

Economics & Genetics

Lecture 5

DR. NIELS RIETVELD, NRIETVELD@ESE.EUR.NL


OFFICE MANDEVILLE T18-29



Power analyses

"Use power calculations to assess under what statistical conditions (e.g., the expected R^2 of a SNP) a GWAS on personal income is feasible. Use reasonable assumptions (e.g., based on earlier genetic associations reported in scientific papers) in your calculations, and use at least 1 scientifically formatted table to report the results."

`pwr.r.test(n = NULL, r = sqrt(0.0002), sig.level = 5e-8, power = 0.80, alternative = "two.sided")`

- *The assumption about the effect of a SNP on personal income deserves your attention and creativity*
- 

Today's agenda

Last week: Using GWAS, we managed to find SNP-outcome associations!

Main questions:

- How can we use GWAS results to identify causal effect in economic models?
- How can polygenic scores be used to understand the interplay between nature and nurture?

Literature:

- Van Kippersluis & Rietveld (2017): "Pleiotropy-robust Mendelian Randomization"
- Slob & Rietveld (2020): "The moderating impact of the genetic predisposition to smoking behaviour on the response to tobacco excise taxes"



Economics & Genetics Lecture 5

RECAP LECTURE 4



“Working horse” model from last week

y_i is the value of the outcome variable for individual i

μ is the intercept (constant)

β_j is the effect of SNP j (assumption of additivity, every allele has the same effect)

x_{ij} reflects the number of reference alleles for a SNP (0, 1, 2)

ε_i is the effect of exogenous residual factors

$i \in [1, \dots, N]$, N usually 5K-10K in a dataset

$j \in [1, \dots, J]$, $J > 1,000,000$ [Overidentification!]

$$y_i = \mu + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i$$

GWAS deals with overidentification (i.e. $J > N$) by running the regression for each SNP separately:

$$y_i = \mu + \beta x_i + \varepsilon_i \text{ for every SNP } j$$

Polygenic risk score

Based on the regression coefficients estimated in a GWAS, $\hat{\beta}_j$, we create the PGS as:

$$PGS_i = \sum_{j=1}^J \hat{\beta}_j x_{ij}$$

SNPs with largest effects get most importance (“weight”)

- SNPs with small (~ 0) effects do not contribute

Polygenic score (PGS, or PRS: “polygenic risk score”): Individual-level genetic susceptibility for a trait

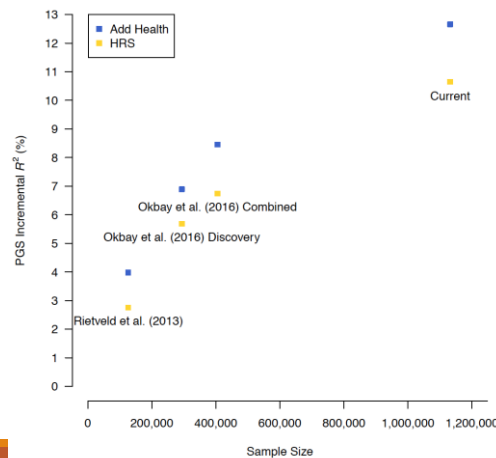
- For example, a high value on the PGS for EA indicates you have a high chance of attaining a high level of education
- Pitfall: PGS combines several (many different!) biological mechanisms (interpretation of effect difficult)

Improvement of PGS for EA over time

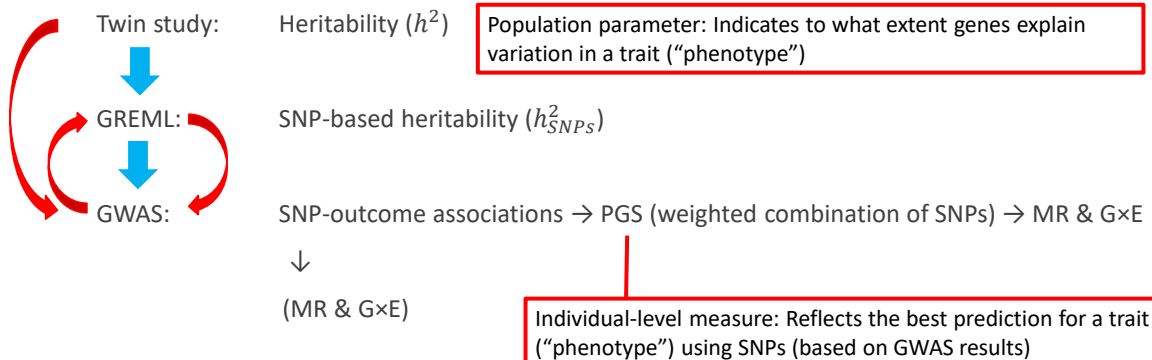
Whereas a SNP explains only 0.02%, the PGS for EA currently explains >10%!

PGSs with such large explanatory power are interesting for economist to include in empirical models

- To identify causal effect (i.e., “Mendelian randomization”, “MR”)
- To investigate the interplay between nature (genes) and nurture (environment) (“GxE”)



Schematic overview methods



Today's agenda


Mendelian randomization

- The value of randomization for estimating causal effects
- How genes can help to estimate causal effects
 - Pitfall: The “pleiotropic” working of genes

G×E interaction analysis

- Pitfall: Distinguishing G×E correlation from G×E interaction
- The policy-relevance of G×E interaction studies

Remember your individual assignment (Section 6): “...Also sketch two interesting directions (provide concrete applications) for further research which could be followed if robust associations between genetic variants and personal income are identified. The first direction should relate to Mendelian randomization (Lecture 5) and the second direction should relate to G×E interaction analysis (Lecture 5). Discuss what valuable knowledge can be obtained when following these two research directions.”



Economics & Genetics Lecture 5

MAIN QUESTION 1



Randomization

Example: Estimate the effect of smoking on health

- We know your health (y_i) and how much you smoke (x_i , e.g., 0, 1, 2, ... cigarettes)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

“Exogeneity assumption”, $E(\varepsilon_i | x_i)$, important for the interpretation of the OLS estimates

- Association vs. Causation

We estimate that β_1 is negative and significantly different from zero...

- Can we conclude that smoking reduces health?
- What is included in error term ε_i (socio-economic status, living circumstances, working environment, etc.)?

Randomization

The trick of randomization: We randomly distribute cigarettes, and force individuals to smoke them!

- Again, we know your health (y_i) and how much you smoke (x_i , e.g., 0, 1, 2, ... cigarettes)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Each person still different in terms of ε_i , but randomization ensures that ε_i is the same on average among those smoking 0, 1, 2, ... cigarettes

- “Exogeneity assumption”, $E(\varepsilon_i | x_i)$, satisfied!
- β_1 can be interpreted as a causal effect, since all other characteristics are, on average, the same

With randomization, we tackle omitted variables bias and reverse causality issues

Let's randomize?

Some limitations:

- Ethics! (randomly assigning the number of cigarettes, whether someone should go to school, etc.)
- Even if you assign exposure x_i , people may not comply (do smokers really stop if you say they shouldn't?)
- Experiments have usually limited duration: Focus on short-term effects
- Estimates are within a selected sample of people (those choosing to take part in an experiment)

Mendelian Randomization

Randomization using genes (i.e., SNPs, PGSs)

- "Natural experiment" (like twin studies)

Mendel's law of "Independent Assortment" (Lecture 2): The inheritance of one pair of genes is independent of the inheritance of the other pair

- Genes are randomly allocated (and fixed at conception), given the genes of the parents (reasoning can be extended to multiple generations back → genetically homogeneous populations)
- No ethical concerns, and neither concerns about compliance, short-term focus, or generalizability

Example motivated by today's literature: The effect of smoking intensity (x_i , the number of cigarettes per day) on BMI (y_i)



MR - Intuition

We need a genetic variable (SNP/PGS) that explains smoking intensity and nothing else

- An "exogenous" variable (a so-called instrumental variable)
- Of course, this SNP may have an effect on BMI y_i through smoking intensity x_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i^y$$

$$x_i = \gamma_0 + \gamma_1 SNP_i + \varepsilon_i^x$$

If you regress x_i on SNP_i , you get γ_1 ("first stage")

$$y_i = \beta_0 + \beta_1(\gamma_0 + \gamma_1 SNP_i + \varepsilon_i^x) + \varepsilon_i^y = (\beta_0 + \beta_1 \gamma_0) + \beta_1 \gamma_1 SNP_i + (\beta_1 \varepsilon_i^x + \varepsilon_i^y)$$

If you regress y_i on SNP_i , you get $\beta_1 \gamma_1$ (reduced-form)

Divide reduced-form by first stage: $\beta_1 \gamma_1 / \gamma_1 = \beta_1$

MR – 3 assumptions

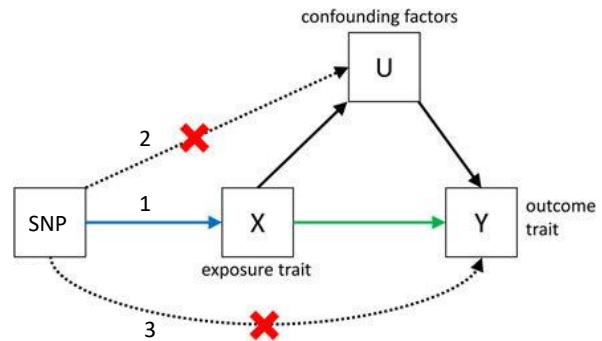
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i^y$$

$$x_i = \gamma_0 + \gamma_1 SNP_i + \varepsilon_i^x$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i^y (+\beta_2 SNP_i)$$

1. Relevance: $\gamma_1 \neq 0$
2. Independence: $E(SNP_i \varepsilon_i^y | x_i) = 0$
3. Exclusion: $\beta_2 = 0$

Bias if assumptions are not satisfied!



The attractiveness of MR

MR is instrumental variable (IV) regression with genetic variants as IV

- GWAS results make it straightforward to select genetic variants that are robustly associated with a treatment of interest (x_i) [Relevance assumption]
- For example, GWAS on smoking behavior identified SNPs related to nicotine *dependency*
- Genetic variants are randomly distributed at conception, conditional on population stratification variables or family-specific effects; Therefore, it is plausible that the independence assumption holds [Independence assumption]

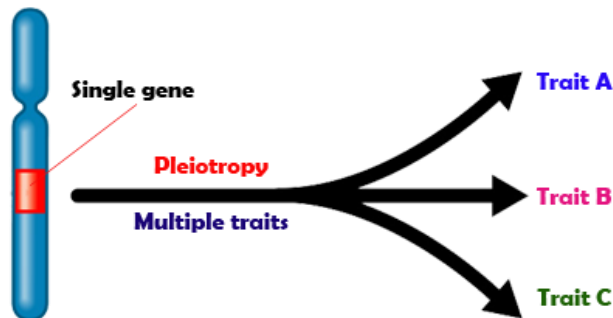
Example based on today's literature: SNP (rs12914385) in CHRNA3 gene (nicotine receptor)

- OLS: Effect of #cigarettes per day on BMI is 0.05 ($p < 0.001$)
- MR: Effect of #cigarettes per day on BMI is -0.24 ($p < 0.001$)

The attractiveness of MR?

But what about the exclusion restriction?

- SNPs may not only influence y through x , but also through other (unobserved) pathways or directly
- Difficult to argue convincingly due to limited knowledge about biological function (e.g., pleiotropy)



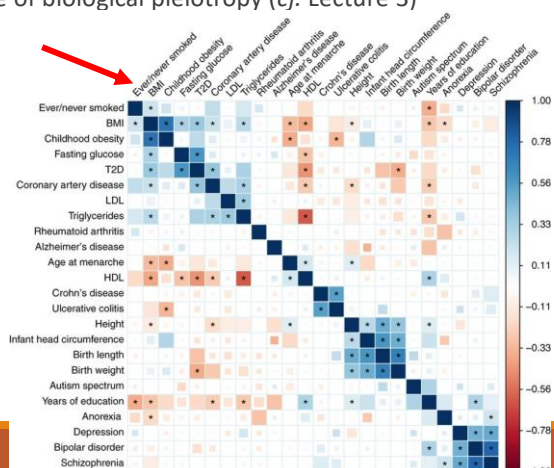
For example: If a SNP not only influences nicotine dependency, but also your educational attainment

The attractiveness of MR?

Genetic correlations do not necessarily imply biological pleiotropy but if genetic correlation $\neq 0$ then you also can't exclude the existence of biological pleiotropy (cf. Lecture 3)

Source: Bulik-Sullivan et al. (2014),
"An atlas of genetic correlations
across human diseases and traits."

For example: A significant genetic
correlation between smoking and
years of education (cf. Lecture 3)



Dealing with pleiotropy

Rapid developments, Pleiotropy-robust Mendelian Randomization (Van Kippersluis and Rietveld, 2017) provides ex-ante correction procedure

- Synthesizes two streams in econometrics to tackle pleiotropy problem in Mendelian Randomization studies

1. Testing the exclusion restriction in sample for which there is no first stage effect (on theoretical ground, you do not expect an effect of the SNP on treatment x_i in this sample: $\gamma_1 = 0$)

$$\text{First stage: } x_i = \gamma_0 + \gamma_1 \text{SNP}_i + \varepsilon_i^x$$

$$\text{Reduced form: } y_i = (\beta_0 + \beta_1 \gamma_0) + \beta_1 \gamma_1 \text{SNP}_i + (\beta_1 \varepsilon_i^x + \varepsilon_i^y)$$

$$\text{Reduced form: } y_i = (\beta_0 + \beta_1 \gamma_0) + 0 + (\beta_1 \varepsilon_i^x + \varepsilon_i^y)$$

$$\text{Reduced form: } y_i = (\beta_0 + \beta_1 \gamma_0) + (\beta_1 \varepsilon_i^x + \varepsilon_i^y) + (\beta_2 \text{SNP}_i)$$

2. Use the estimate obtained in this subsample to directly control for the effect (" β_2 ") of the instrument (i.e., SNP) on the outcome (y_i)

Example 1

The effect of smoking intensity on BMI, with SNP (rs12914385) in CHRNA3 gene (nicotine receptor) as IV

- OLS: Effect of #cigarettes per day on BMI is 0.05 ($p < 0.001$)
- MR: Effect of #cigarettes per day on BMI is -0.24 ($p < 0.001$)

A sample in which there can be no effect of this SNP on the #cigarettes per day? (no first-stage effect)

- Non-smokers! (SNP has no relationship with smoking initiation)
- Among non-smokers: Effect of SNP on BMI (β_2) is small and insignificant! (Exclusion restriction seems to hold)
- PRMR: Effect of #cigarettes per day on BMI is -0.26 ($p = 0.003$)

Caveat:

- First stage should be truly zero in subsample in which direct effect is estimated (something we cannot test)

Example 2

The effect of prostate cancer on self-reported health

We know some SNPs that influence the development of prostate cancer

What would be a good “zero-first-stage” group to test the relation between these SNPs and self-reported health?

One possibility: The subsample of females

- OLS: Prostate cancer negatively associated with self-reported health ($-0.17, p < 0.001$)
- MR: Prostate cancer negatively associated with self-reported health but insignificantly ($-1.26, p = 0.54$)
- Among females, SNPs negatively associated with self-reported health ($-0.01, p = 0.002$)
- PRMR: Prostate cancer positively associated with self-reported health but insignificantly ($4.51, p = 0.10$)

The popularity of MR

<https://www.youtube.com/watch?v=LoTgfGotaQ4>



However...

Pleiotropy is really problematic for MR studies → We need more biological knowledge to assess whether the exclusion restriction holds (especially for behavioral applications)

- Power larger if first stage effect γ_1 is larger: Polygenic scores (PGSs) preferable over individual SNPs → However, exclusion restriction is even more difficult to assess as PGSs capture different biological mechanisms

In the mean time: Methods like PRMR may help (but finding subsamples is really difficult)

- Or perform other tests to gauge the plausibility of validity of the instrument (*Presentation Paper 3*)

Answering the first main question

How can we use GWAS results to identify causal effect in economic models? → Mendelian Randomization!

- Great potential, but one big caveat: Violation of the exclusion restriction (due to pleiotropy)
- Proper biological understanding needed, or methods that accounts for violation of exclusion restriction
- PRMR is such a method, but (sub)sample needed in which first stage effect is truly zero

Remember your individual assignment (Section 6) : “...Also sketch two interesting directions (provide concrete applications) for further research which could be followed if robust associations between genetic variants and personal income are identified. The first direction should relate to Mendelian randomization (Lecture 5) ...”

- For example, an otherwise non-genetic study estimating the effect of income on outcome “Y” using a SNP associated with income as IV (elaborate on potential and caveats)

Economics & Genetics

Lecture 5

MAIN QUESTION 2



GxE interactions


The classical twin study *decomposes* trait variance (cf. Lecture 1):

$$Var(y) = Var(a) + Var(c) + Var(e)$$

y – trait (e.g., educational attainment), a – additive genetic effects, c – common (family) environment, e – unique environment

(Narrow-sense: *Only additive genetic effects*) heritability $h^2 = Var(a) / Var(y)$

Importantly:

- The variance components are assumed to be independent (very strong assumption)
 - We do not measure the environment: It's everything we cannot attribute to genes
- 

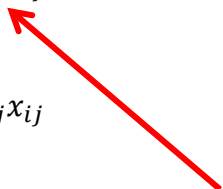
G×E interactions

- How can we be sure that GWAS findings not only reflect the environment? (Bliss, 2018: “Social by Nature: The Promise and Peril of Sociogenomics”)
- Use polygenic scores to understand the interplay between nature and *measured* nurture

◦ GWAS:
$$y_i = \mu + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i \quad (1 \text{ SNP per regression})$$

◦ PGS:
$$PGS_i = \sum_{j=1}^J \hat{\beta}_j x_{ij}$$

β_j may capture biological and environment mechanisms (remember e.g., the grasshoppers and the “is parenting overrated” debate from Lecture 1)



Gene-environment interaction

Gene–environment interaction (G×E) studies analyze how genetic effects may be different in different environments

- Do genes influence traits only in certain environments?
- Can environments help to moderate genetic susceptibilities?

Examples

- Are only children who grow up in advantageous environments able to reach their genetic potential in terms of cognitive skills?
- Does an advantageous environment cushion genetic susceptibility to risky health behaviors? (e.g., smoking)

Relevance: Public policy cannot change genes, but can manipulate environments! (i.e., induce environmental variation)

Pitfall: Gene-environment correlations

Gene-environment interactions are different from gene-environment correlations (which may be intertwined in the PGS!)

- Reactive “gene-environment correlation”: The environment responds differently to individuals with different genes (e.g., giving books to kids who enjoy reading, different treatment males/females)
- Active “gene-environment correlation”: Individuals with different genes seek out or create different environments (e.g., those with light skin avoid sunny environments, MZ twins reared apart often very similar)
- Passive “gene-environment correlation”: Parental genes affect both a child’s genes and the rearing environment (e.g., social class)

Setup G×E model

To avoid bias from gene-environment correlation, we should be sure that “G” (the PGS) is independent (“exogenous”) from “E”

- Therefore: Focus on unique “exogenous shocks” in the environment (“random” events, such as policy changes/discontinuities)

$$y_i = \alpha + \beta_1 \cdot PGS_i + \beta_2 \cdot E_i + \beta_3(PGS_i \times E_i) + \varepsilon_i$$

y_i indicates one’s educational attainment

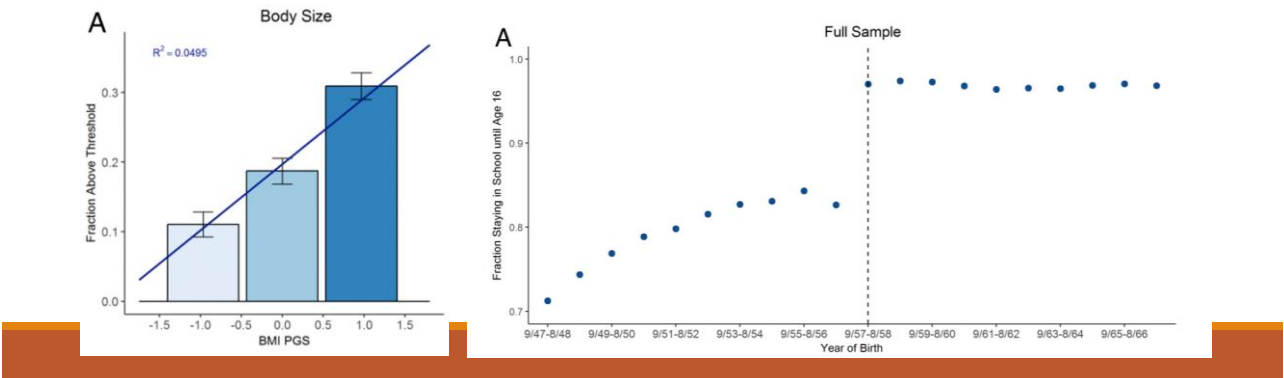
PGS_i is the genetic predisposition to achieve high education

E_i is a binary indicator of being affected by a policy to increase the compulsory schooling age

Example 1

Barcellos, Carvalho & Turley (2018): “Education can reduce health differences related to genetic risk of obesity”

- G: PGS for BMI (with explanatory power, Panel A)
- E: Compulsory schooling age reform in the United Kingdom (“ROSLA 1972”, Panel B)

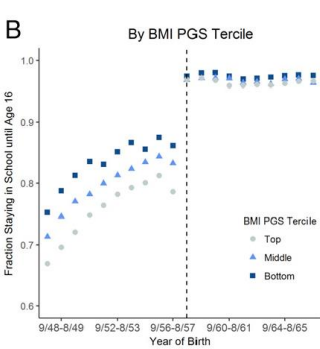


Example 1

	Body size
Interaction with BMI PGS	
BMI PGS × Edu16	−0.057*** (0.011)
Edu16	−0.060* (0.035)
BMI PGS	0.124*** (0.010)
P value for H ₀ : no effect of education	7.97 × 10 ^{−10}

Main effect Edu16 negative &
Interaction effect negative:
Effect schooling reform largest
for those with highest PGS
values

The schooling reform reduced
the gap in unhealthy body size
between those in the top and
bottom terciles of genetic risk
of obesity from 20 to 6
percentage points



Policy implications

Barcellos et al. (2018): “Our results challenge the notion of genetic determinism and underscore the role that social policy can have in mitigating possible health differences arising from genetic background”

- The effect of the PGS for BMI can be moderated
- A “good environment” may help those with a high genetic predisposition for obesity

Disclaimer: Gene-based prediction at individual level is almost impossible (Lecture 4), so we can only conclude something about (sub)population averages

- Still: The effectiveness of policies can be evaluated for certain genetic subgroups in the population
- Pointer to Lecture 6: Do you want governments to perform these kind of gene-based policy evaluations?

Example 2

Slob & Rietveld (2020): “The moderating impact of the genetic predisposition to smoking behaviour on the response to tobacco excise taxes”

- Smoking is a risky health behavior, but despite heavy tobacco excise taxes many people remain smoking
- Is the effect of smoking taxation dependent on someone’s genetic susceptibility for nicotine dependency?

G = PGS for smoking

E = Tobacco excise taxes imposed in US states

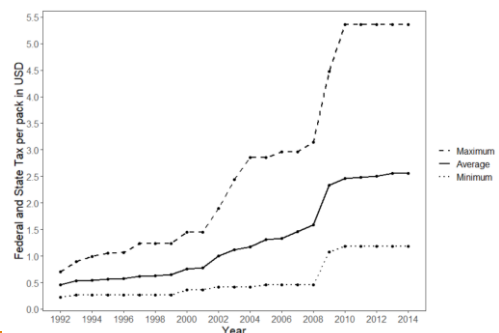


Figure 1: Tobacco excise taxes levied per pack of 20 cigarettes.

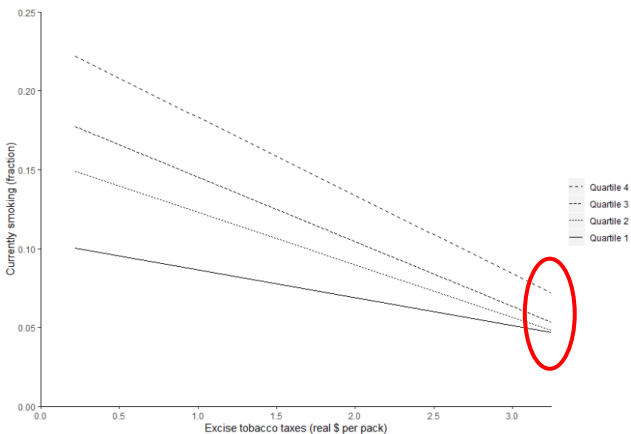
Example 2

Individuals with a high PGS value for smoking are less likely to smoke when tobacco excise taxes are high:

Table 2: Results of the regressions explaining an individual’s current smoking status.

	(1)	(2)
-Log(Tax)	0.066***	0.066***
	(0.005)	(0.005)
<i>PGS</i> _{Smoking initiation}	0.038***	0.036***
	(0.004)	(0.004)
-Log(Tax) × <i>PGS</i> _{Smoking initiation}		0.012***
		(0.003)

Example 2



The higher tobacco excise taxes, the lower the percentage of smokers in a state

Even in states with the highest tobacco excise taxes, there are smokers

- Are some smokers “insensitive” to further tax increases?

Figure 2: The relationship between excise tobacco taxes and the likelihood of smoking in each quartile of the distribution of the polygenic score for smoking initiation.

Example 2

G×E interaction effects for the intensity of smoking (*not for smoking cessation*)

Table 3: Results of the regressions explaining an individual's smoking intensity.

	Full sample		Subsample of current smokers	
	(1)	(2)	(4)	(5)
-Log(Tax)	1.585*** (0.115)	1.587*** (0.115)	3.481*** (0.394)	3.428*** (0.386)
$PGS_{\text{Smoking intensity}}$	0.358*** (0.0532)	0.318*** (0.0482)	1.075*** (0.159)	0.962*** (0.154)
$-\text{Log(Tax)} \times PGS_{\text{Smoking intensity}}$		0.204** (0.0591)		0.376 (0.188)

Explanations and implications

Individuals with the highest genetic predisposition to smoking respond most *strongly* to taxes

- Shouldn't you expect that those with a strong genetic predisposition to smoking are least responsive to changes in tobacco excise taxes?
- PGS for smoking relates to processing of rewards in the brain: (monetary) rewards of not smoking increase when tobacco prices increase?

Taxes alone are not going to make all individuals stop smoking

- Addiction vs. recreational use
- E.g., Behavioral therapies may be needed

Again: G×E studies challenge the notion of genetic determinism

- Environmental conditions can moderate the effect of genetic predispositions

Answering the second main question

How can polygenic scores be used to understand the interplay between nature and nurture?

- Avoid bias from reactive, active, or passive correlation between genes and environments
- G×E studies exploiting exogenous variation may provide pointers about whether the effect of policies is different for genetically different individuals

Remember you individual assignment (Section 6) : “...Also sketch two interesting directions (provide concrete applications) for further research which could be followed if robust associations between genetic variants and personal income are identified ... the second direction should relate to G×E interaction analysis (Lecture 5).”

- For example, what would be an interesting “exogenous” environment you want to interact a PGS for individual income with to explain outcome “Y” (elaborate on potential and caveats)

What comes next?

Tutorial: No new exercise set

Lecture 6

- Are we already realizing the promises of genoconomics?
- Is there any value in genetic profiling for socio-economic traits?
- How could policy makers “deal” with “genetic luck”?

Presentation 3 (Zoom)

The presenting group will be chosen randomly
(<https://www.random.org/integers/>)

- Format: 15 minutes + discussion
- The other teams that reviewed the same paper take the lead in the discussion

Please to Zoom now:

- <https://zoom.us/join>, Meeting ID: 972 4750 2187, Password: 902246
- <https://eur-nl.zoom.us/j/97247502187?pwd=d3hXeXd2WjliR1h5ZHZuL2E3SERSZz09>



Economics & Genetics Lecture 5

DR. NIELS RIETVELD, NRIETVELD@ESE.EUR.NL

OFFICE MANDEVILLE T18-29

