Optimal data collection design in machine learning

Introduction

A major current challenges in data science concerns:

1. Optimizing the acquisition costs (time and money) of big data to create novel opportunities for data analysts (Sivarajah et al., 2017).

Can we collect less data to reach the same desired statistical properties?

Overview

RQ: is having "many but bad" examples always worse –in terms of minimization of the generalization error– than having "few but good" examples in a balanced fixed effects context with correlated errors?

Main model

Fixed Effects GLS (FEGLS) output model taken from Wooldridge (2010) work 10:

$$y_{n,t} := \eta_n + \underline{\beta}' \underline{x}_{n,t}, \text{ for } n = 1, \ldots, N, t = 1, \ldots, T\,, \tag{1}$$

Outputs $y_{n,t}$ are **unavailable** directly. Only noisy measurements $z_{n,t}$ are available:

$$z_{n,t} := y_{n,t} + \varepsilon_{n,t}\,, \text{ for } n = 1, \ldots, N, t = 1, \ldots, T\,, \tag{2}$$

$\varepsilon_{n,t}$ identically distributed and $\varepsilon_{n,t} \not\perp \varepsilon_{n,s}$. We assume **strict exogeneity** of the explanatory variables conditional on η_n (Wooldridge, 2010). Notice that:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \text{ where } X_n = \begin{bmatrix} \underline{x}'_{n,1} \\ \underline{x}'_{n,2} \\ \vdots \\ \underline{x}'_{n,T} \end{bmatrix} \text{ and } \underline{x}_{n,t} = ([x_{n,t,1} \ldots x_{n,t,p}])'$$

being $\underline{x}_{n,t}$ the vector of features for unit n at time t.

- How are errors correlated?

- How are errors correlated?
- An AR(1) model is assumed (Bhargava et al., 1982; Im et al., 1999):

$$\Lambda := \sigma^2\Psi := \mathrm{Var}\left(\underline{\varepsilon}_n\right) = \mathbb{E}\{\underline{\varepsilon}_n\underline{\varepsilon}_n'\} =$$
$$= \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho^2 & \rho & 1 \end{bmatrix} \in \mathbb{R}^{T\times T}, \quad (3)$$

Λ matrix is

- **symmetric**
- positive semi-definite
- idempotent
- **positive-definite**
- eigenvalues multiplicity

After demeaning (Appendix) the resulting covariance matrix $\mathbb{E}\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}'_n\}$–being $\ddot{\varepsilon}$ the demeaned error term– has the expression

$$\Omega := \sigma^2 \Phi := \mathrm{Var}\left(\ddot{\underline{\varepsilon}}_n\right) = \mathbb{E}\{\ddot{\underline{\varepsilon}}_n \ddot{\underline{\varepsilon}}'_n\} = Q_T \mathbb{E}\{\underline{\varepsilon}_n \underline{\varepsilon}'_n\} Q'_T = \\ = Q_T \Lambda Q'_T = \sigma^2 Q_T \Psi Q'_T \,, \tag{4}$$

Ω matrix is

- **symmetric**
- **positive semi-definite**
- idempotent
- positive-definite
- eigenvalues multiplicity

PROBLEM:

Matrix (4) is rank deficient ($rank(\Omega) = T - 1 < T$). Cannot be inverted.

These produce the same estimate of $\underline{\beta}$

(a) project Eq. (2) onto L

(set of vectors orthogonal to $\underline{1}_T$) (b) apply ordinary GLS

SOLUTIONS:

1. drop one of the time periods from the analysis (see, e.g. Im et al., 1999, Theorem 4.3)

2. Use the Moore-Penrose pseudoinverse of Ω, denoted as Ω^+ (a)

PROBLEM:
Matrix (4) is rank deficient
$(rank(\Omega) = T - 1 < T)$.
Cannot be inverted.

SOLUTIONS:

1. drop one of the time periods from the analysis (see, e.g. Im et al., 1999, Theorem 4.3)

2. Use the Moore-Penrose pseudoinverse of Ω, denoted as Ω^+ (a)

These produce the same estimate of $\underline{\beta}$

(a) project Eq. (2) onto L
(set of vectors orthogonal to $\underline{1}_T$) (b)
apply ordinary GLS

Definition 1

Generalization error or expected risk for the i^{th} unit ($i = 1, ..., N$), conditioned on the training input data,

$$R_i\left(\{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right) := \mathbb{E}\left\{\left(\hat{\eta}_{i,FEGLS} + \underline{\hat{\beta}}'_{FEGLS}\underline{x}_i^{test} - \eta_i - \underline{\beta}'\underline{x}_i^{test}\right)^2 \middle| \{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T}\right\} \tag{5}$$

Notice that R_i is defined on test data being a performance index.

In the next slides we will rewrite R_i conveniently.

Details on computations leading to the following expression of R_i are reported in the paper. Here it is worth noticing that the generalization error can be split into 6 components, namely:

$$
\begin{aligned}
&R_i\left(\left\{\underline{x}_{n,t}\right\}_{n=1,\dots,N}^{t=1,\dots,T}\right) = \\
&= \frac{\sigma^2}{T^2}\underline{1}_T' X_i \left(\sum_{n=1}^N \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} X_i' \underline{1}_T + \frac{\sigma^2}{T^2}\underline{1}_T' \Psi \underline{1}_T \\
&- \frac{2\sigma^2}{T^2}\underline{1}_T' X_i \left(\sum_{n=1}^N \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \ddot{X}_i' \Phi^+ Q_T \Psi \underline{1}_T \\
&+ \sigma^2 \mathbb{E}\left\{\left(\underline{x}_i^{\text{test}}\right)' \left(\sum_{n=1}^N \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \underline{x}_i^{\text{test}} \,\middle|\, \left\{\underline{x}_{n,t}\right\}_{n=1,\dots,N}^{t=1,\dots,T}\right\} \\
&- \frac{2\sigma^2}{T}\underline{1}_T' X_i \left(\sum_{n=1}^N \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \mathbb{E}\left\{\underline{x}_i^{\text{test}}\right\} \\
&+ \frac{2\sigma^2}{T}\left(Q_T \Psi \underline{1}_T\right)' \Phi^+ \ddot{X}_i \left(\sum_{n=1}^N \ddot{X}_n' \Phi^+ \ddot{X}_n\right)^{-1} \mathbb{E}\left\{\underline{x}_i^{\text{test}}\right\}, \qquad (6)
\end{aligned}
$$

In the work we derived a large-sample approximation of (6) with respect to T, for fixed N, which has its major applications in macroeconometrics.

The large-sample approximation is based on the following major results (see paper for the assumptions under which they hold):

In the work we derived a large-sample approximation of (6) with respect to T, for fixed N, which has its major applications in macroeconometrics.

The large-sample approximation is based on the following major results (see paper for the assumptions under which they hold):

- $$\operatorname*{plim}_{T \to +\infty} \frac{1}{T} \underline{1}'_T X_i = \left(\mathbb{E} \left\{ \underline{x}_{i,1} \right\} \right)', \tag{7}$$

In the work we derived a large-sample approximation of (6) with respect to T, for fixed N, which has its major applications in macroeconometrics.
The large-sample approximation is based on the following major results (see paper for the assumptions under which they hold):

- $$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}\underline{1}'_T X_i = \left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)', \tag{7}$$

- $$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}\ddot{X}_i'\Phi^{+} Q_T \Psi \underline{1}_T = \underline{0}_p\,. \tag{8}$$

In the work we derived a large-sample approximation of (6) with respect to T, for fixed N, which has its major applications in macroeconometrics.
The large-sample approximation is based on the following major results (see paper for the assumptions under which they hold):

- $$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}\underline{1}'_T X_i = \left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)', \tag{7}$$

- $$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}\ddot{X}_i'\Phi^+ Q_T \Psi \underline{1}_T = \underline{0}_p\,. \tag{8}$$

- If $\lim_{T\to\infty}\|\Phi^+ - Q_T\Psi^{-1}Q'_T\|_2 = 0$ holds, then

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T}\sum_{n=1}^{N}\ddot{X}_n'\Phi^+\ddot{X}_n = A_N\,, \tag{9}$$

When (7)-(8)-(9) hold, then we can write the large-sample approximation of the generalization error $R_i\left(\{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T}\right)$ w.r.t. T as:

$$
\begin{aligned}
(6) \quad &\simeq \quad \frac{\sigma^2}{T}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A_N^{-1}\mathbb{E}\left\{\underline{x}_{i,1}\right\} + \frac{\sigma^2}{T}\frac{1+\rho}{1-\rho} \\
&\quad + \frac{\sigma^2}{T}\mathbb{E}\left\{\left(\underline{x}_i^{test}\right)' A_N^{-1}\underline{x}_i^{test}\right\} - 2\frac{\sigma^2}{T}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)' A_N^{-1}\mathbb{E}\left\{\underline{x}_i^{test}\right\} \\
&= \quad \frac{\sigma^2}{T}\left(\frac{1+\rho}{1-\rho} + \mathbb{E}\left\{\left\|A_N^{-\frac{1}{2}}\left(\mathbb{E}\left\{\underline{x}_{i,1}\right\} - \underline{x}_i^{test}\right)\right\|_2^2\right\}\right), \qquad (10)
\end{aligned}
$$

Blue and *cyan* terms of (6) disappear due to (8).

Time to optimize...

Aim: Optimize (10) when

$$\sigma^2 = kc^{-\alpha}$$

being $c \in [c_{min}, c_{max}]$ the cost per example with upper bound C on total supervision cost NTc. $T \in [\frac{C}{c_{max}}, \ldots, \frac{C}{c_{min}}]$.

Time to optimize...

Aim: Optimize (10) when

$$\sigma^2 = kc^{-\alpha}$$

being $c \in [c_{min}, c_{max}]$ the cost per example with upper bound C on total supervision cost NTc. $T \in [\frac{C}{c_{max}}, \ldots, \frac{C}{c_{min}}]$.

Idea: The higher the cost per example, the greater the precision of supervision.

Time to optimize...

Aim: Optimize (10) when

$$\sigma^2 = kc^{-\alpha}$$

being $c \in [c_{min}, c_{max}]$ the cost per example with upper bound C on total supervision cost NTc. $T \in [\frac{C}{c_{max}}, \ldots, \frac{C}{c_{min}}]$.

Idea: The higher the cost per example, the greater the precision of supervision.

Scenarios:

1. "decreasing returns to scale*": $0 < \alpha < 1$
2. "increasing returns to scale*": $\alpha > 1$
3. "constant returns to scale*": $\alpha = 1$

$*$ of the precision with respect to the cost per example

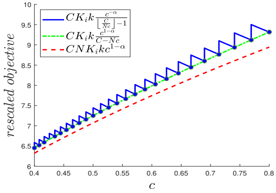
Can we further simplify the analysis?

Actual optimization problem:

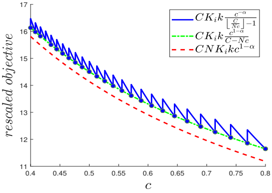
$$\underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} \, K_i k \frac{c^{-\alpha}}{\left\lfloor \frac{C}{Nc} \right\rfloor} \,. \tag{11}$$

However...

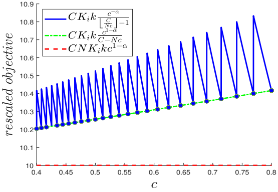
Folllowing Gnecco and Nutarelli (2019[4]), the objective function of the optimization problem (11), rescaled by the multiplicative factor C, can be approximated, with a negligible error in the maximum norm on $[c_{\min}, c_{\max}]$, by $NK_i kc^{1-\alpha}$.



(a) $\alpha = 0.5$



(b) $\alpha = 1.5$



(c) $\alpha = 1$

Figure 1: Plots of the rescaled objective functions $CK_ik\frac{c^{-\alpha}}{\lfloor\frac{C}{Nc}\rfloor-1}$, $CNK_ikc^{1-\alpha}$, and $CNK_ik\frac{c^{1-\alpha}}{C-Nc}$

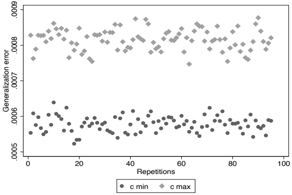
Final optimization problem

$$\underset{c \in [c_{\min}, c_{\max}]}{\text{minimize}} \, N K_i k c^{1-\alpha}\,, \tag{12}$$

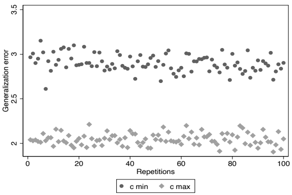
whose optimal solutions c° have the following expressions:

1. if $0 < \alpha < 1$ ("decreasing returns of scale"): $c^\circ = c_{\min}$;
2. if $\alpha > 1$ ("increasing returns of scale"): $c^\circ = c_{\max}$;
3. if $\alpha = 1$ ("constant returns of scale"): $c^\circ =$ any cost c in the interval $[c_{\min}, c_{\max}]$.

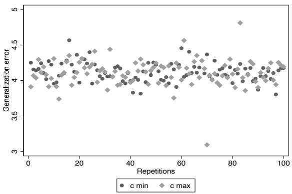
Notice: No assumptions of the probability distribution of the input examples is needed!



(a) $\alpha = 0.5$



(b) $\alpha = 1.5$



(c) $\alpha = 1$

Figure 2: Empirical approximations of the generalization error in the various scenarios $\alpha = 0.5$, $\alpha = 1.5$, $\alpha = 1$

Appendix

Demeaning step

Common practice in F.E.: eliminate unobserved individual heterogeneity (η_n in Eq. (1)). How?

De-meaning using $Q_T \in \mathbb{R}^{T \times T}$

where

$$Q_T := I_T - \frac{1}{T}\underline{1}_T \underline{1}_T'$$

Q_T matrix is

- **symmetric**
- **positive semi-definite**
- **idempotent**
- positive-definite
- **eigenvalues multiplicity**

Proof of Eq.(7), sketch

$$\underset{T\to+\infty}{\mathrm{plim}} \frac{1}{T} \underline{1}'_T X_i = \left(\mathbb{E}\left\{\underline{x}_{i,1}\right\}\right)',$$

Proof.

i) Replace the empirical average of $\underline{x'}_{n,t}$ with the common expected value[1]

ii) Apply Chebyshev's weak law of large numbers. □

[1] since $\mathbb{E}\left\{\underline{x}_{i,t}\right\}$ is the same $\forall t$ we arbitrarily chose $t = 1$.

Proof of Eq.(8), sketch

$$\underset{T \to +\infty}{\text{plim}} \frac{1}{T} \ddot{X}_i' \Phi^+ Q_T \Psi \underline{1}_T = \underline{0}_p \,.$$

Proof.

- Step 1: define $\underline{v}_T := Q' \Phi^+ Q \Psi \underline{1}_T = Q' \Phi^+ \underline{u}_T$
- Step 2: rewrite the argument of the *plim* using $\underline{v}_T$:

$$\frac{1}{T} \ddot{X}_i' \Phi^+ Q \Psi \underline{1}_T = \frac{1}{T} X_i' Q' \Phi^+ Q \Psi \underline{1}_T = \frac{1}{T} X_i'$$

- Step 3: Notice that in $\frac{1}{T} X_i'$ is a weighted average with weights $\underline{v}_T \to$ some law of large number must apply (Bai, Cheng,and Zhang (1997, Theorem 2.1));
- Step 4: check if $\underline{v}_T$ and $\frac{1}{T} X_i'$ satisfy the requirements of Bai, Cheng,and Zhang (1997, Theorem 2.1) (they do)

$\square$

Proof of Eq.(9), sketch

$$\operatorname*{plim}_{T\to+\infty} \frac{1}{T} \sum_{n=1}^{N} \ddot{X}_n' \Phi^+ \ddot{X}_n = A_N \,,$$

Proof.

- Step 1: prove that $\ddot{X}_n' \Phi^+ \ddot{X}_n = X_n' Q' \Phi^+ Q X_n = X_n' \Phi^+ X_n = X_n' \left[\Phi^+ - Q\Psi^{-1}Q'\right] X_n + X_n' Q \Psi^{-1} Q' X_n$.;
- Step 2: prove that $\operatorname*{plim}_{T\to+\infty} \frac{1}{T} X_n' \left[\Phi^+ - Q\Psi^{-1}Q'\right] X_n = 0_{p\times p}$.;
- Step 3: prove that $\operatorname*{plim}_{T\to+\infty} \frac{1}{T} X_n' Q \Psi^{-1} Q' X_n = \frac{1+\rho^2}{1-\rho^2} \mathbb{E}\left\{\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)\left(\underline{x}_{n,1} - \mathbb{E}\left\{\underline{x}_{n,1}\right\}\right)'\right\}$.;
- Step 4: Combine Steps 1,2 and 3. Then sum over N.

$\square$

Set-up for simulations

For each c, an **empirical approximation** of the generalization error is computed:

$$\sum_{i=1}^{N} \mathbb{E}\left\{ \left(\hat{\eta}_{i,FEGLS} + \underline{\hat{\beta}}'_{FEGLS} \underline{x}_i^{test} - \eta_i - \underline{\beta}' \underline{x}_i^{test} \right)^2 \Big| \{\underline{x}_{n,t}\}_{n=1,\dots,N}^{t=1,\dots,T} \right\}$$

$$\simeq \quad \frac{1}{N^{test}} \sum_{i=1}^{N} \sum_{h=1}^{N_i^{test}} \frac{1}{\mathcal{N}^{tr}} \sum_{j=1}^{\mathcal{N}^{tr}} \left(\hat{\eta}^j_{i,FEGLS} + \left(\underline{\hat{\beta}}^j_{FEGLS} \right)' \underline{x}_{i,h}^{test} - \eta_i - \underline{\beta}' \underline{x}_{i,h}^{test} \right)^2 \tag{13}$$

[2] Can provide further details at the end of the discussion

Set-up for simulations

For each c, an **empirical approximation** of the generalization error is computed:

$$\sum_{i=1}^{N} \mathbb{E}\left\{ \left(\hat{\eta}_{i,FEGLS} + \underline{\hat{\beta}}'_{FEGLS} \underline{x}_i^{test} - \eta_i - \underline{\beta}' \underline{x}_i^{test} \right)^2 \middle| \{\underline{x}_{n,t}\}_{n=1,\ldots,N}^{t=1,\ldots,T} \right\}$$

$$\simeq \quad \frac{1}{N^{test}} \sum_{i=1}^{N} \sum_{h=1}^{N_i^{test}} \frac{1}{\mathcal{N}^{tr}} \sum_{j=1}^{\mathcal{N}^{tr}} \left(\hat{\eta}^j_{i,FEGLS} + \left(\underline{\hat{\beta}}^j_{FEGLS} \right)' \underline{x}_{i,h}^{test} - \eta_i - \underline{\beta}' \underline{x}_{i,h}^{test} \right)^2 \quad (13)$$

(13) is based on $\mathcal{N}^{tr}$ training sets and N_i^{test} test examples for each unit i ($i = 1, \ldots, N$), hence on a total number $N^{test} = \sum_{i=1}^{N} N_i^{test}$ of test examples.[2]

[2] Can provide further details at the end of the discussion

Details on simulations set-up

Fair comparison (since T depends on c):

- The number of rows in each matrix X_n is increased when c is reduced from $c_{\max}$ to $c_{\min}$, by increasing the number of observations T.
- For a fair comparison, when doing this, the rows already present in each matrix X_n **are kept fixed**.
- Finally, the **same** test examples (generated independently from the training sets) are used to assess the performance of the fixed effects generalized least squares estimates for different costs per example c.

We choose:

1. $N = 20$,
2. $p = 5$ (for the number of features),
3. $c_{\min} = 2$, $c_{\max} = 4$,
4. $\mathcal{N}^{tr} = 100$ (for the number of training sets),
5. $N_i^{test} = 50$ for the number of test examples per unit (hence the total number of test examples is $N^{test} = 1000$)

The number of training examples per unit is $T = 50$ for $c = c_{\min}$, and $T = 25$ for $c = c_{\max}$.[3] Without loss of generality, the constant k of the variance of the supervision cost is assumed to be equal to 1.

[3] In this way, the (upper bound on the) total supervision cost is $C = 2000$ for both cases.

The components of $\underline{\beta}$ are generated randomly and independently according to a uniform distribution on $[-1, 1]$:

$$\underline{\beta} = [-0.8562, 0.6837, 0.2640, -0.0038, -0.0598]' . \tag{14}$$

The fixed effects η_n (for $n = 1, \ldots, N$) are generated similarly for each unit;

For both training and test sets, the input data associated with each unit are generated as realizations of a multivariate Gaussian distribution with mean $\underline{0}$ and covariance matrix $\mathrm{Var}\left(\underline{x}_{n,t}\right) = \mathrm{Var}\left(\underline{x}_i^{test}\right) = A_x A_x'$, where the elements of $A_x \in \mathbb{R}^{p \times p}$ have been randomly and independently generated according to a uniform probability density on the interval [0,1].

References I

- *Corporate governance of big data: Perspectives on value, risk, and cost* Tallon, Paul P., Computer, volume 46, number 6 pages 32–38, 2013, IEEE

- *Big Data Tools: Advantages and Disadvantages* Baig, Maria Ijaz and Shuib, Liyana and Yadegaridehkordi, Elaheh, Journal of Soft Computing and Decision Support Systems, volume 6, number 6 pages 14–20, 2019

- *On the trade-off between number of examples and precision of supervision in regression problems*, Gnecco G., Nutarelli, F. (2019) in Proceedings of the 4th International Conference of the International Neural Network Society on Big Data and Deep Learning (INNS BDDL 2019), Sestri Levante, Italy, pp. 1-6.

References II

On the trade-off between number of examples and precision of supervision in machine learning problems., Gnecco G., Nutarelli, F. (2019) Optimization Letters. https:// doi. org/ 10. 1007/ s11590- 019- 01486-x.

Optimal trade-off between sample size and precision of supervision for the fixed effects panel data model., Gnecco G., Nutarelli, F. (2020). In Proceedings of the 5^{th} International Conference on machine Learning, Optimization Data science (LOD 2019), Certosa di Pontignano (Siena), Italy. Lecture Notes in Computer Science, vol. 11943, pp. 1-12.

Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model. Gnecco, G., Nutarelli, F., Selvi, D. (2020). Soft Computing, 24, 15937–15949.

References III

- *Optimal data collection design in machine learning: the case of the fixed effects generalized least squares panel data model.* Gnecco, G., Nutarelli, F. and Selvi, D., 2021. Machine Learning, pp.1-36.
- *Econometric analysis of cross section and panel data.* Wooldridge, J.M., 2002. MIT press. Cambridge, MA, 108.
- *Big data: New tricks for econometrics.* Varian, H.R., 2014. Journal of Economic Perspectives, 28(2), pp.3-28.
- *Machine learning methods that economists should know about.* Athey, S. and Imbens, G.W., 2019. Annual Review of Economics, 11, pp.685-725.
- Daron Acemoglu and Joshua Linn. *Market size in innovation: theory and evidence from the pharmaceutical industry.* The Quarterly journal of economics, 119(3):1049–1090, 2004.

References IV

- Zeina Alsharkas. *Firm size, competition, financing and innovation.* International Journal of Management and Economics, 44(1):51–73, 2014.
- Ram Bala, Pradeep Bhardwaj, and Pradeep K Chintagunta. *Pharmaceutical product recalls: Category effects and competitor response.* Marketing Science, 36(6):931–943, 2017.
- Natarajan Balasubramanian and Jeongsik Lee. *Firm age and innovation. Industrial and Corporate Change*, 17(5):1019–1047, 2008.
- George Ball, Jeffrey Thomas Macher, and Ariel Dora Stern. *Recalls, innovation, and competitor response: Evidence from medical device firms. 2018.*

References V

- Margaret E Blume-Kohout and Neeraj Sood. *Market size and innovation: Effects of medicare part d on pharmaceutical research and development.* Journal of public economics, 97:327–336, 2013.
- Bowe C.*Merck quarterly profits hit by vioxx recall, 2005.* [Online; accessed 15-April-2021].
- Rodrigo A Cerda. *Endogenous innovations in the pharmaceutical industry. Journal of Evolutionary Economics, 17(4):473–515, 2007.*
- Jessie Cheng. *An antitrust analysis of product hopping in the pharmaceutical industry.* Colum. L. Rev., 108:1471, 2008.
- Abdulkadir Civan and Michael T Maloney. *The effect of price on pharmaceutical r&d.* The BE Journal of Economic Analysis Policy, 9(1), 2009.

References VI

- Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. *The price of innovation: new estimates of drug development costs.* Journal of health economics, 22(2):151–185, 2003.
- Pierre Dubois, Olivier De Mouzon, Fiona Scott-Morton, and Paul Seabright. *Market size and pharmaceutical innovation.* The RAND Journal of Economics, 46(4):844–871, 2015.
- Mark Duggan and Fiona Scott Morton. *The effect of medicare part d on pharmaceutical prices and utilization.* American Economic Review, 100(1):590–607, 2010.
- M.Provost et al. *Pharmaceutical antitrust law in european union*. Dechert LLP, 2019.
- Paul A Geroski and Chris F Walters. *Innovative activity over the business cycle.* The Economic Journal, 105(431):916–928, 1995.

References VII

Carmelo Giaccotto, Rexford E Santerre, and John A Vernon. *Drug prices and research and development investment behavior in the pharmaceutical industry.* The Journal of Law and Economics, 48(1):195–214, 2005.

Bronwyn H Hall and Nathan Rosenberg. *Handbook of the Economics of Innovation*, volume 1. Elsevier, 2010.

Kelsey Hall, Tyler Stewart, Jongwha Chang, and Maisha Kelly Freeman. *Characteristics of FDA drug recalls: A 30-month analysis. American Journal of Health-System Pharmacy*, 73(4):235–240, 2016.

Christian Hansen, Jerry Hausman, and Whitney Newey. *Estimation with many instrumental variables.* Journal of Business Economic Statistics, 26(4):398–422, 2008.

References VIII

- Iraj Hashi and Nebojša Stojčić. *The impact of innovation activities on firm performance using a multi-stage model: Evidence from the community innovation survey 4.* Research Policy, 42(2):353–366, 2013.
- Venit J.S. Hawk, B.E. and Huser H.L. *Recent developments in EU merger control. antitrust.* (15):24, 2000.
- Elena Huergo and Jordi Jaumandreu.*How does probability of innovation change with firm age?* Small Business Economics, 22(3-4):193–207, 2004.
- Boyan Jovanovic. *Product recalls and firm reputation.* Technical report, National Bureau of Economic Research, 2020.
- Alfred Kleinknecht and Bart Verspagen. *Demand and innovation: Schmookler re-examined.* Research policy, 19(4):387–394, 1990.

References IX

Steven Klepper and Franco Malerba. *Demand, innovation and industrial dynamics: an introduction.* Industrial and Corporate Change, 19(5):1515–1520, 2010.

Srinivas Kolluru and Pundarik Mukhopadhaya. *Empirical studies on innovation performance in the manufacturing and service sectors since 1995: A systematic review.* Economic Papers: A journal of applied economics and policy, 36(2):223–248, 2017.

Margaret K Kyle and Anita M McGahan. *Investments in pharmaceuticals before and after trips.* Review of Economics and Statistics, 94(4):1157–1172, 2012.

Jean O Lanjouw. *Patents, price controls, and access to new drugs: how policy affects global market entry.* Technical report, National Bureau of Economic Research, 2005.

References X

Frank R Lichtenberg. *Pharmaceutical innovation as a process of creative destruction. Knowledge Accumulation and Industry Evolution: The Case of Pharma-Biotech*, page 61, 2006.

Wei Lin and Jeffrey M Wooldridge. *Testing and correcting for endogeneity in nonlinear unobserved effects models.* In Panel Data Econometrics, pages 21–43. Elsevier, 2019.

Chin-jung Luan, Chengli Tien, and Yi-chuang Chi. *Downsizing to the wrong size? a study of the impact of downsizing on firm performance during an economic downturn*, The International Journal of Human Resource Management, 24(7):1519–1535, 2013.

References XI

- Franco Malerba. *Innovation and the evolution of industries. In Innovation, Industrial Dynamics and Structural Transformation*, pages 7–27. Springer, 2007. 41 [42] Anthony Markham. Lurbinectedin: first approval. Drugs, pages 1–9, 2020. [43] Linda Martin, Melissa Hutchens, Conrad Hawkins, and Alaina Radnov. How much do clinical trials cost?, 2017.
- Kamel Mellahi and Adrian Wilkinson.*A study of the association between level of slack reduction following downsizing and innovation output.* Journal of Management Studies, 47(3):483–508, 2010.
- David Mowery and Nathan Rosenberg. *The influence of market demand upon innovation: a critical review of some recent empirical studies.* Research policy, 8(2):102–153, 1979.

References XII

- Igho J Onakpoya, Carl J Heneghan, and Jeffrey K Aronson. *Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis.* Critical reviews in toxicology, 46(6):477–489, 2016.
- Ariel Pakes and Mark Schankerman. *The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources.* In R&D, patents, and productivity, pages 73–88. University of Chicago Press, 1984.
- Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. *The productivity crisis in pharmaceutical R&D.* Nature reviews Drug discovery, 10(6):428–438, 2011.
- Jorge V Pérez-Rodríguez and Beatriz GL Valcarcel. *Do product innovation and news about the r&d process produce large price changes and overreaction? the case of pharmaceutical stock prices.* Applied Economics, 44(17):2217–2229, 2012.

References XIII

- W Price and II Nicholson. *Making do in making drugs: Innovation policy and pharmaceutical manufacturing.* BCL Rev., 55:491, 2014.
- Bastian Rake. *Determinants of pharmaceutical innovation: the role of technological opportunities revisited.* Journal of Evolutionary Economics, 27(4):691–727, 2017.
- David Roodman. *How to do xtabond2: An introduction to difference and system gmm in stata.* The stata journal, 9(1):86–136, 2009.
- Frederic M Scherer. *Demand-pull and technological invention: Schmookler revisted.* The Journal of Industrial Economics, pages 225–237, 1982.
- Jacob Schmookler. *Invention and economic growth.* Harvard University Press, 2013.

References XIV

Vishal B Siramshetty, Janette Nickel, Christian Omieczynski, Bjoern-Oliver Gohlke, Malgorzata N. Drwal, and Robert Preissner. *Withdrawn—a resource for withdrawn and discontinued drugs.* Nucleic acids research, 44(D1):D1080–D1086, 2016.

Gregory N Stock, Noel P Greis, and William A Fischer. *Firm size and dynamic technological innovation.* Technovation, 22(9):537–549, 2002.

Paul Stoneman. *Soft innovation: economics, product aesthetics, and the creative industries.* Oxford University Press, 2010.

George Symeonidis. *Innovation, firm size and market structure: Schumpeterian hypotheses and some new themes* 1996.

Terence N. *Merck pulls vioxx painkiller from market, and stock plunges, 2004* [Online; accessed 15-April-2021].

References XV

Sriram Thirumalai and Kingshuk Sinha. *Product recalls in the medical device industry: An empirical exploration of the sources and financial consequences.* Management Science, 57:376–392, 2011.

Carl H Tong, Lee-Ing Tong, and James E Tong. *The vioxx recall case and comments.* Competitiveness Review: An International Business Journal, 2009.

Anish Vaishnav. *Product market definition in pharmaceutical antitrust cases: Evaluating cross-price elasticity of demand.* Colum. Bus. L. Rev., page 586, 2011.