Universität Zürich, BOEC0344

Summary Intermediate Econometrics

Emanuel Nussli

FS 2020

1 OVB

Violated assumptions: $E[\epsilon \mid X_{1i}, X_{2i}] = 0$ and correct model specification.
True model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \epsilon_i$
Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$
OVB: $E[\hat{\beta}_1 \mid X] = \alpha_1 + \alpha_2 \cdot \rho$ with ρ being the regression coefficient of a regression of X_2 on X_1.

Resulting Bias:
overestimated: estimate is larger than the true parameter (in absolute values): $|\hat{\beta}_1| > |\alpha_1|$
underestimated: estimate is smaller than the true parameter (in absolute values): $|\hat{\beta}_1| < |\alpha_1|$

Positive and negative bias:
positive bias: $\alpha_2 \cdot \rho \in (0, \infty)$
negative bias: $\alpha_2 \cdot \rho \in (-\infty, 0)$

2 Panel Data

Panel data with k regressors: $(X_{1it}, X_{2it}, ..., X_{kit}, Y_{it}), i \in (1, .., n), t \in (1, ..., T)$
balanced panel: no missing observations, all variables are observed for all entities and time periods

Central ideas of panel data regressions: we control for factors that:
1. vary across entities but not over time
2. are unobserved and therefore not included in our regression
3. could cause omitted variable bias if omitted

Estimation strategies:
Changes specification:
$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 Z_i + u_{it}$ with Z_i being a factor that does not change over time and that is not observed but causes OVB. We eliminate Z_i by using $T = 2$ years.
$Y_{it} - Y_{it+n} = \beta_0 + \beta_1 X_{1it} + \beta_2 Z_i + u_{it} - (\beta_0 + \beta_1 X_{1it+n} + \beta_2 Z_i + u_{it+n}) = \beta_1 \cdot (X_{1it} - X_{1it+n}) + u_{it} - u_{it+n}$

consequences:
1. no intercept
2. new error term neither correlated with X_{1it+n} nor with X_{1it}
3. Difference equation can be estimated by OLS

Fixed Effects Regression: $T > 2$
$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 Z_i + u_{it}$ with $(\beta_0 + \beta_2 Z_i)$ not changing over time. We define $(\beta_0 + \beta_2 Z_i) = \alpha_i$
That means that we have an intercept for each $state_i$. That means that we have i different intercepts and i individual lines with the same slope β_1.

Binary Regressor way of wiriting Fixed Effects Regressions:
$Y_{it} = \beta_0 + \gamma_1 X_{1it} + ... + \beta_{1it} + u_{it}$ with $X_{1it} = 1$ if entity is entity 1, = 0 otherwise
Do not forget that you need a base dummy in order to prevent perfect multicollinearity.

Entity-demeaned OLS regression:
1. Start with FE regression: $Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$
2. $\frac{1}{T}\sum_{t=1}^{T} Y_{it} = \frac{1}{T}\sum_{t=1}^{T} X_{it} + \frac{1}{T}\sum_{t=1}^{T} u_{it} \rightarrow$ mean of x always on regression line.
3. Devation from means: $Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it} + u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}$ or

$\tilde{Y_{it}} = \beta_1 \tilde{X_{it}} + \tilde{u_{it}} \rightarrow$ very easy in STATA, only one command. For STATA commands, see other file. For $i = 1$ and $t = 1982$, $\tilde{Y_{it}}$ is the difference between the fatality rate in Alabama in 1982 and the average of the fatality rate in Alabama over all the years.

Important Stuff

- Helpful: Areg command automatically demeans the data and estimates via entitiy-demeaned OLS.
- Do not report the intercept $\rightarrow$ arbitrary, allows the mean change in $Y \neq 0$.
- Clustered SEs: *areg y x, absorb(state) cluster(state)* or *xtreg y x, fe vce(cluster state)*
- confounding variable: influences both dependent and independent variables

Up until now, we discussed state fixed effects. Those are effects that vary across states but are constant across time. But of course, that concept needs to be extended to time fixed effects.

Time fixed effects: effects that are constant across states but vary across time. That results in the following (equivalent) regression specification:
$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$ with λ_t being the fixed effects for the periods $t \in (1, ..., T)$.
HOW TO:
"year-demeaned" OLS:
1. Deviate (Y_{it}, X_{it}) from year averages.
2. Estimate OLS with that "year-demeaned" data. "T-1 binary regression" OL
1. create $B_2, ..., B_T$
2. Regress Y on X, $B_2, ..., B_T$ using OLS

State and Time Fixed Effects:
Some omitted variables are constant over time but vary across entities (state FE) while others are constant across entities but vary over time (time FE), we need both state and time fixed effects.

Estimation:
FE formulation: $Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_i + u_{it}$
Binary regressor formulation: $Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + ... + \gamma_n DN_i + \delta_2 B2_t + ... + \delta_T BT_t + u_{it} \rightarrow$ base variables left out.

Assumptions and asymptotic normal distribution for FE regressions (only for state FE for neatness)
1. $E[u_{it} \mid X_{i1}, ..., X_{iT}, \alpha_i] = 0$ meaning that the error term has mean zero, given α_i and the entire history of Xs for that state. $\rightarrow$ no lagged effects and no feedback from u to X. That means we dont want that u e.g depends on X_{it}. This assumption is violated if the current value of the error term is correlated with the past, the present or the future values of the independent variables.
2. $(X_{i1}, ..., X_{iT}, u_{i1}, ..., u_{iT}), i \in (1, ..., n)$ are i.i.d draws from their joint distribution. $\rightarrow$ random

sampling from population and doing it over time. That means that they are distributed identically but independently of the variables for another state. Observations are not required to be i.i.d over time!
3. and 4. $(X_{it}, u_{it}) : E[X_{it}^j] < \infty \; \forall \; j \in \{1, 2, 3, 4\}$ and $E[u_{it}^j] < \infty \; \forall \; j \in \{1, 2, 3, 4\}$ and no perfect multicollinearity. The last two are the same as in standard linear regression.

Standard errors with FE:
u_{it} can be autocorrelated and heteroskedastic. It can be autocorrelated and assumption 2. allows the X_i to be correlated within the different states. Autocorrelation violates our assumption 1. and we have to address that. We need to address that which is why we use clustered SE, which allow correlation in clusters but not across clusters. Autocorrelated errors do not offer as much information as not autocorrelated errors. We therefore use HAC (heteroskedasticity and autocorrelation consistent standard errors), and clustered SE belong in that category. If there are more than 40 clusters, inference can be done as usual.

Limitations and challenges
1. Need variation in X within states

2. Lag effects are important
3. We lose a lot of interesting information
4. Role of measurement error might increase

3 Regression with a Binary Dependent Variable

Linear Probability Model: $Y_i = \beta_0 + \beta_1 X_i + u_i$
What does it mean when $\hat{Y} = 0.26$ if Y_i is binary. It means that for 26% of the Xs, it holds: $Y_i = 1$.
$E[Y \mid X] = 1 \cdot P(Y = 1 \mid X) + 0 \cdot P(Y = 0 \mid X) = P(Y = 1 \mid X)$
Using assumption of uncorrelated errors, we get: $E[Y_i \mid X_i] = E[\beta_0 + \beta_1 X_i + u_i \mid X_i] = \beta_0 + \beta_1 X_i = P(Y = 1 \mid X)$
Marginals: linear, as $\beta = \frac{P(Y=1|X=x+\Delta x)-(P(Y=1|X=x)}{\Delta x}$: exact with linear models, aprroximative with nonlinear models such as probit and logit.
→ compare difference in p.p to average: difference = $\Delta p.p$. What is that in relative terms? $\frac{\Delta p.p - average p.p}{average p.p}$
→ use heteroskedastic-robust standard errors

Probit model: $P(Y \mid X) = F(\beta_0 + \beta_1 X_i) = \phi(\beta_0 + \beta_1 X_i)$. Works also if the model is multivariate: $P(Y \mid X_1, X_2) = \phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ etc. The same goes for the Logit and the LPM.
Logit model: $P(Y \mid X) = F(\beta_0 + \beta_1 X_i) = \frac{exp(\beta_0+\beta_1 X_i)}{1+exp(\beta_0+\beta_1 X_i)} = \Lambda(\beta_0 + \beta_1 X_i)$
→ the Logit and Probit models are both estimated via MLE, which produces consistent results.

Where does the Probit model come from?
Latent variable model: $\tilde{Y} = \beta_0 + \beta_1 X + u$ with $\tilde{Y}$ being a latent (unobserved or partially unobserved variable) and $Y = 1 \; \forall \; \tilde{Y} > 0$. Hence: $P(Y = 1 \mid X) = P(\tilde{Y} > 0 \mid X) = P(\beta_0 + \beta_1 X + u > 0 \mid X) = P(u > -(\beta_0 + \beta_1 X) \mid X) = 1 - \phi(-(\beta_0 + \beta_1 X)) = \phi(\beta_0 + \beta_1 X)$

Important: β_1 is the effect on the z-score of a unit change in X_1, if X_2 is held constant.

Marginals: continous X
$\frac{\delta P(Y=1|X)}{\delta X} = \beta_1 \cdot \phi(\beta_0 + \beta_1 X_i)$ → probit model. Depends on all variables. Mostly, you enter the mean or median for the missing values. Same goes for the Logit. Another method is to fix x_j and let x_i goes from its min to max and plot the change of the marginal effect of x_i across the entire range.
$\frac{\delta P(Y=1|X)}{\delta X} = \beta_1 \cdot \frac{exp(\beta_0+\beta_1 X_i)}{[1+exp(\beta_0+\beta_1 X_i)]^2}$ → logit model
STATA: *probit y x, r* and then *margins,dydx(*)* or *margins, dydx(*) atmeans*. The first options gives the average margin and the second the margin at the average of X. The first one is usually very similar to the LMP.

Marginals: discrete X

Probit: $\phi(\beta_0 + \beta_1) - \phi(\beta_0)$
Logit: $\Lambda(\beta_0 + \beta_1) - \Lambda(\beta_0)$

Nonlinear least squares
It is possible to estimate the nonlinear models and asymptotically consistent and normally distributed but produce larger variances than the MLE and are therefore rarely used in practice. Formally, the procedure for the probit is a follows: $\min_{b_0,...,b_k} \sum_{i=1}^{n} [Y_i - \phi (b_0 + b_1 X_{1i} + ... + b_k X_{ki})]^2$

Maximum Likelihood Estimator for Bernoulli Trials
$P(Y_1 = y_1) = p^{y_1} \cdot (1-p)^{1-y_1} \rightarrow$ first trial
Joint density of $(Y_1, Y_n) \overset{iid}{=} [p^{y_1} \cdot (1-p)^{1-y_1}] \cdot [p^{y_2} \cdot (1-p)^{1-y_2}] \cdot .... \cdot [p^{y_n} \cdot (1-p)^{1-y_n}] = p^{\sum_{i=1}^{n} y_i} \cdot (1-p)^{n-\sum_{i=1}^{n} y_i} = f(p; Y_1, ..., Y_n)$
$ln(f(p; Y_1, ..., Y_n)) = (\sum_{i=1}^{n} Y_i) \cdot ln(p) + (n - \sum_{i=1}^{n} Y_i) \cdot ln(1-p)$
$\frac{\delta ln(f(p;Y_1,...,Y_n))}{\delta p} \overset{!}{=} 0 = (\sum_{i=1}^{n} Y_i) \cdot \frac{1}{p} + (n - \sum_{i=1}^{n} Y_i) \cdot \frac{-1}{1-p}$
$\frac{\bar{Y}}{1-\bar{Y}} = \frac{\bar{p}}{1-\bar{p}} \rightarrow \hat{p} = \bar{Y} \rightarrow$ as. normally distributed and inference as usual.The MLE estimator is the most efficient estimator of p and much stronger than the Gauss-Markov Theorem. To emphasize that we need large n, STATA uses the t-statistic instead of the z-statistic.
Important: The likelihood function is always smaller or equal one and its only 1 if it predicts with certainty what you will draw $\rightarrow$ only possible for $p \in \{0, 1\}$. That is why the log-likelihood function $f(\cdot) \in (-\infty, 0]$.

Measures of Fit in Logit and Probit
1. Fraction correctly predicted: $P(Y_i) > 0.5$ if $Y_i = 1$ and vice versa. Problem: Does not depend on the quality of the prediction as 51% is treated the same as 99%.
2. Pseudo R^2. It compares the fit of MLE function with all regressors to the situation with none regressors.
$pseudo - R^2 = 1 - \frac{ln(f_{maxprobit})}{ln(f_{max_Bernoulli})}$. The R^2 is not useful here as it is not possible to get an R^2 of one because not all data can lie in a line as the dependent variable is discrete. It would be possible if the regressors were also discrete. R^2 is therefore not useful here.

Percentage Points vs. Percentage effect
Effect of secondary schooling: 0.000646 - 0.000488 = 0.000158. That is 0.0158 p.p but an increase of $\frac{0.000158}{0.000488}$, which is by 32%.

4 Instrumental Variables

Threats to **internal** validity: problems with validity within our model
1. OVB from an unobserved variable
2. Simultaneous causality bias / reverse causality
3. Mesurement error bias (X measured with error)
Those problems lead to $E[u \mid X] \neq 0$ and IV can solve all of those problems!

$Y_i = \beta_0 + \beta_1 X_i + u_i$ while IV breaks X into two parts: one part is correlated with u and one part is not correlated with u. The information about X that are uncorrelated with u are obtained via an instrument.
endogenous variable: variable that is correlated with u_i
exogenous variable: variable that is uncorrelated with u_i

2SLS:
1. $cor(Z_i, X_i) \neq 0$
2. $cor(Z_i, u_i) = 0$
First stage: $X_i = \pi_0 + \pi_1 Z_i + v_i$ (1) by OLS, obtain predicted values $\hat{X}$
Second stage: regress Y_i on $\hat{X}_i$ and obtain β^{2SLS}, an asymptotically consistent estimator.
We can also use the **reduced form** to estimate the causal effect: $Y_i = \gamma_0 + \gamma_1 Z_i + w_i$

Getting the right estimator: $Y_i = \beta_0 + \beta_1 Z_i + u_i$ and $cov(Y_i, Z_i) = cov(\beta_0 + \beta_1 Z_i + u_i, Z_i) \overset{linearity}{=} cov(\beta_0, Z_i) + \beta_1 \cdot cov(X_i, Z_i) + cov(u_i, Z_i) = \beta_1 \cdot cov(X_i, Z_i)$. Therefore, we have: $\beta_1 = \frac{cov(Y_i, Z_i)}{cov(X_i, Z_i)}$
reduced form: $\gamma_1 = \frac{cov(Z_i, Y_i)}{var(Z_i)}$
First stage: $\pi_1 = \frac{cov(Z_i, X_i)}{var(Z_i)}$
Therefore: $\beta_1 = \frac{\gamma_1}{\pi_1}$. Intuiton: §
plim $\beta_1^{2\hat{S}LS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{cov(Y,Z)}{cov(X,Z)} = \beta_1 \longrightarrow$ consistent and $\hat{\beta}_1^{2SLS} \sim N(\beta_1, \sigma^2_{\hat{\beta}_1^{2SLS}})$

Inference is done as ususal (when large samples are available). The standard errors are not correct though, as they do not take into account that the first stage was also estimated. STATA computes the correct SEs (and we use heteroskedastic robust SEs as usual).

The general (multivariate) IV Regression Model
1. multiple endogenous regressors $(X_1, ..., X_n)$

2. multiple exogenous variables $(W_1, ..., W_n)$ as controls
3. multiple instruments $\rightarrow$ reduce σ_{2SLS} and increase R^2 of the first stage which allows for more variation in $\hat{X}$

Identification
$\beta_1, ..., \beta_k$ are said to be:
exactly identified if $m = k$ with $m :=$ number of instruments
over-identified if $m > k$. If the effects are linear, we can test if the instruments are valid. Otherwise, there are multiple good instruments.
under-identified if $m < k$

That yields in: $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \beta_{k+1} W_{1i} + ... + \beta_{k+r} W_{ri} + u_i$. We call the W_i the included exogenous regressors (controls) and the Z_i the excluded exogenous variables (instruments).

Control variables
Quite often, instruments Z are only exogenous when having introduced a control variable W. Formally: $E[u_i \mid W_i, Z_i] = E[u_i \mid W_i] \rightarrow$ no dependence on Z_i. In that case, the W_i **do not have to be exogenous**.
Important: always regress on **all** of the controls and **all** of the instruments.

LATE; local average treatment effect: IVs are of local nature $\rightarrow$ no linear effect, the marginals change with our input.

Checking Assumptions
1. Relevance: $X_i = \pi_o + \pi_1 Z_{1i} + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + ... + \pi_{m+k} W_{ki} + u_i \rightarrow$ at least 1 $\pi_i \in (\pi_1, ..., \pi_m) \neq 0$. **Important**: There must be at least one instrument per X and there must be different exogenous variation for each X to prevent multicollinearity.
weak instrument: $\pi_1, ..., \pi_m$ are zero or close to zero $\rightarrow$ weak instruments are not approximately normally distributed, not even asymptotically. To establish a rule for weakness, regress X_i on $Z_1, ..., Z_m, W_1, ..., W_k$ and test with an F-test whether $Z_1, ..., Z_m$ are jointly insignificant. If F-stat > 10, then you are okay. STATA: *test instrument* or if $m = 1$: $F_{stat} = (t_{stat})^2$
More formally (w.o proof): $E[\hat{\beta}_1^{2SLS}] - \beta_1 \approx \frac{(\beta_1^{OLS} - \beta_1)}{(E[F]-1)}$. That means that the bias of 2SLS is, relative to the bias of OLS, just over 0.1, which is small enough in most applications.
2. Exogeneity of instrument: $cor(Z_i, u_i) = 0, ..., cor(Z_{mi}, u_i) = 0$. If $m > k$ and the effects are not heterogeneous (no LATE), exogeneity can be tested.
Showing that exogeneity fails (we can not show that exogeneity is fulfilled but we can show when it is violated):
$W_{1i} = \beta_0 + \beta_1 Z_i + \beta_2 W_{2i} + ... + \beta_r W_{ri} + u_i \rightarrow$ if $\beta_1 \neq 0$, exogen. failed.
3. The X_i,W_i,Z_i and Y_i have finite fourth moments.

What to do if you have weak instruments
1. Find better instruments, duh
2. Drop the weak ones as that improves the F-statistic
3. Use liml (limited information maximum likelihood) and CIs with Anderson-Rubin and Moreira's Cond. Likelihood Ratio CIs → if we have $m > k$ and no LATE (homogeneity), we can (partially) test for exogeneity of the instrument. See J-Test in slides.

LHS-test for evidence against exogeneity
Regress: $W_i = \beta_0 + \beta_{Zi} + \beta_2 W_{2i} + ... + \beta_r W_{ri} + u_i$ → obviously not the control that is included to make sure that Z_i is exogenous. If β_1 is significant, the **Exclusion restriction** fails.

Measurement error:
$Y_i = \alpha + \beta \cdot \tilde{S}_i + e_i$ with m_i being **random** measurement error. $S_i = \tilde{S}_i + m_i$. That means: $E[m_i] = 0$ and $Cov(m_i, \tilde{S}_i) = 0$.
The coefficient on S_i is the following: $b = \frac{Cov(Y_i,S_i)}{Var(S_i)} = \frac{Cov(Y_i,\tilde{S}_i+m_i)}{Var(\tilde{S}_i+m_i)} = \frac{Cov(\alpha+\beta\cdot\tilde{S}_i+e_i,\tilde{S}_i+m_i)}{Var(\tilde{S}_i+m_i)} = \beta \cdot \frac{Cov(\tilde{S}_i,\tilde{S}_i)}{Var(\tilde{S}_i+m_i)} = \beta \cdot \frac{Var(\tilde{S}_i)}{Var(\tilde{S}_i)+Var(m_i)} \leq \beta$. That means that our biased estimate (given $Var(m_i) > 0$ will be $\beta \cdot x$, with $x \in [0, 1]$. That reduction of the effect, the bias towards 0 is called **attenuation bias**. Adding more controls makes this bias even worse.

How does one find good instruments?
1. Search for variables in another equation → supply, demand stuff.
2. Look for exogenous variation (as if randomly assigned) → those methods are very much the same thing

5 Matching

Idea: We want the effect of D on Y → take two very similar data points, except in X, pair them and compare their dependent variable Y. Repeat for all data points. You can causally estimate the effect of X in doing that.

Assumptions
1. Conditional Independence: $Y \perp D \mid X$: Peoples decision to be part of the treatment can be due to their observables X but on nothing else. → untestable assumption. Plausible if the analyst has "the correct Xs" (many and highly predictive ones).
2. Common support assumption: $0 < P(D = 1 \mid X = x) \equiv p(x) < 1 \; \forall \; x$ that are observed. We need a matching pair $\forall x \in X$. Testable, is a data requirement.

Discrete Variables: The Cell estimator
1. Make a table according to all possible covariates
2. Isolate the effect of the treatment for all combinations of covariates.
3. Aggregate the information of the cells into one number: the ATE: weigh the effect of the treatment with the frequency of the combination of covariates. → very non-parametric (most non-parametric possible) (no functional form) and very easy to estimate.

Continous Variables
Problem: No untreated people at $X = x$. Solution: Take people close to x and average them. That is a semi-parametric approach. We prefer not having to arbitrarily choose a function.

Methods of choosing the right data for the missing data points:
Generally: Trade-off of variance against bias. Remember the MSE: $MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\theta, \hat{\theta})]^2$. The bigger the bandwidth, the bigger the bias and the smaller the variance. In this case (attention as terms usually used as synonyms) variance is precision and bias accuracy.
1. Nearest neighbor matching: $\hat{Y}_0(x) = Y_0(x_{MIN}(x))$

2. Caliper Matching: $\hat{Y}_0(x_i) = \sum_{k \in C^0(p_i)} w_{ik} \cdot Y_0(x_i)$. $C^0(p_i)$ is the set of nearest neighbors of i which are treated. $w_{ik} \in [0, 1] with \sum_{k \in C^0(p_i)} w_{ik} = 1$.
k-nearest neighbors are the k nearest neighbors and the weights are linearly distributed.
Caliper is in terms of distance, i.e: $C^0(p_i) = \{k : |p_i - p_k| < \delta\}$ and the weights are also linearly

distributed, i.e $w_{ik} = \frac{1}{k}$.

→ kernel-based matching addresses the problem of equal weights as data closer to our point should - intuitively - be weighted with more importance. The weights are calculated as follows: $w_{ik} = \kappa(\frac{p_i - p_k}{\delta})$ with κ being the kernel function (Gaussian, Cosine, etc.) and δ the bandwidth parameter (variance of gaussian etc.).

Curse of Dimensionality
Suppose we have a lot of cont. variables $X_1, ..., X_n$. It is quite clear that matching with all those atttributes is quite difficult. Imagine having 100 variables concerning physical attributes; finding two people with all the "same" attributes in all but the treatment variable, is difficult. To still be able to match, we loosen our idea of "same" or "close" which will generate bad matches → curse or dilemma of dimensionality.

Propensity Score Matching
As matching in high-dimensional data sets would computationally be too extensive or just not possible, we introduce the propensity score.
Propensity score:

$$(Y_0, Y_1) \perp D \mid X \rightarrow (Y_0, Y_1) \perp D \mid p(x); p(x) = P(D = 1 \mid X = x)$$

The propensity score measures the probability of being in the treatment group given $X = x$.

Inverse Propensity Score Weighting
Has the effect of making the distributions of X for both groups (treated and untreated) the same.
Formally: $F(x \mid D = 0) = F(x \mid D = 1)$
Weights for untreated group $w_j = \frac{1}{1-p(x_j)}$
Weights for treated group $w_j = \frac{1}{p(x_j)}$ → large weight to unusual data points, given their treatment status.

How to get the Prop. Score $p(x)$
1. discrete data: $\forall\, X = x \rightarrow p(x) = \frac{n_{x,D=1}}{n_x}$ → STATA
2. continuous data: curse of dimensionality! → linearity: OLS/Probit/Logit → STATA. Propensity Score can be tested → STATA.

Checks concerning Propensity score
Check if the distribution for both groups (treated and untreated) are indeed the same. Check the moments. STATA for discrete X: *tab education treatment [aw=propScoreWeight], nofreq col*

6 Experiments and Quasi-Experiments

Experiment: Designed and consciously implemented by human researchers.
Quasi-experiment or natural experiment: randomization → as if randomly assigned.
Potential Outcomes: always only one observed, the other outcome is called counterfactual.

ATE
Different people obv. react differently to treatments. We define the average treatment effect (ATE) to be the population mean off all the individual treatment effects.

Ideal RCT:
$Y_i = \beta_0 + \beta_1 D_i + u_i$ with D_i being randomly assigned. β_1 is therefore the causal effect.
$(\bar{Y}^{treated} - \bar{Y}^{control})$ = differences estimator.
Derivation expectations: $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i) = Y_i(0) + [Y_i(1) - Y_i(0)]D_i$
We have: $Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]D_i$. We rearrange and add and subtract $E[Y_i(0)]$. Hence: $Y_i = E[Y_i(0)] + [Y_i(1) - Y_i(0)]D_i + [Y_i(0) - E[Y_i(0)]]$. We have, as we have a model with just one dummy, that: $\beta_{0i} = E[Y_i(0)]$ and $\beta_{1i} = Y_i(1) - Y_i(0)$ and $u_i = Y_i(0) - E[Y_i(0)] = 0$. We get the simple regression of the form $Y_i = \beta_0 + \beta_1 D_i + u_i$ with β_1 being the causal effect of the treatment.

Introducing additional regressors
Reasons to introduce control variables W_i: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 W_i + u_i$
1. If D_i is random, then there is no correlation between D_i and W_i. But SEs can be reduced.
2. Randomization based on covariates: $E[u_i \mid D_i, W_i] = E[u_i, W_i]$ → once controlled for W_i, the error term is independent of D_i.
To estimate such a model, a regression with interaction is used: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 W_i + \beta_3 D_i \cdot W_i + u_i$. That allows that the effect differs for different groups (effect heterogeneity). Remark: Only controls that are **predetermined** are to be included in the regression.

Internal Validity: problems with $cor(D_i, u_i)$
1. Failure to randomize (small sample, random correlation) → always check if treatment and control group are balanced.
2. Failure to follow treatment protocol, partial compliance. If not all subjects do what they are told, the effect is not causal. Interesting side note: If one knows about treatment actually received X_i and initial random assignment, the causal effect can be estimated via IV-regression. One uses initial assignment Z_i as an instrument for treatment actually received X_i. Z_i is relevant ($cor(Z_i, Y_i) \neq 0$) and exogenous ($cor(u_i, Z_i) = 0$).

3. Attrition: subjects drop out
4. Experimental effects: treatment D is associated with extra effort → Hawthorne effect

External Validity
1. Nonrep. sample
2. Nonrep. treatment
3. Nonrep. implementation
4. General equilibrium effects → treatment depends on scale

And they are on top of all the other problems often physically, ethically and financially not feasible.

7 Difference-in-Difference

Generally:
$\hat{\beta_3}^{DiD} = (\bar{Y}^{treat,after} - \bar{Y}^{treat,before}) - (\bar{Y}^{control,after} - \bar{Y}^{control,before})$
While this DiD is based on means, the regression version (see below) gives (if we dont include controls) the exact same estimate. But we are able to include controls, which is always good. If controls are included, then the estimates differ.

General things about natural experiments (quasi experiments)
Source of randomization but not done explicitly.
They are usually free, no ethical concerns and offer big samples. A con is that they are hard to find. There are 2 kinds of quasi-experiments:

- Treatment D_i as if randomly assigned
  $Y_i = \beta_0 + \beta_1 Treat_i + \beta_2 After_i + \beta_2 Treat_i \cdot After_i + u_i$ with β_3 being the treatment effect and β_1 capturing the systematic differences in the two groups, as the treatment is not completely randomly assigned (meaning $After_i$ is constant). β_2 captures a common time trend (must be the same for the control group as for the treatment group).
  We can write that in regression form (differences estimator): $\Delta Y_i = \beta_2 + \beta_3 D_i + u_i$. We thus are left with the average change over time (β_2) and the treatment effect (β_3).
  Adding control variables: $\Delta Y_i = \beta_2 + \beta_3 D_i + \beta_4 W_{1it} + ... + \beta_{3+r} W_{rit} + u_{it}$.
- A variable Z_i influences receipt of the treatment D_i as if randomly assigned, which allows us the use Z_i as an instrumental variable for D_i.

Critical Assumptions

- The difference in the groups (treated vs. untreated) do not differ over time → otherwise $\beta_1 Treat_i$ would not cancel.
- The post treatment control does not differ for the treated and the untreated group. Example: A change in the labor market that takes place simultaneously as our treatment has to effect both our groups the same! Simply put, the "trend" that both groups were following pre-treatment must be the same. That is called the **Common Trend Assumption**

8 Regression Discontinuity Design: RDD

Receipt of treatment T depends partially or fully on whether some variable R is over some threshold $\hat{R}$. Close by $\hat{R}$ it is random whether one receives the treatment.

Sharp design: $P(T = 1 \mid R > \hat{R}) = 1$ and $P(T = 1 \mid R < \hat{R}) = 0$. Not very common as there is a need for a strict rule that determines treatment, e.g voting procedures. Here reduced form estimate is the same as the treatment effect (IV-estimate).
Fuzzy design, essentially a special case of IV: Some people do not get the treatment even though they are above the threshold and vice versa. Important here is that $cov(X_i, u_i) \neq 0$ as the receipt of the treatment does not entirely depend on $X_i > \hat{R}$. To get an unbiased estimator we need to include an instrument. That works as follows: $Z_i = 1 \; \forall \; W_i < w_0$.
$\rightarrow$ rescale the effect of discontinuity on the outcome by the fraction that is treated.

Causal effect on Y
$\lim_{R \rightarrow \bar{R}^+} E[Y \mid R = \bar{R}] - \lim_{R \rightarrow \bar{R}^-} E[Y \mid R = \bar{R}] = \tau$
How do we estimate τ?
1. Polynomial of R. If the data look linear $\rightarrow$ use a linear regression (same principle as always).
2. Dummy $D_i = 1 \; \forall \; R > \bar{R}$
3. Interaction between D and the polynomial of R, i.e: $Y = \beta_0 + \beta_1 R + \beta 2D + \beta_3 D \cdot R + \epsilon$ with β_2 being the treatment effect. The interaction, as always, allows the slope of our regression to change after $\bar{R}$. That allows for a flexible relation between R and Y. When experimenting with different degree polynomials, we want little sensitivity in our estimation when changing the degree.

Connection to Instrumental Variables
reduced form of previous example: $Y = \beta_0 + \beta_1 R + \beta_2 D + \beta_3 D \cdot R + \epsilon$ with reduced form coefficient being β_2.
first stage of previous example: $T = \pi_0 + \pi_1 R + \pi_2 D + \pi_3 D \cdot R + u$ with first stage coefficient being π_2. The first stage is often not observed meaning we do not see the influence of the threshold-crossing on the treatment but only the effect of the threshold crossing (dummy D) on the outcome variable.
The **IV-estimator** would be $\frac{\hat{\beta}_2}{\hat{\pi_2}}$. For example: The jump in the outcome variable is 3 ($\hat{\beta}_2$) but the share of treated only changes from 0.6 to 0.7, which means $\hat{\pi_2} = 0.1$. That means that the IV estimate is $\frac{3}{0.1} = 30$.

Critical assumptions, important as always
1. similar baseline characteristics of both entities: regress pre-treatment characteristics on R and observe whether there is a "jump" in $\hat{R}$. More accurate is estimating that regression and observing whether the coefficient of $D(D = 1 \; \forall \; R > \hat{R})$ is insignificant and small.
2. No sorting around $\hat{R}$: no manipulating to get the treatment. Histograms around $\hat{R}$ to see whether the frequency is constant.

Threats to internal validity
Same as before with addition of instrumental invalidity (relevance and exogeneity).

Fuzzy treatment effect
E.g if we see a jump of 3 in the outcome variable but the share of treated only changes from 0.6 to 0.7, what would be the treatment effect. $0.1 \longrightarrow 3$ implies that $1 = 10 \cdot 0.1 = 10 \cdot 3 = 30$ is the treatment effect. We call that rescaling.

9 Power Calculations and Multiple Hypothesis Testing

Power Calculations n :sample size of each group
y :outcome of interest
θ :true effect size, we dont know that (we guess)
$d = \frac{1}{n}\sum_{i\in T} y_i - \frac{1}{n}\sum_{j\in C} y_j$
Power = $P\left(\frac{d}{\hat{\sigma_d}/\sqrt{n}} > t_\alpha\right) = 1 - P\left(\frac{d-\theta}{\hat{\sigma_d}/\sqrt{n}} < t_\alpha - \frac{\theta}{\hat{\sigma_d}/\sqrt{n}}\right) \approx 1 - \phi\left(t_\alpha - \frac{\theta}{\hat{\sigma_d}/\sqrt{n}}\right)$. We use the normal approximation with standardisation.
We often assume: $\hat{\sigma_d} = \sqrt{\sigma_{yT}^2 + \sigma_{yC}^2} = \sqrt{2}\sigma_y$ if $\sigma_{yT} = \sigma_{yC}$.
Effect size: often in units of SD: example: $\Delta(\mu_C, \mu_T) = 3$ and $SE = 10$ for both groups. The effect size in terms of the SE is $0.3 = \frac{3}{10}$. Cohen's rule: small effect: $0.2\sigma_y$
, medium effect: $o.5\sigma_y$ and large effect: $0.8\sigma_y$.

Multiple Hypothesis Testing
1. Pre-register: counteract bias with transparency
2. Indexing to reduce number of primary outcomes → example: health is the dependent variable. Health is determined via numerous indicators and checking all of them separately would require harsh adjustment via a method of controlling for multiple hypothesis. One therefore introduces an index that unites health into one entity: $Index_i = \sum_{j=1}^{J} \frac{M_{ij} - \mu_{j,control}}{\sigma_{j,control}}$ with M_{ij} being the measure j of person i. Adjust subtraction and addition relative to the meaning of the measure j.
3. Multiple Hypothesis testing: If there H^i hypotheses with $i \in 2, ..., m$, we still need to correct for that.

Bonferroni-correction
Reject H_0^i if $p_i < \frac{\alpha}{m}$ with $i \in [1, m]$. FWER (family error rate) correction that limit type I error to α. It is a very conservative way of correcting and therefore sacrifices power.

Holm-correction
1. Sort hypotheses by p-value: from lowest p-value to highest
2. Find lowest H_0^z with $z \in [1, m]$ such that $p_z > \frac{\alpha}{m+1-z}$
3. Reject all H_0^i with $i < z$ and do not reject those with $i \geq z$
This is also FWER correction and also guarantees that $P(Type1) < \alpha$. Less conservative, higher power.

Benjamin-Hochberg
1. Sort hypotheses by p-value: from lowest p-value to highest
2. Find highest H_0^z with $z \in [1, m]$ such that $p_z < \frac{\alpha}{m+1-z}$
3. Reject all H_0^i with $i \leq z$ and do not reject those with $i > z$.

This is a FDR correction (false discovery rate). Depends not only on α but on other data characteristics. With large m, there are falsely rejected H_0.

10 Additional Stuff

Log-Lin models and the (approximated) interpretation of β_i:

- Lin-Lin model: with $\Delta X = 1 \rightarrow \Delta Y = \beta_1 \cdot \Delta X$
- Log-Lin model: with $\Delta X = 1 \rightarrow \Delta Y = 100 \cdot \beta_1\%$ or exact: $100 \cdot (e^{\beta_1} - 1)$ with approx. being reasonably accurate with $\beta_1 \in [-0.1, 0.1]$
- Lin-Log model: with an increase of X of 1 percent, we have $\Delta Y = \frac{\beta_1}{100} \cdot \Delta X\%$
- Log-Log model: $\Delta Y\% = \beta_1 \Delta Y\%$

Power calculations with required sample size calculation:
$\alpha = 0.05, Power = 0.8$ and $\theta = 0.1\sigma_D$ with $\sigma_D = \sqrt{\sigma_{YT}^2 + \sigma_{YC}^2} \approx \sqrt{2}\sigma_Y$
Power $\approx 1 - \phi\left(t_\alpha - \frac{\theta}{\sigma_Y/\sqrt{n}}\right) = 1 - \phi\left(1.645 - \frac{0.1\sigma_Y}{\sqrt{2}\sigma_Y}\sqrt{n}\right) \rightarrow \phi\left(1.645 - \frac{0.1}{\sqrt{2}}\sqrt{n}\right) = 0.2 \rightarrow$ standard normal table: $1.645 - \frac{0.1}{\sqrt{2}}\sqrt{n} = -0.845 \rightarrow n = 310$