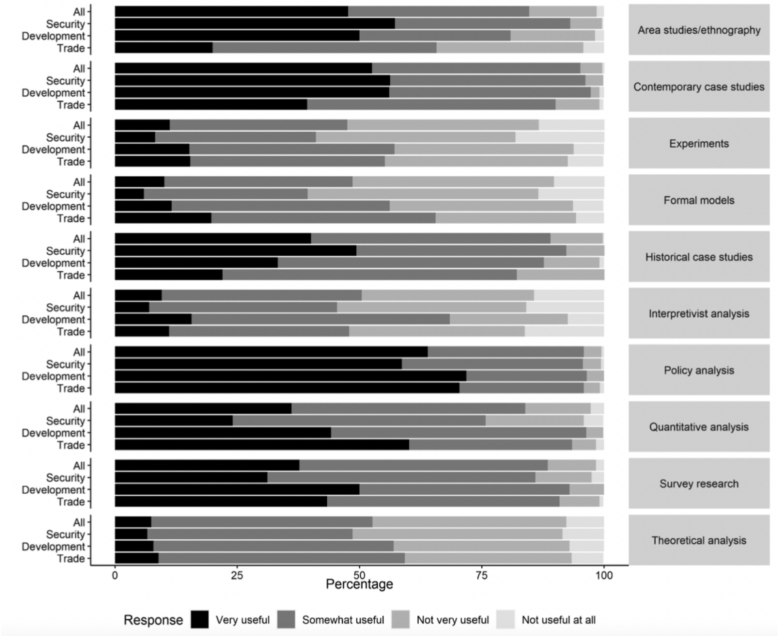


Useful researchers

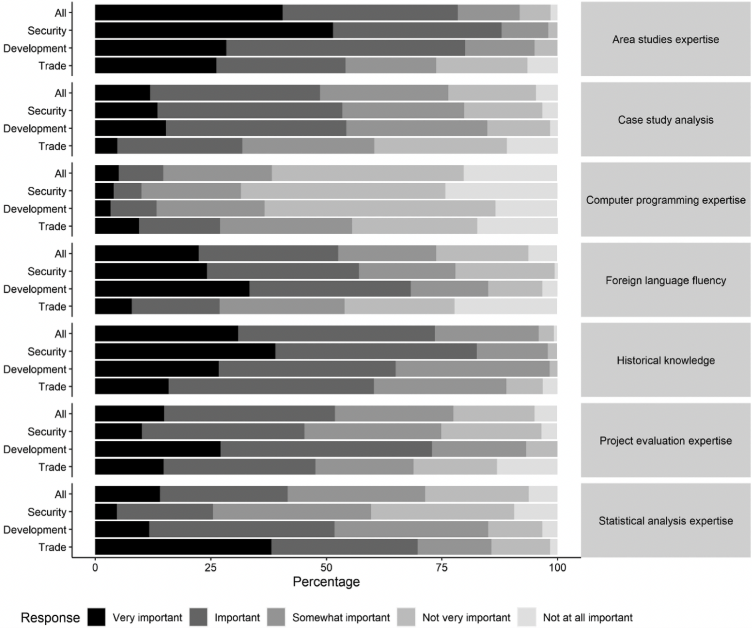
- How can researchers contribute the most?



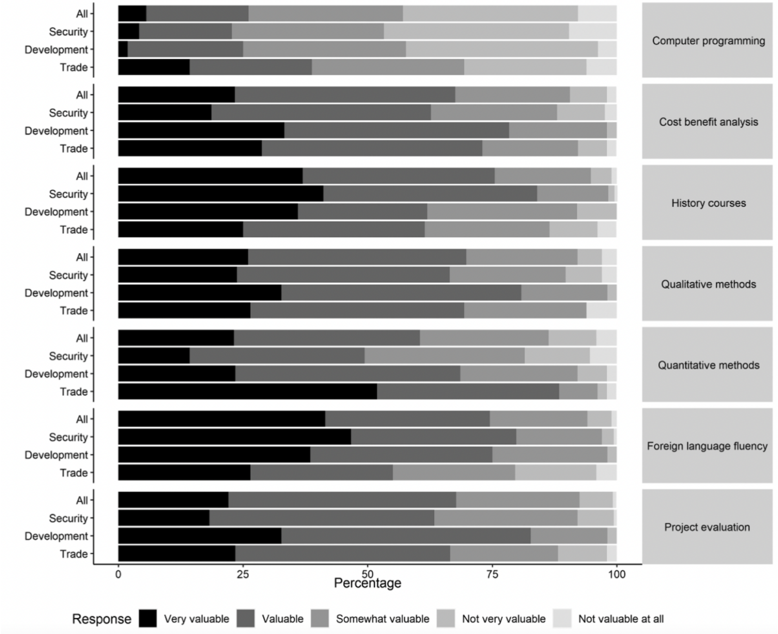
Useful skills. . .



Useful skills. . .



Working for the government. . .



Using political research

- How can research be useful (Byman & Kroenig 2016)
  - Practical and useful recommendations.
  - Focus - clarify complex situations.
  - Time relevant research.
- Scenarios for applying academic insights:
  - Challenge existing government knowledge - shocks (9.11, Soviet collapse).
  - Policy failures (Iraq insurgency outbreak).
  - Missing baseline knowledge (Somalia intervention).

Becoming useful political advisor

- Concrete steps:
  - Networking and personal connections.
  - 'Inject' research into bureaucracy.
  - Concise and clear reports in nonacademic outlets.
- Tamper expectations:
  - What is being relevant?
  - Not likely to drastically shape policy.
  - Influence the **deliberation** process.
- What's in it for policymakers?
  - Offer *contrarian arguments* to accepted view.

Data science in the real world

- Data analysis $\rightarrow$ set of tools to understand the world.
- The core role of probability.
- Apply complex concepts like repeated sampling.
- Bayesian logic and saving lives.
- Movies:
  - Prediction by the numbers: (Link)
  - Tails you win: (Link)