Text as Instruments*

James Nesbit†

March 6, 2021

This paper provides a theoretical framework to justify and guide the use of text data in the estimation of quantitative economic models. Previous work that utilizes text data has implicitly assumed that the text and traditional data are driven by common exogenous unobservables, such as structural macroeconomic shocks or financial news. This link has been used informally to infer the responses of variables of interest to variation in the text. In contrast, this paper introduces a model that formalizes an explicit link between text data and the unobservables in the econometric model, and therefore justifies formally the use of text data to augment existing econometric methods. The model highlights the difficulty in finding valid instruments from text, but derives sufficient conditions to identify terms that are exogenous. We show that the generalized log odds ratio of a set of certain terms can be used an instrument. As long as the size of documents do not grow too relative to the time dimension, estimation of the text instruments does not distort HAC standard errors. We use our results to identify monetary policy shocks, in the presence of other macroeconomic shocks, using FOMC transcripts. We also quantify the effects of fresh news on equity prices.

Keywords: Machine Learning, Text Analysis, Monetary Policy Shocks.

*I am very grateful to my advisors Timothy Christensen and José Luis Montiel Olea for their support. I would also like to thank Elliott Ash, David Childers, Jean-Jacques Forneron, Alfred Galichon, Elena Manresa, Michael McMahon, Konrad Menzel, Anna Mikusheva, Mikkel Plagborg-Møller, Serena Ng, Isabelle Perrigne, Alex Torgovitsky, Quang Vuong, seminar participants at NYU and the 2020 Econometric Society World Congress. The usual disclaimer applies.

†Department of Economics, New York University, NY 10003 (email:jmn425@nyu.edu).

1 Introduction

Text data is being increasingly used in applied economics research to infer latent information that is otherwise unobservable using standard data sources. The *implicit* assumption of this research is that there exists exogenous unobservable random variables that generate both the text and the tabular data. This link between text and data is used *informally* to infer the responses of variables of interest to changes in these unobservables through variation in the composition of the text.

These exogenous unobservables can be structural macroeconomic shocks which drive the evolution of macroeconomic variables, as well as the generation of policy documents. Romer and Romer (2004) uses transcripts and other materials from Federal Open Market Committee (FOMC) meetings to identify exogenous monetary policy, and Romer and Romer (2010) uses presidential speeches and Congressional reports to identify exogenous fiscal policy. These exogenous unobservables can also be market or company specific news which drives equity prices and newspaper reporting. Tetlock (2007) uses a popular column in the Wall Street to quantify the effect that sentiment has on market returns. Ke, Kelly, and Xiu (2020) uses a more comprehensive collection of Wall Street Journal articles to quantity the effect of news on individual equity returns.

This paper builds a model of text that formalizes an explicit link between text and quantitative data. Our starting point is a simple IV model with a single right hand side endogenous variable. We posit the existence of a random variable that can be used as an instrument, but is *unobserved* and can only be inferred from an auxiliary corpus of text. This instrument appears as a taste shifter in a random utility model of text generation, where documents are formed as an independent sequence of terms chosen from a vocabulary. The utility of a term depends on this unobserved instrument as well as the variables from the quantitative model. This creates tension, as the probability of most terms are relevant to the endogenous variable (one of the requirements of an instrument), but are not exogenous.

As a motivating example, consider the problem of identifying monetary policy shocks from transcripts of the FOMC meetings using a Local Projections IV (Ramey 2016). There are two other shocks in the model, an output shock and an inflation shock, and we assume that policy makers plan the policy responses to these shocks during these

meetings. Suppose we were to use the frequency of the term '*monetary policy*' or '*interest rates*' as an instrument for our monetary policy shocks. We would expect that there would be increased discussion about '*monetary policy*' when there is a monetary policy shock, so this frequency is relevant. However, we would also expect that the frequency of '*monetary policy*' would increase when there is an inflation or output shock, as policy makers discuss the policy response to these shocks. Hence the frequency of '*monetary policy*' would not be exogenous, and not a valid instrument.

Under a parametric distributional assumption, the conditional probabilities of terms take logit form, and we show that the log odds ratio of the likelihood of two special terms, which are not shifted by the unobservables from the quantitative model, are valid instruments. We argue that in order to find such terms, we should include additional events to the information set of our text model. These include events observable in the text, which we refer to as *context*, such as the presence or absence of certain terms within the same document, or the similarity of the current documents with others in the corpus. Returning to our example, we condition on the context that terms similar to '*inflation*' and '*output*' do not appear nearby discussion of '*monetary policy*'. Conditional on this event, the log odds ratio of '*monetary policy*' and '*interest rates*' is a valid instrument.

We show that this result can be generalized to the *generalized log odds* of a set of terms. This is a weighted sum of log probabilities such that the weights sum to zero. Under a generalization of our previous assumption, we show that the generalized log odds of a set of terms that is not shifted by the unobservables from the quantitative model is a valid instrument. This generalization is natural as many synonyms of a term may appear in a corpus.

Estimation is a two stage procedure: instruments are estimated using the conditional frequencies of a set of terms satisfying our assumptions, then used as a plug-in estimator. We show consistency and asymptotic normality of our two stage estimator.If the size of each document, N, does not grow too slowly relative to the time dimension T, $\sqrt{\log(T)/N} \to 0$, then there is no generated regressors problem and conventional HAC covariance estimators can be used. Our estimation procedure is straightforward and computationally simple.

We present two empirical exercises. The first empirical application is to identify monetary policy using transcripts from the FOMC meetings. We use the generalized log

odds of synonyms of '*monetary policy*', and condition on the context that synonyms of '*inflation*' and '*output*' do not appear nearby. Our estimated impulse responses do not suffer from the price puzzle (Sims (1986)), inflation decreases on impact due to a contractionary monetary policy shock. Additionally the effect output falls on impact and gradually returns to baseline.

The second empirical application is to quantity the effects of 'fresh' or contemporaneous news on individual equity returns. We use the sentiment of an article as our right hand endogenous variable and argue that it is endogenous due to reporting on past events, or 'stale' news. We use 'sentiment-charged words' from Ke, Kelly, and Xiu (2020) as an instrument, conditioning on the context that discussion of the past, terms like '*yesterday*', '*last week*', or the three weekends prior to an article's publication date do not appear in the same article. We find that by controlling for the effect of stale news, the effect that sentiment has on individual equity returns is far greater.

Related literature: There is an applied economics literature that informally links text and data to infer the responses to latent information in the text. Early work in macroeconomics uses the 'narrative record' to identify structural shocks (Friedman and Schwartz (1963) and Romer and Romer (1989, 2004, 2010)), where human readers quantify the latent variables. More recently, Baker, Bloom, and Davis (2016) and Engle et al. (2019) have used statistics derived from the text, but the relationship between text and data is 'one-sided'; these statistics enter into the quantitative model, but there is no sense in which the quantitative variables enter into the generation of text. In our model of text generation, variables from the quantitative model enter into utility of choosing a term. This allows us to document the difficulty in finding statistics from the text that are exogenous, and also provides guidance in constructing valid instruments.

Our model of text generates choice probabilities that take logit form. This form is commonly used in machine learning and statistics and is motivated as a generalized linear model with a logistic link, Taddy (2013, 2015) and Gentzkow, Shapiro, and Taddy (2019), or a neural language model with a softmax transform Bengio et al. (2003) and Mnih and Hinton (2007). Often these models are used for prediction, where the goal is to predict as accurately as possible given a set of covariates. Gentzkow, Shapiro, and Taddy (2019) differs from the other papers, in that the goal

is not merely prediction, but to infer the difference in term choice due to observable speaker level characteristics, in this case the political affiliation of the speaker. In contrast, we are not interested in explaining the variation in text, but rather using the variation in the text to infer responses of variables of interest to unobservables.

Conditioning on context is a technique used in natural language processing and machine learning, Shannon (1948), Mikolov et al. (2013b), and Mikolov et al. (2013a). Like these papers, we condition on context to allow terms to have more specific meanings than they would unconditionally.[1] Unlike these papers, we use the added specificity of context to find terms that are exogenous. These NLP models are interested in an entire vocabulary for a prediction task, and use generic conditioning events: n-gram models will condition on the previous $n-1$ words and word vector models will condition on the words that appear in a neighborhood around a given word. The contexts we use in our empirical application instead condition on the absence, rather than the presence of terms nearby. Additionally, we advocate for using contexts that are less generic and more tailored to the particular application.

Our inference results connect to the literature on generated regressors and two step estimation, Pagan (1984), Murphy and Topel (1985), and Newey and McFadden (1994). Wooldridge (2010) shows that when instruments are generated from a model with a finite dimensional parameter that, under certain moment conditions, estimation of the instrument does not effect second stage inference. Our results instead rely on a rate condition, but require weaker moment conditions for the instruments. Bai and Ng (2008) show first stage estimation of a factor model does not affect second stage inference, as long as $\sqrt{T}/N \to 0$.[2] Our paper derives similar results but with the rate condition that $\sqrt{\log(T)/N} \to 0$. Additionally we extend this type of analysis to a strong mixing DGP, which nests the i.i.d. case as a special case.

We contribute to the literature on the responses to monetary policy shocks (Romer and Romer (1989, 2004), Gertler and Karadi (2015), and Nakamura and Steinsson (2018)). Romer and Romer (1989, 2004) read through FOMC materials and manually generate a measure of the planned stance of monetary policy and compare that to actual monetary policy. The method used in this paper does not rely on human

1. The use of context is often motivated by Firth (1957); the specific meaning of a term is defined by "the company it keeps".

2. The rate $\sqrt{T}/N \to 0$ is for a linear model. For a non-linear model, the required rate condition is $T^{5/8}/N \to 0$.

readings, instead generating a measure of monetary policy from the frequencies of certain terms in a given context. We compare the impulse responses of interest rates, inflation and output with our instrument and the Romer and Romer (2004) instrument and find that the sign patterns of our responses better match economic theory, and the confidence intervals are tighter.

Our second empirical exercise quantifies the effect of fresh news on equity prices. There is a large literature on the effects of business news on asset prices (Tetlock (2007), Loughran and McDonald (2011), and Ke, Kelly, and Xiu (2020)), as well as a number of commercial vendor platforms. Most similar to our application is Ke, Kelly, and Xiu (2020) (henceforth KKX) who develop a topic model that jointly models returns and sentiment to find the words that best predict returns. KKX also investigate the effect on fresh news of equity returns, by developing a measure of similarity between articles on consecutive days to measure the freshness of news. We instead condition on the context that terms indicating stale news are not present. Like KKX, we find that fresh news has a much larger effect on asset returns than stale news.

The remainder of the paper is as follows. Section 2 outlines the quantitative and text models. Section 3 demonstrates how the text model can be used to find instruments. Section 4 discusses estimation and inference. Section 5 presents the first empirical application, identifying monetary policy shocks using FOMC transcripts. Section 6 presents the second empirical application, quantifying the effect of fresh news on equity returns. Section 7 concludes. Appendix contains proofs and additional results.

2 Model

Our main econometric specification is a linear model with a single right-hand side endogenous regressor. This is the typical starting point of much applied work; it is the linear IV in microeconometrics and Local Projections-IV in macroeconomics (Stock and Watson 2018).

For $t = 1, \ldots, T$, denote $y_t \in \mathbb{R}$ as the outcome variable, $Y_t \in \mathbb{R}$ the right-hand side

endogenous variable and $e_t \in \mathbb{R}$ the unobservables. Assume that

$$y_t = Y_t\theta_0 + e_t, \tag{1}$$

where $\mathbb{E}[Y_t e_t] \neq 0$ for $t = 1, \ldots, T$. We could also include a vector of exogenous covariates (including a constant), $x_t \in \mathbb{R}^{d_x}$ satisfying $\mathbb{E}[x_t' e_t] = 0$ for $t = 1, \ldots, T$, in which case (1) is our specification after partialling out the effect of x_t.

Our object of interest is θ_0, which cannot consistently be estimated using OLS due to the endogeneity of Y_t. One way to overcome this endogeneity problem is to use an instrument. We assume the existence of such an exogenous variable z_t^*, which can be used for an instrument; that is z_t^* is *relevant*, $\mathbb{E}[z_t^* Y_t] \neq 0$ and *exogenous*, $\mathbb{E}[z_t^* e_t] = 0$ for all $t = 1, \ldots, T$.

The departure from standard IV framework is that z_t^* cannot be directly observed. Instead the instrument z_t^* appears as a shifter in a reduced form random utility model of text generation, that determines the likelihood of each term, conditional on some observed covariates. The unobserved instrument makes terms more or less likely to appear in certain 'contexts'.

2.1 Text Model

At each time period t, there exists a *document*, each of which contains N_t *words*, each denoted as $w_{t,n}$.[3] A word is one of V *terms* in a vocabulary.[4] Note that a word refers to the index with a document (the choice), whereas term refers to a particular member of the vocabulary that is chosen (the chosen alternative).

The corpus is modelled as a discrete choice problem, where a speaker chooses a sequence of terms independently, conditional on covariates. The utility that a speaker receives from choosing term $w_{t,n}$, given an information set $\mathcal{F}_t$ is

$$U_{w_{t,n}} = x_t'\gamma + \alpha_{1,w} z_t^* + \alpha_{2,w} e_t + \xi_{w,t} + \epsilon_{w_{t,n}}, \tag{2}$$

3. If there exist a corpus of documents at time t, these are concatenated into a single document.

4. A term can be a single term—a unigram—or a combination of n terms—an n-gram. It is also common to 'preprocess' a document to transform it into an ordered collection of words. For details, see Gentzkow, Kelly, and Taddy (2019).

where $\xi_{w,t}$ reflects exogenous changes in term choice over time and is uncorrelated with the other variables in the model; that is $\mathbb{E}[\xi_{w,t}Y_t] = 0$ and $\mathbb{E}[\xi_{w,t}e_t] = 0$ for all $t = 1, \ldots, T$. The random utility shock $\epsilon_{w_{t,n}}$ is assumed to be Gumbel (appropriately centred) and is independent across $w_{t,n}$. This leads to the familiar result from random utility theory that the probability of choosing term w in document t, conditional on $\mathcal{F}_t$ is

$$p_t(w) := P(w_{t,n}|\mathcal{F}_t) = \frac{\exp\{x_t'\gamma + \alpha_{1,w}z_t^* + \alpha_{2,w}e_t + \xi_{w,t}\}}{\sum_{w'=1}^{V}\exp\{x_t'\gamma + \alpha_{1,w'}z_t^* + \alpha_{2,w'}e_t + \xi_{w',t}\}}. \tag{3}$$

The information set at time t, $\mathcal{F}_t$, contains $z_t^*, e_t, \xi_{w,t}$ as well as other observable events that we will refer to as the *context*.[5] These observable events allow us to construct very specific meanings for terms. Recalling our example from the introduction, context is crucial to finding terms that are exogenous.

In our empirical applications, Sections 5 and 6, the context will be the absence of certain 'troublesome' terms nearby (in our example from the introduction, these terms where synonyms of '*inflation*' and '*output*'). These troublesome terms often indicate discussion about the confounding unobservables e_t; so by conditioning on the absence of these terms, we can find discussion that is free from e_t. Context can capture all manner of observable events and thus is very general; we discuss in Section 6 how article novelty has been used implicitly as context in other work.

3 Valid Instruments

In order to construct valid instruments from the text, we focus our attention on terms whose variation is driven by the unobserved instrument z_t^*, but not by the confounding latent variables e_t.

Assumption 1. There exist two terms w_1 and w_2 such that

1. $\alpha_{1,w} \neq 0$ for $w = w_1$ *or* $w = w_2$.
2. $\alpha_{2,w} = 0$ for $w = w_1$ *and* $w = w_2$.

Assumption 1.1 states that there exist *two* terms w_1 and w_2, such that the likelihood of least one of them is shifted by the unobserved instrument z_t^*. Assumption 1.2

5. We use the term *context* to connect with the NLP and machine learning literatures.

states that neither the likelihood of w_1 nor w_2 depends on the confounding unobservables e_t.

Define the *log odds ratio* of w_1 and w_2 as

$$z_t(w_1, w_2) := \log\left(\frac{p_t(w_1)}{p_t(w_2)}\right). \tag{4}$$

Proposition 1. Let w_1, w_2 satisfy Assumption 1, and $\alpha_{1,w_1} \neq \alpha_{1,w_2}$. Then $z_t(w_1, w_2)$ is an instrument for Y_t, for $t = 1, \ldots, T$.

Proof. See Appendix A. □

The frequency of certain terms (as opposed to the log odds ratio of two terms) is often used in applied economic research that uses text data. Unfortunately, a similar result to Proposition 1 for conditional choice probabilities in (3), cannot be established under Assumption 1. As discussed above, the denominator in (3) contains e_t, which is not merely an artifact of the logit form of conditional probability, but instead it represents a tradeoff speakers face when choosing what to say. For example Consider choosing between two terms, w_1 and w_2. Suppose that w_1 satisfies Assumption 1, it is relevant, $\alpha_{1,w_1} > 0$, and exogenous, $\alpha_{2,w_2} = 0$; and w_2 is the opposite, $\alpha_{1,w_2} = 0$, $\alpha_{2,w_2} > 0$. Suppose that at time t, we have $Y_t = 0$ and $e_t < 0$. While the utility from speaking w_1 is unaffected by e_t, the utility from w_2 falls. In relative terms, the speaker gets more utility from uttering w_1 and hence $p_t(w_1)$ is higher. Therefore the probability of w_1 is affected by e_t, and we fail to satisfy exogeneity.

This suggests that using the exogeneity of the frequency of certain terms may be difficult to establish. We do show that under considerably stronger assumptions than those presented in Assumption 1, we can establish a similar result for conditional choice probabilities. The additional results and a discussion is presented in Appendix G.

The additional condition in statement of Proposition 1, $\alpha_{1,w_1} \neq \alpha_{1,w_2}$, ensures that the likelihood of the two words w_1 and w_2 are not shifted by z_t^* by exactly the same magnitude. This is a mild condition.

3.1 Generalized Log Odds

There are many pairs of terms that satisfy Assumption 1. This is particularly true as there are often synonyms that convey similar meanings in a given context. Similarly, there are multiple possible configurations of log odds ratios that satisfy Proposition 1 and hence can be used as instruments.[6] We consider the following generalization of the log odds ratio. Define a subset of the vocabulary, $\mathcal{J} \subseteq \{1, \ldots, V\}$ of size J. We will generalize Assumption 1 for $J > 2$.

Assumption 2. Suppose $J > 2$, and

1. $\alpha_{1,w_j} \neq 0$ for at least one $j \in \mathcal{J}$.
2. $\alpha_{2,w_j} = 0$ for all $j \in \mathcal{J}$.

Assumption 2.1 states that the likelihood of at least one of terms in $\mathcal{J}$ must be shifted by the unobserved instrument z_t^*. Assumption 2.2 states that the likelihood of all of the terms in $\mathcal{J}$ do not depend on the confounding unobservable e_t

Define the *generalized log odds* as

$$z_t(\mathcal{J}, \boldsymbol{\omega}) := \sum_{j \in \mathcal{J}} \omega_j \log(p_t(w_j)), \tag{5}$$

where $\boldsymbol{\omega} = \omega_1, \ldots, \omega_J$, $\omega_j \in \mathbb{R}$, is a vector of weights. Note that one can recover the log odds ratio from (4) by setting $\omega_1 = 1$, $\omega_2 = -1$ and $\omega_{j'} = 0$ for all $j' \in \mathcal{J} \backslash \{i, j\}$.

Proposition 2. Let $\mathcal{J}^*$ satisfy Assumption 2. For any $\boldsymbol{\omega}$ such that $\sum_{j \in \mathcal{J}^*} \omega_j = 0$ and $\sum_{j \in \mathcal{J}^*} \omega_j \alpha_{1,w_j} \neq 0$, $z_t(\mathcal{J}^*, \boldsymbol{\omega})$ is an instrument for Y_t for $t = 1, \ldots, T$.

Proof. See Appendix A. □

The additional requirement that $\sum_{j \in \mathcal{J}^*} \omega_j \alpha_{1,w_j} \neq 0$ is the generalization of the condition $\alpha_{1,w_1^*} \neq \alpha_{1,w_2^*}$, in the statement of Proposition 1.

Proposition 2 holds true for any vector of weights $\boldsymbol{\omega}$, such that $\sum_{j \in \mathcal{J}} \omega_j = 0$, $z_t(\mathcal{J}^*, \boldsymbol{\omega})$. While these weights can be chosen by the econometrician, we can find the optimal weights (in the sense of minimizing asymptotic variance), by using two

6. For example $z_t(w_1, w_2)$ and $z_t(w_2, w_1)$ both satisfy Assumption 1 and are valid instruments.

stage least squares. In particular if we include in the first stage $z_t(w_j, w_J)$ for $j = 1, \ldots, J-1$, the optimal weights are just the first stage regression coefficients. The optimality of these weights follows from the classic result of the asymptotic efficiency of 2SLS in the class of IV estimators (Wooldridge 2010)[p. 102].

This procedure will yield the same asymptotic variance as including all possible combinations of log odds ratios of terms in $\mathcal{J}^*$ in the first stage. However, the number of instruments in this case would be all possible permutations of terms in $\mathcal{J}^*$, which is $(J-1)!$, which quickly becomes infeasible for even modest J.

4 Estimation and Inference

In this section, we will discuss our estimation procedure. At each time t, we estimate the generalized log odds using the corpus of documents published at that time. This estimator will be used as a plug-in for instruments in the second stage. We will show that under an assumption about size of documents relative to the time dimension, we do not have to adjust our inference for the generated instruments. Throughout this section we will use the notation $N = \min_t N_t$.

Given a collection of terms $\mathcal{J}$ and weights $\boldsymbol{\omega}$, the *generalized log odds estimator* at time t is

$$\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) = \sum_{j \in \mathcal{J}} \omega_j \log \left(\frac{1}{N_t} \sum_{w=1}^{N_t} \mathbf{1}_{w_{t,n}=w_j} \right). \tag{6}$$

This estimator is consistent and asymptotically normal.[7]

Define the maximum weight given to a log conditional probability as $\omega_M = \max_{j \in \mathcal{J}} \omega_j$, and define L as the lower bound of the conditional probabilities on all $j \in \mathcal{J}$.[8] We have the following large deviation bound for our generalized log odds estimator.

Proposition 3. For any $\mathcal{J}$ and $\boldsymbol{\omega}$, such that $\omega_M < \infty$,

1. $P\left(\sup_t |\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) - z_t(\mathcal{J}, \boldsymbol{\omega})| > \epsilon\right) \leqslant T2 \exp\left(-2N \left(\frac{\epsilon L}{J|\omega_M|}\right)^2\right)$.

7. See Appendix F.

8. Formally, $L \geqslant 0$ is such that $L < \min_{j \in \mathcal{J}}\{p(w_j)\}$. Given that the conditional probabilities are derived from a logit discrete choice model in (3), $L > 0$.

2. $\sup_t |\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) - z_t(\mathcal{J}, \boldsymbol{\omega})| = O_p\left(\sqrt{\log(T)/N}\right)$.

Proof. See Appendix B. □

Estimation of the outcome equation (1) is complicated by the fact that we only have an estimate of our instruments $\hat{z}_t$. If we were to observe z_t, then we could use the *infeasible* 2SLS estimator

$$\hat{\theta}(z_t) = (\boldsymbol{Y}' P_{\boldsymbol{Z}} \boldsymbol{Y})^{-1} \boldsymbol{Y}' P_{\boldsymbol{Z}} y, \tag{7}$$

where the dependence on the instruments z_t has been made explicit. The infeasible 2SLS estimator is consistent and asymptotically normal[9] with variance

$$\boldsymbol{V}(\hat{\theta}(z_t)) = (\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zY})^{-1}\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zY}(\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zY})^{-1}, \tag{8}$$

where $\mathbf{Q}_{Yz} = T^{-1}\sum_{t=1}^{T}\mathbb{E}[Y_t z_t']$, $\mathbf{Q}_{zz} = T^{-1}\sum_{t=1}^{T}\mathbb{E}[z_t z_t']$ and

$$\boldsymbol{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[e_t^2 z_t z_t'] + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[e_t e_{t-\tau}(z_t z_{t-\tau}' + z_{t-\tau}z_t')].$$

We do not observe the actual value of the instruments, but instead only an estimator of them, $\hat{z}_t$, estimated *independently* at each time period t. The *feasible* 2SLS estimator is

$$\hat{\theta}(\hat{z}_t) = (\boldsymbol{Y}' P_{\hat{\boldsymbol{Z}}} \boldsymbol{Y})^{-1} \boldsymbol{Y}' P_{\hat{\boldsymbol{Z}}} y,$$

where $\hat{\boldsymbol{Z}}$ is the matrix of instruments with $\hat{z}_t$ instead of z_t.

In order to estimate $\hat{\theta}(\hat{z}_t)$, we will place minimal assumptions on the DGP. Denote a sequence of random variables $\{X_t, t \in \mathbb{Z}\}$ on a given probability space $(\boldsymbol{\Omega}, \mathcal{F}, P)$ and let $\mathcal{F}_a^b$ be the σ-algebra of events generated by $(X_a, Z_{a+1}, \ldots, X_b)$. Define the strong mixing coefficient

$$\alpha(m) = \sup_T \sup_{F \in \mathcal{F}_{-\infty}^T, G \in \mathcal{F}_{T+m}^{\infty}} |P(F \cap G) - P(F)P(G)|.$$

A sequence for which $\alpha(m) \to 0$ as $m \to \infty$ is called strong or α-mixing (Rosenblatt

9. This result follows from steps 2-4 of the proof of Theorem 1.

(1956)). Mixing processes allow considerable dependence and heterogeneity, but are sufficiently well behaved that they admit laws of large numbers and central limit theorems.

We will adapt the conditions and proof concepts of Domowitz (1982) to the 2SLS case, and include additional assumptions to ensure that first estimation of instruments does not distort second stage inference.

Assumption 3.

1. The random sequence $\{Y_t, e_t, z_t^*, \xi_t\}$ is α-mixing with $\alpha(m) = O(m^{-\lambda})$, $\lambda > 2(r+\delta)/(r+\delta-1)$, for positive constants $r > 1$ and $0 < \delta \leqslant r$.
2. For all $t = 1, \ldots, T$
   - $\mathbb{E}[|(z_t^*)^3|^{r+\delta}] < \infty$,
   - $\mathbb{E}[|\xi_{w_j,t}^3|^{r+\delta}] < \infty$ for all $j \in \{1, \ldots, J\}$,
   - $\mathbb{E}[|Y_t^3|^{r+\delta}] < \infty$,
   - $\mathbb{E}[|e_t^3|^{r+\delta}] < \infty$.
3. The average moment matrix $\mathbf{Q}_{zz} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[z_t z_t']$ is such that $\det(\mathbf{Q}_{zz}) > 0$ for sufficiently large T.
4. $\mathbb{E}[|z_t^* e_t|^{2(r+\delta)}] < \infty$ and $\mathbb{E}[|\xi_{w_j,t} e_t|^{2(r+\delta)}] < \infty$, for all $t = 1, \ldots, T$ and $j \in \{1, \ldots, J\}$.
5. Define $\mathbf{\Omega}_{a,m} = \text{var}(T^{-1/2}\sum_{t=a+1}^{a+m} z_t e_t)$. There exists a matrix $\mathbf{\Omega}$ such that $\det(\mathbf{\Omega}) > 0$ and $\lambda' \mathbf{\Omega}_{a,m} \lambda - \lambda' \mathbf{\Omega} \lambda \to 0$ and $m \to \infty$ uniformly in a for any nonzero $J-1$ vector λ.
6. For all $t = 1, \ldots, T$
   - $\mathbb{E}[|(z_t^*)^2 e_t^2|^{r+\delta}] < \infty$,
   - $\mathbb{E}[|\xi_{w_j,t}^2 e_t^2|^{r+\delta}] < \infty$ for all $j \in \{1, \ldots, J\}$.

Assumption 3 are standard requirements to invoke the WLLN and CLT for α-mixing random variables in order to show consistency, asymptotic normality and the consistency of the HAC covariance estimator.

Theorem 1. Let Assumption 3 hold, $\sqrt{\log(T)/N} \to 0$, and let $\mathcal{J}^*$ satisfy Assumption 2. For any $\boldsymbol{\omega}$ such that $\sum_{j \in \mathcal{J}} \omega_j = 0$, we have

$$\sqrt{T}(\hat{\theta}(\hat{z}_t(\mathcal{J}^*, \boldsymbol{\omega})) - \theta_0) \xrightarrow{d} N(0, \boldsymbol{V}(\hat{\theta}(z_t))).$$

Additionally if $l \to \infty$ as $T \to \infty$, such that $l = O(T^{\gamma})$, $0 < \gamma < \delta/(r+\delta)$, then

$\boldsymbol{V}(\hat{\theta}(z_t))$ can be consistently estimated by

$$\hat{\boldsymbol{V}}(\theta(\hat{z}_t)) = (\boldsymbol{Y}'P_{\hat{\boldsymbol{Z}}}\boldsymbol{Y})^{-1}\boldsymbol{Y}'P_{\hat{\boldsymbol{Z}}}\hat{\boldsymbol{\Omega}}(\hat{z}_t)P_{\hat{\boldsymbol{Z}}}\boldsymbol{Y}(\boldsymbol{Y}'P_{\hat{\boldsymbol{Z}}}\boldsymbol{Y})^{-1}$$

where

$$\hat{\boldsymbol{\Omega}}(\hat{z}_t) = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2\hat{z}_t\hat{z}_t' + \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}\hat{e}_t\hat{e}_{t-\tau}(\hat{z}_t\hat{z}_{t-\tau}' + \hat{z}_{t-\tau}\hat{z}_t').$$

Proof. See Appendix B. □

The rate assumption in Theorem 1, $\sqrt{\log(T)/N} \to 0$, requires that T not grow too quickly relative to N. In situations where the text is a corpus of policy documents, like in Section 5, this is easily satisfied as the size of these documents, N tends to be large and macro time series, T tend to be small. In situations where the text are newspaper articles, N tends to be shorter and T larger. However, T would still need to be much larger than N for our rate assumption to not be satisfied.

A popular source of text data is Twitter, where by design tweets are limited to 280 characters.[10] In this setting, we imagine documents growing large by the number of tweets at a given time t becoming large, and documents being formed by concatenating these tweets together.

Typically, we would need to adjust the variance of our estimator to account for the presence of generated regressors (Pagan 1984; Murphy and Topel 1985). When the generated regressors are instruments (as opposed to covariates) under the stronger assumption that $\mathbb{E}[e_t|z_t^*] = 0$, one can ignore first stage estimation (Wooldridge 2010)[p. 139]. This occurs when the generated instrument has the same index as the second stage estimation (i.e. not a panel structure), and the first estimation involves estimating a finite dimensional parameter. Our setting differs in that the generated instruments are estimated using cross sectional data (documents), independently at each time t, with the weaker assumption that $\mathbb{E}[e_t z_t^*] = 0$ for all $t = 1, \ldots, T$. This comes at the cost of requiring $\sqrt{\log(T)/N} \to 0$.

The results of this section and Section 3 suggest a straightforward estimation pro-

10. This is true as of August 2020. The original character limit on tweets was 140, based on the 160 character limit of SMS. The limit was modified to 280 characters in 2017.

cedure. Instruments can be estimated using only conditional frequencies, and the optimal generalized log odds ratio can be estimated using 2SLS. In addition, conventional variances and their estimators can be used without requiring adjustment.

4.1 Monte Carlo Simulations

In this section, we will present Monte Carlo simulations to investigate the finite sample properties of our estimator. We estimate the outcome equation with an intercept and the endogenous variable, with parameters $\theta_0 = [1, 2]$ and the endogenous variable is generated by a first stage $Y_t = [1, \gamma][1, z_t^*]' + \nu_t$, with $\gamma = 2$ and $z_t^* \sim N(1, 1)$. $(e_t, \nu_t)'$ are distributed normally mean 0, unit variance and correlation 0.75.

As for the text model for each term w, $\xi_{w,t}^* \sim N(0, 1)$, $\alpha_{1,w}$ and $\alpha_{2,w}$ are standard normal. For terms that will be used in the construction of our $\mathbb{E}[\alpha_{1,w}] = 2$, so that they appear sufficiently often, and $\alpha_{2,w} = 0$, so that they are exogenous.

We perform $1,000$ Monte Carlo repetitions, and set the user defined parameters similar to our main empirical application, identifying monetary policy shocks; the size of the vocabulary is $13,000$ and we use 10 terms in the construction of our instrument. We report 4 statistics, Rate $= \sqrt{\log(T)/N}$, the bias of $\hat{\theta}$, the MSE and the 95% coverage rate. We use three different values of $N = 800$, $10,000$, $26,000$ and three different values of $T = 100$, 750, $1,500$, for 9 total combinations. The results are reported in Table 1.

The results are in line with our theory. As the Rate $= \sqrt{\log(T)/N}$ decreases (the lower rows of the table), the bias and MSE decrease. 95% coverage tends to decrease slightly, however.

5 Monetary Policy Shocks

In this section, we will present our first empirical application, identifying monetary policy shocks in a Local Projections IV model, using instruments derived from FOMC meeting transcripts.

Our objects of interest are the responses of inflation, output and interest rates to

Table 1: Monte Carlo Results

	$T = 100$				$T = 750$				$T = 1500$			
N	Rate	Bias	MSE	Cov	Rate	Bias	MSE	Cov	Rate	Bias	MSE	Cov
800	0.0759	0.0145	0.0039	0.9560	0.0910	-0.0109	0.0008	0.9830	0.0956	-0.0051	0.0003	0.9570
10000	0.0215	0.0289	0.0049	0.9160	0.0257	0.0062	0.0006	0.9400	0.0270	0.0009	0.0004	0.9560
26000	0.0133	0.0221	0.0038	0.9170	0.0160	0.0052	0.0010	0.9230	0.0168	0.0024	0.0003	0.9340

Each simulation involves $1,000$ Monte Carlo repetitions, on a corpus with a vocabulary of $13,000$ terms. 50 terms are used in the construction of our instrument. 'Rate' = $\sqrt{\log(T)/N}$, 'Bias' is the sample average of $\hat{\theta}$ minus θ_0, 'MSE' is Mean Squared Error, and 'Cov' is the 95% coverage rate.

a unit contractionary monetary policy shock. We assume that the other structural shocks in the model are an inflation shock and an output shock. We will consider a system of three macroeconomic variables, the Federal Funds rate (i_t), GDP Deflator (p_t) and GDP (gdp_t)—which will we collect as $y_t = [i_t, p_t, gdp_t]'$—and three structural macroeconomic shocks, a monetary policy shock ($\epsilon_{1,t}$), an inflation ($\epsilon_{2,t}$) and an output shock ($\epsilon_{3,t}$).

We will write this system as a series of local projections,

$$y_{i,t+h} = \theta_{h,i1} y_{1,t} + u^h_{i,t+h}, \tag{9}$$

where $u^h_{i,t+h} = \sum_{j=-\infty}^{h} \epsilon_{1,t+j} + \epsilon_{2,t+j} + \epsilon_{3,t+j} - \epsilon_{1,t}$ and $\theta_{h,i1}$ is the h period ahead impulse response of y_i to an increase in interest rates by 1 unit. Our objects of interest are $\theta_{h,i1}$ for $h = 0, \ldots, H$ and $i = 1, 2, 3$. We cannot consistently estimate $\theta_{h,i1}$ by using OLS, as $y_{1,t}$ is correlated with $u^h_{i,t+h}$, and is thus endogenous.

A solution to this problem is to use Local Projection IV (LP-IV), in which we find an instrument for $y_{1,t}$. We will do so by using transcripts from the FOMC meetings.

5.1 Data

The text data are the transcripts of the FOMC meetings from March 29, 1976 to December 17, 1996.[11] We focus on this time period for a number of reasons. First, full transcripts of the meeting do not exist prior to March 29 1976, only minutes of the meetings. Secondly, we wish to compare our monetary policy shock series with the series of Romer and Romer (2004), which end in 1996. Finally, there are large changes in communication in FOMC meeting due to a change in transparency that occurred in 1993 (see Egesdal, Gill, and Rotemberg (2015) and Hansen, McMahon, and Prat (2018)). There are a total of 182 meetings within this time period. The macroeconomic data is quarterly, so we concatenate all meeting transcripts within a quarter, resulting in a total of 84 documents.[12]

We perform the following preprocessing steps. First, we separate the spoken text from the name of the speaker, to create a labelled set of 'interjections', of which there

11. The transcripts can be obtained directly from the Federal Reserve website.
12. The FRED codes for i_t, p_t and gdp_t are respectively GDPDEF, GDP, and DFF.

are 85,414 in the entire sample, and an average of 1,017 interjections per document. We then perform a number of standard 'preprocessing' steps. We transform the interjections into lower case and remove punctuation and strings that consist solely of numbers. Next we remove stop words, commonly used terms (such as '*the*', '*a*', '*an*') which do not contribute to a document meaning, and finally stem the terms (replacing terms with their 'root', e.g. '*accounts*', and '*accounting*' become '*account*').[13] This gives us an average word count per document of 26,100, and a total vocabulary of 13,344 terms.

5.2 Identification

In order to construct an instrument using the generalized log odds, we need to find a set of terms in a shared context that are relevant to monetary policy shock, but exogenous to the output and inflation shocks.

If we were to look for terms unconditionally, this would be a difficult task. For example consider a term like '*monetary policy*'. Clearly this term will be relevant to the monetary policy shock, but it is unlikely that, unconditionally, this term will be exogenous to the nuisance shocks. Consider the interest rate equation from (9); interest rates respond not only to monetary policy shocks, but the output and inflation shocks as well (this is rationalized by a Taylor rule). Therefore, if there is an inflation or an output shock, we would expect the central bank to adjust interest rates in response. Importantly we would expect that the FOMC members would be discussing '*monetary policy*' in response to these confounding shocks. Hence discussion about '*monetary policy*' varies with both inflation and output shocks, and the frequency of the term '*monetary policy*' would not be exogenous.

However by leveraging the idea of a context, we can construct instruments that are exogenous. In particular, the context we consider is that synonyms of '*output*' and '*inflation*' are not mentioned within that interjection, and the relevant set of terms are those synonyms of '*monetary policy*'. The key identifying assumption is that if policy makers are discussing '*monetary policy*' (or terms like it), but not discussing '*inflation*' or '*output*', this is indicative of the monetary policy shock, but not on an

13. We use the stop word list from the popular python package `nltk` and well as its implementation of the Porter stemmer (Porter (1980)).

inflation or output shock.

In order to properly specify our context and our set of instrument terms $\mathcal{J}^*$, we will formalize the construction of the terms above. We could use synonyms for the terms '*monetary policy*', '*output*' and '*inflation*', by using a thesaurus (Hale (1998)) or WordNet (Miller et al. (1990), Resnik (1995), and Miller (1995)). Instead, we will use an approach similar to Egesdal, Gill, and Rotemberg (2015) (henceforth EGR), who use the Oxford Dictionary of Economics (Hashimzade, Myles, and Black (2017)), denoted ODE, in order to measure the similarity between different terms. As a technical dictionary the ODE can be used to construct a more precise measurement of the similarity between terms used in the FOMC transcripts. The ODE contains $3,614$ terms, which after preprocessing results in $3,535$ unique stems.[14] Of these terms, $1,524$ appear in the FOMC transcripts.

To measure the similarity between terms in the ODE, EGR compute the pairwise cosine similarity between each term's definition. The cosine similarity is a useful tool, because it is easy to compute and satisfies a number of desirable axioms that one would like a similarity metric to satisfy.[15] However, it cannot easily distinguish between sentences that have similar meanings but different terms. The canonical example are the following two sentences "Obama speaks to the media in Illinois" and "The President greets the press in Chicago", which while conveying almost the same meaning, have no terms in common and thus would have a cosine similarity of zero.

Instead, we use the *Word Mover's Distance* (Kusner et al. (2015)), denoted WMD, to measure the similarity of term's definitions in the ODE. The WMD relies on 'word embeddings' (techniques to embed documents in low dimensional vector space), and is defined as the 'cost' to transform one document into another.[16]

14. For example '*account(s)*' and '*accounting*' both have the stem '*account*'. To break these ties, we keep the definition that has the highest similarity with either of '*monetary policy*', '*output*' and '*inflation*'.

15. EGR show that the 'generalized cosine similarity' satisfies five desirable axioms: addition, monotonicity, synonym invariance, within-word similarity, cross-word similarity. The 'vanilla' cosine similarity satisfies all axioms other than cross-words similarity and synonym invariance. The generalized cosine similarity is not appropriate for this exercise as it requires nonnegative, positive definite weighting matrix to weight the similarity between documents. In their exercise, they construct this weighting matrix from the ODE, which is exactly the exercise we are trying to accomplish. See Section 3 of EGR.

16. Details on Cosine Similarity and WMD are provided in Appendix D.

We compute the pairwise WMD between the definitions of all terms in the dictionary (that are present in the FOMC transcripts) and the definitions of the three terms, '*monetary policy*', '*output*' and '*inflation*'.[17] We rank the WMD, and choose terms with the lowest WMD as the synonyms for '*monetary policy*', '*output*' and '*inflation*'. It is possible for there to be overlap between terms that are synonyms for '*monetary policy*' and '*output*' or '*inflation*'. These terms are clearly not suitable as instruments, and we remove these terms until we have a sufficient number of terms. We chose to use 50 synonyms of '*output*' and '*inflation*', and 10 synonyms of '*monetary policy*'. The higher number of synonyms for '*output*' and '*inflation*' are to ensure that our instruments are in fact exogenous.

These synonyms are plotted in a wordcloud in Figure 1, and are also listed in Appendix D.1. The size of the terms in the wordcloud are inversely proportion to their WMD to either '*monetary policy*' (top), '*output*' (middle) or '*inflation*' (bottom).

Therefore our instrument is constructed in the following way: $\mathcal{J}^*$ is the set of synonyms of '*monetary policy*', and the context is none of the set of synonyms of '*inflation*' or '*output*' appear within the same interjection. $73,114$ or 86% of the interjections satisfy our context. $3,368$ (4%) of interjections both satisfy our context and contain a synonym of '*monetary policy*'.[18]

5.3 Results

We estimate the LP-IV model and compute impulse responses up to 12 quarters after impact. The results are presented in Figure 2.

The responses of interest rates, inflation and output demonstrate their expected sign patterns, with interest rates increasing on impact, then gradually declining and inflation and output falling on impact, and gradually returning to baseline. Our impulse responses do not exhibit the prize puzzle, an an unexpected monetary

17. Because we will be looking at the frequencies of synonyms of '*monetary policy*' in interjections that do not contain synonyms for '*output*' or '*inflation*', this may result in terms with zero frequency in the entire corpus. This occurs when a '*monetary policy*' appears in the transcript, but only in interjections that contain 'bad' synonyms. We remove these terms.

18. To handle situations where frequencies of terms are zero, and the $\log(p_t(w))$ is undefined, we smooth the multinominal frequencies. See Appendix E for details.

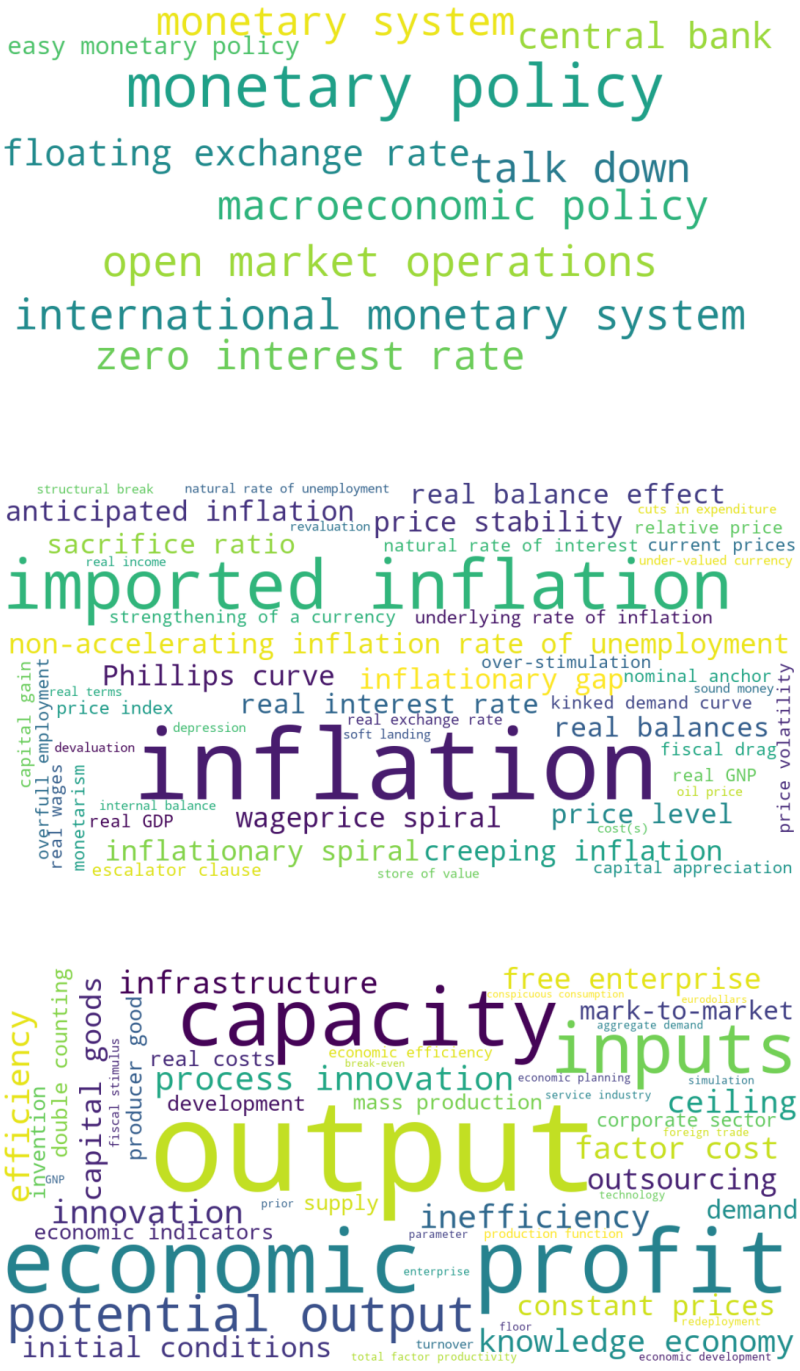

Figure 1: Synonyms for '*monetary policy*' (top), '*inflation*' (middle) and '*output*' (bottom). The size of terms is inversely related to the WMD between a term and it's target.

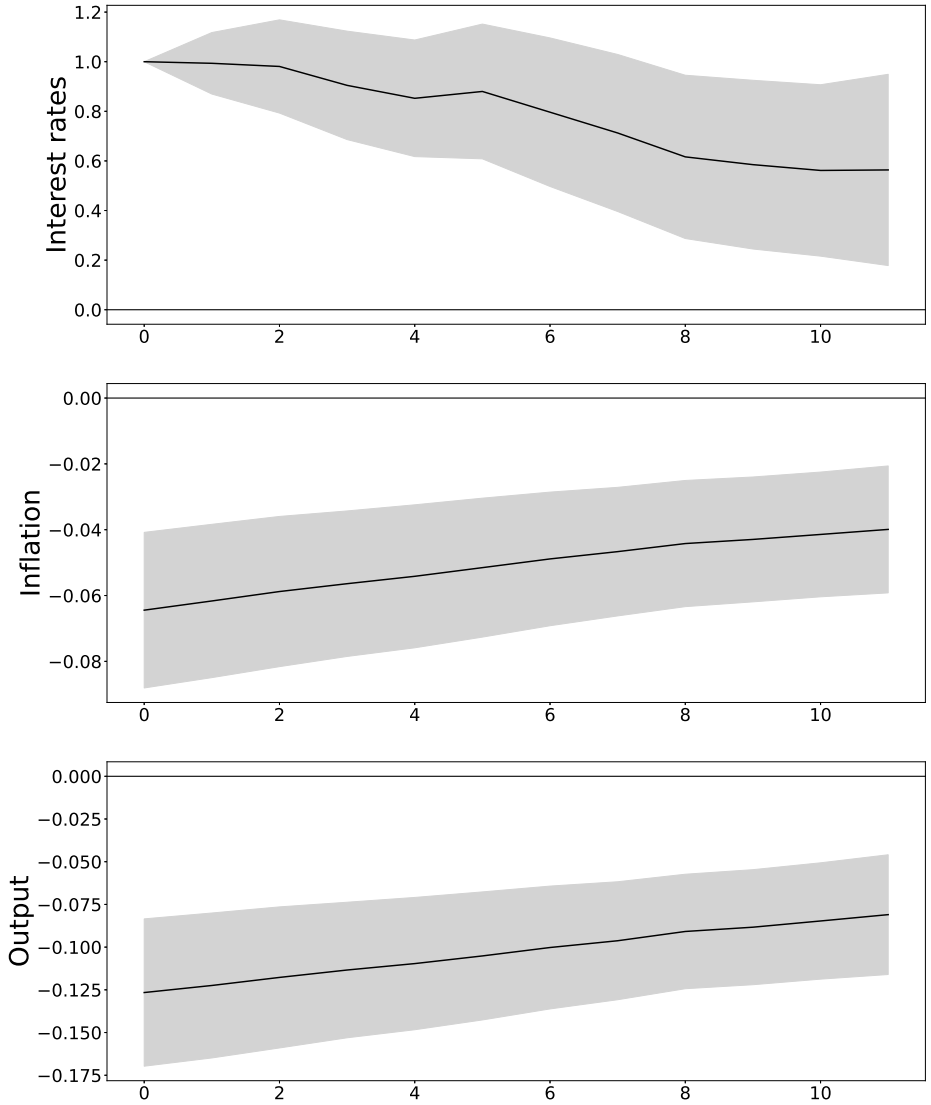

Figure 2: Impulse responses of interest rates (top), inflation (middle) and output (bottom) to a one unit contractionary monetary policy shock identified using generalized log odds ratio. Horizon is quarters. Shaded area represents the 95% confidence interval.

tightening often leads to a counter intuitive increase in inflation (Sims (1986)). We can compare these impulse responses to those identified using the Romer and Romer (2004) monetary policy shock series, presented in Figure 3 in Appendix C.[19]

We do note that both inflation and output fall on impact before gradually increasing, indicating a very quick adjustment of these macro to the monetary policy shock. We would except the response of inflation and output to be less stark and more gradual. This is likely due to the simplicity of the model, both in terms of only having 3 variables and very limited dynamics. This is an area for future work.

6 Effect of News on Equity Returns

In this section, we will present our second empirical application quantifying the effect of contemporaneous or 'fresh' news on equity returns.

Denote y_t as the return of a company at date t, x_t a vector of covariates, and Y_t as the 'sentiment' of newspaper articles about the company published at time t. The model is identical to (1) (with covariates added)

$$y_t = x_t'\beta_0 + \theta_0 Y_t + e_t.$$

The use of sentiment as a measure of news is standard in the literature, dating back to the seminal work of Tetlock (2007). However, it is difficult to use sentiment to estimate the causal effect of fresh news on equity prices, θ_0, as Y_t is endogenous, due to the effect of 'stale' news. The sentiment of articles about a company published at time t captures discussion about current events, but also captures news events that have happened in the past, which we define as 'stale'. Information is known to disseminate through the market slowly, and hence the return at time t depends not only on fresh news, but also on stale news, which is captured in e_t. Hence Y_t and e_t are correlated.

To solve this endogeneity problem, we will construct an instrument from the frequencies of 'sentiment-charged terms' by conditioning on the context that 'discussions of the past' do not appear in the same article. We will expand on this in Section 6.2.

19. A comparison of our text instrument and the Romer and Romer (2004) monetary policy shock series is presented in Figure 4 in Appendix C.

6.1 Data

The data set is a corpus of financial and economic news extracted from Reuters, used in Ding et al. (2014). It spans from 20 October 2006 to 19 November 2013, with the number of articles totalling 105, 375. First, we remove a small number of articles with blank text, and then match the articles with the 'main' firm of interest. We identify the main firm of interest, by extracting identifiers from the url of the articles and matching them.[20] We extract any company that has at least 100 occurrences of an identifier (note that some companies have multiple identifiers, "boeing" and "dreamliner" or "apple" and "iphone", which we combine), leaving a total of 86 companies. We further remove articles that have more than one company identifier. The total number of articles that can be matched uniquely to a single firm is 25, 852.

We concatenate all articles posted about a firm at time t and match these documents with the firm's adjusted close-to-close returns posted at time t, retrieved from CRSP.[21] Articles posted on weekends and trading holidays are attributed to the next closest trading day. Some companies have news articles posted about them but do not have returns associated with the date, as the stock has been removed from trading (for example due to bankruptcy, of which there were several during this time period); these articles are dropped from the sample. After dropping these articles and concatenating articles into documents, the total number of document-return pairs in our sample is 14, 593.

6.2 Identification and Results

Our endogenous measure of the sentiment of an article is the fitted 'sentiment score' of KKX, a measure from 0 to 1, where a sentiment score of 1 indicates that an

20. For example, the article hosted at https://www.reuters.com/article/us-energy-bp-idUSWLA488420061024, we identify using "us-energy-bp", simplifying to "bp".

21. This differs from the strategy of KKX, who match articles posted on day t with a companies three day returns from market close on day $t-2$ to market close on day $t+1$. They follow this strategy because they do not know the timing by which new information is impounded in prices: if prices adjust slowly to news it makes sense to align with future returns, and if articles are a restatement of past news it makes sense to align with past returns.

Positive		Negative	
Word	Score	Word	Score
undervalue	0.596	shortfall	0.323
repurchase	0.573	downgrade	0.382
surpass	0.554	disappointing	0.392
upgrade	0.551	tumble	0.402
rally	0.548	blame	0.414
surge	0.547	hurt	0.414
treasury	0.543	plummet	0.423
customary	0.539	auditor	0.424
imbalance	0.538	plunge	0.429
jump	0.538	waiver	0.429

Table 2: Top 10 positive and negative sentiment-charged terms from KKX Table A2. The full list of top 50 50 top positive and negative sentiment-charged is presented in Tables 4 and 5 in Appendix C.

article has maximally positive sentiment and a sentiment score of 0 is maximally negative. To construct this sentiment score, they use a supervised learning framework to identify which terms are most predictive of returns—which they label 'sentiment-charged terms—and they also compute the term's weight. These are split into two categories, positive and negative sentiment terms depending on whether the term increases or decreases a company return. They report 50 terms in each category. The measure of sentiment that we use, for a particular company in document t, is the weighted sum of the frequency of these highly charged sentiment terms, using KKX weights.[22] The 10 top positive and negative sentiment-charged terms are presented in Table 2.[23]

Our instrument will be the generalized log odds ratio of the *conditional* probabilities of the same sentiment-charged terms, where the context we condition on is that the terms '*yesterday*', '*last week*' ,'*day before yesterday*', '*in a row*', and the three weekdays prior to the article do not occur within that article.[24] Of the total 25, 852

22. We do standard processing on the documents to identify the frequencies of these highly charged sentiment terms. This includes: removing punctuation and whitespace, converting to lower case, using the Miller (1995) lemmatizer and the Porter (1980) stemmer from the `NTLK` Python package.

23. The full list of the 50 top positive and negative sentiment-charged terms are presented in Tables 4 and 5 in Appendix C.

24. For example if the article was posted on a Thursday, the three prior weekdays would

articles in the sample, 16,159 or 62.5% of articles satisfy this particular context.

Table 3 presents results of our regressions. In the first two columns, we present the reduced form regression of equity returns against the KKX sentiment score, without company fixed effects in the first column and with company fixed effects in the second column. The coefficient on sentiment can be interpreted as the effect on the return of an equity due to an article changing from maximally negative (a sentiment value of 0) to maximally positive (a sentiment value of 1). In this specification, we cannot distinguish between news that is fresh and news that stale news.

The third and fourth columns present the IV regression, with our generalized log odds as an instrument, without and with company fixed effect respectively. We can interpret the coefficient on sentiment as the effect on equity from a change in sentiment, when we are only considering fresh news. Importantly, we find a much larger effect of news on equity returns when considering only fresh news as opposed to fresh and stale news.

Table 3: The Effect of Fresh News on Returns

	Reduced Form		IV	
	No FE	FE	No FE	FE
Sentiment	0.507***	0.574***	0.999***	1.145***
	(0.105)	(0.0975)	(0.288)	(0.314)
F - First stage			40.02	38.16
Observations	14593	14593	14593	14593

The effect of fresh news on equity returns. The 'Reduced Form' columns are the reduced form regression of a company's equity returns on KKX sentiment score, without and with company fixed effects respectively. The 'IV' columns are the instrumental variables regression, where the instrument is our generalized log odds ratio of positive and negative sentiment-charged terms in the context that no discussion of past events has occurred within the article. Standard errors clustered at the company level are reported in parenthesis. Significance denoted as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

be '*Monday*', '*Tuesday*', '*Wednesday*'.

Our results are very closely in line to those of KKX, despite using a different identification scheme and smaller data set with less frequent reporting. KKX find that the effect of fresh news on returns is 70% larger than stale news, whereas we find that the effect of fresh news is 100% larger than fresh and stale news combined in the FE specification.

Our approach complements the work of KKX who distinguish between fresh and stale news by computing article novelty. They define Article novelty as the maximum cosine similarly between an article at time t and any article published about the same firm within the last 5 trading days. They classify articles as fresh news if they have a sufficiently high article novelty (they are sufficiently dissimilar to all articles within the last 5 days) and stale news otherwise. Through the lens of our text model, this is similar to defining the context as article novelty being above a certain threshold.

We instead condition on observables within the same article (instead of articles within the last 5 days). This is useful when the corpus is less comprehensive.[25] Their context will capture fresh news when the news is reported often, whereas we can still identify fresh news even when reporting is more sporadic. On the other hand, our context depends on a dictionary of past tense terms and may misclassify stale news as fresh, if none of our pre-specified 'past tense terms' are present, but the article is similar to previous articles within the last 5 days. In these situations, the KKX novelty measure may perform better.[26]

Our context can be generalized to identify a particular topic of news, beyond simply fresh and stale news. For example if we were interested in news related to climate change risk (like in Engle et al. (2019)), but not related to policy uncertainty (like in Baker, Bloom, and Davis (2016)), one could use as instruments terms like '*weather*' and as a context that terms like *'Federal Reserve*' do not appear in the same article. Traditional topic models (Hansen, McMahon, and Prat (2018) and Ke, Olea, and Nesbit (2020)) can be used to find relevant topics, but struggle to identify exogenous topics. Our model can handle both relevance and exogeneity, but the terms defining topics need to generated by the researcher.

25. KKX use the *Dow Jones Newswire* corpus which contains $10,364,18$ articles, whereas we only have access to $105,375$ articles.

26. In particular one could improve upon the KKX novelty measure by using the Word Mover's distance as a measure of article novelty. See Section 5.2 and Appendix D for discussion.

7 Conclusion

This paper presents a theoretical framework that formalizes an explicit link between text data and unobservables in a quantitative econometric model. Our model of text highlights that finding terms in the text that can be used as instruments is challenging, but offers a path forward by conditioning on particular contexts.

The paper uses the model to provide several new insights about constructing instruments using text data. First, we show that under weak conditions, the generalized log odds of certain terms in the appropriate context can be used as an instrument. Additionally the probabilities of terms (as opposed to the log odds), which have been used in the literature, require much stronger assumptions to be used as instruments.

Estimation is a two step estimation procedure, we first estimate the instruments, then use them as plug-in estimators. We provide bounds on the worst case estimation error of the generalized logs odds ratio, and use this to show that the estimation of the instruments in the first stage does not affect inference in the second stage when $\sqrt{\log(T)/N} \to 0$. Instruments can be computed using conditional frequencies, and simply running 2SLS, without the need to adjust HAC covariance estimators.

We present two empirical applications. The first to identify a monetary policy shock using the FOMC transcripts. We find that our impulse responses do not suffer from the price puzzle, have the sign patterns that theory would suggest and have tight confidence intervals compared to alternatives. The second application quantifies the effect of fresh news on equity returns. We find that by controlling for the effect of stale news, the effect of news on equity returns is far larger.

There remain interesting avenues for future research. The text model we develop captures that text and data are jointly determined and that the data affects the text and vice versa, but it is not a structural model of text generation. Formally modelling the choice problem that agents face in communicating would allow a better understanding of how agents respond to unobservables. Additionally, the contexts that we use in applications are related to absence of certain terms nearby. We feel this is only scratching the surface of the possible contexts that could be useful in economic research.

References

Andreas, Jacob, and Dan Klein. 2015. "When and Why are Log-Linear Models Self-Normalizing?" In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 244–249.

Bai, Jushan, and Serena Ng. 2008. "Extremum Estimation When the Predictors Are Estimated From Large Panels." *Annals of Economics and Finance* 9 (2): 201–222.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring Economic Policy Uncertainty." *The Quarterly Journal of Economics* 131 (4): 1593–1636.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3:1137–1155.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5:135–146.

Chen, Stanley F, and Joshua Goodman. 1999. "An Empirical Study of Smoothing Techniques for Language Modeling." *Computer Speech & Language* 13 (4): 359–394.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171–4186. June.

Ding, Xiao, Yue Zhang, Ting Liu, and Junwen Duan. 2014. "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP):* 1415–1425.

Domowitz, Ian. 1982. "The Linear Model With Stochastic Regressors and Heteroscedastic Dependent Errors." *Center for Mathematical Studies in Economics and Management Science Discussion Paper,* no. 543: 257–75.

Domowitz, Ian, and Halbert White. 1982. "Misspecified models with dependent observations." *Journal of Econometrics* 20 (1): 35–58.

Egesdal, Michael, Michael Gill, and Martin Rotemberg. 2015. "How Federal Reserve Discussions Respond to Increased Transparency." *Available at SSRN 2676429.*

Engle, Robert F, Stefano Giglio, Bryan T Kelly, Heebum Lee, and Johannes Stroebel. 2019. "Hedging Climate Change News." *NBER Working Paper w25734.*

Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930-1955." *Studies in Linguistic Analysis:* 1–32.

Friedman, Milton, and Anna J Schwartz. 1963. "A Monetary History of the United States, 1867–1960." *NBER Books.*

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87 (4): 1307–1340.

Gertler, Mark, and Peter Karadi. 2015. "Monetary Policy Surprises, Credit Costs, and Economic Activity." *American Economic Journal: Macroeconomics* 7 (1): 44–76.

Hale, Michael Mc. 1998. "A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity." *arXiv preprint cmp-lg/9809003.*

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics* 133 (2): 801–870.

Hashimzade, Nigar, Gareth Myles, and John Black. 2017. *A Dictionary of Economics.*

Ke, Shikun, José Luis Montiel Olea, and James Nesbit. 2020. "A Robust Machine Learning Algorithm for Text Analysis." *Working Paper.*

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu. 2020. "Predicting Returns with Text Data." *NBER Working Paper w26186.*

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. "From Word Embeddings to Document Distances." In *International Conference on Machine Learning,* 957–966.

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (1): 35–65.

McLeish, Don L. 1975. "A Maximal Inequality and Dependent Strong Laws." *Annals of Probability* 3, no. 5 (October): 829–839.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. "Efficient Estimation of Word Representations in Vector Space." In *1st International Conference on Learning Representations, ICLR 2013.*

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. "Distributed Representations of Words and Phrases and their Compositionality." In *Advances in Neural Information Processing Systems,* 26:3111–3119.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41.

Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. "Introduction to WordNet: An On-Line Lexical Database." *International Journal of Lexicography* 3 (4): 235–244.

Mnih, Andriy, and Geoffrey Hinton. 2007. "Three New Graphical Models for Statistical Language Modelling." In *Proceedings of the 24th International Conference on Machine Learning,* 641–648.

Murphy, Kevin M, and Robert H Topel. 1985. "Estimation and Inference in Two-Step Econometric Models." *Journal of Business & Economic Statistics* 3 (4): 370–379.

Nakamura, Emi, and Jón Steinsson. 2018. "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect." *The Quarterly Journal of Economics* 133 (3): 1283–1330.

Newey, Whitney K, and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." *Handbook of Econometrics* 4:2111–2245.

Pagan, Adrian. 1984. "Econometric Issues in the Analysis of Regressions with Generated Regressors." *International Economic Review* 25 (1): 221–247.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543.

Porter, Martin F. 1980. "An Algorithm for Suffix Stripping." *Program* 14 (3): 130–137.

Ramey, Valerie A. 2016. "Macroeconomic Shocks and Their Propagation." In *Handbook of Macroeconomics,* 2:71–162. Elsevier.

Resnik, Philip. 1995. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1,* 448–453.

Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review* 94 (4): 1055–1084.

———. 2010. "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." *American Economic Review* 100 (3): 763–801.

———. 1989. "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." *NBER Macroeconomics Annual* 4:121–170.

Rosenblatt, Murray. 1956. "A Central Limit Theorem and a Strong Mixing Condition." *Proceedings of the National Academy of Sciences* 42 (1): 43–47.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

Sims, Christopher A. 1986. "Are Forecasting Models Usable for Policy Analysis?" *Quarterly Review of the Federal Reserve Bank of Minneapolis* 10 (1): 2–16.

Stock, James H, and Mark W Watson. 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments." *The Economic Journal* 128 (610): 917–948.

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108 (503): 755–770.

———. 2015. "Distributed Multinomial Regression." *The Annals of Applied Statistics* 9 (3): 1394–1414.

Tetlock, Paul C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance* 62 (3): 1139–1168.

White, Halbert, and Ian Domowitz. 1984. "Nonlinear Regression with Dependent Observations." *Econometrica: Journal of the Econometric Society:* 143–161.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

Appendix

A Proofs in Section 3

Proof of Proposition 1

Proof. Under Assumption 1

$$z_t(w_1, w_2) = z_t^*(\alpha_{1,w_1} - \alpha_{1,w_2}) + (\xi_{w_1,t} - \xi_{w_2,t}). \tag{10}$$

We need to show that $z_t(w_1, w_2)$ satisfies the conditions to be an instrument: it is relevant, $\mathbb{E}[z_t(w_1, w_2)Y_t] \neq 0$ and exogenous, $\mathbb{E}[z_t(w_1, w_2)e_t] = 0$, for $t = 1, \ldots, T$.

Relevance: For $t = 1, \ldots, T$

$$\begin{aligned}
\mathbb{E}[z_t(w_1, w_2)Y_t] &= \mathbb{E}[(z_t(\alpha_{1,w_1} - \alpha_{1,w_2}) + (\xi_{w_1,t} - \xi_{w_2,t}))Y_t] \\
&= \mathbb{E}[z_t^* Y_t](\alpha_{1,w_1} - \alpha_{1,w_2}) + \mathbb{E}[\xi_{w_1,t}Y_t] - \mathbb{E}[\xi_{w_2,t}Y_t] \\
&= \mathbb{E}[z_t^* Y_t](\alpha_{1,w_1} - \alpha_{1,w_2}) \\
&\quad \text{(by assumption } \mathbb{E}[\xi_{w,t}Y_t] = 0 \text{ for all } w).
\end{aligned}$$

By assumption z_t^* can be used as an instrument for Y_t, hence $\mathbb{E}[z_t^* Y_t] \neq 0$. Also $(\alpha_{1,w_1} - \alpha_{1,w_2}) \neq 0$ by assumption. Together these imply $\mathbb{E}[z_t(w_1, w_2)Y_t] \neq 0$.

Exogeneity: For $t = 1, \ldots, T$

$$\begin{aligned}
\mathbb{E}[z_t(w_1, w_2)e_t] &= \mathbb{E}[(z_t^*(\alpha_{1,w_1} - \alpha_{1,w_2}) + (\xi_{w_1,t} - \xi_{w_2,t}))e_t] \\
&= \mathbb{E}[z_t^* e_t](\alpha_{1,w_1} - \alpha_{1,w_2}) + \mathbb{E}[\xi_{w_1,t}e_t] - \mathbb{E}[\xi_{w_2,t}e_t] \\
&= \mathbb{E}[z_t^* e_t](\alpha_{1,w_1} - \alpha_{1,w_2}) \\
&\quad \text{(by assumption } \mathbb{E}[\xi_{w,t}e_t] = 0 \text{ for all } w).
\end{aligned}$$

Once again by assumption z_t^* can be used as an instrument for Y_t and hence $\mathbb{E}[z_t^* e_t] = 0$. This implies $\mathbb{E}[z_t(w_1, w_2)e_t] = 0$. □

Proof of Proposition 2

Proof. Let $\boldsymbol{\omega}$ be such that $\sum_{j\in\mathcal{J}}\omega_j = 0$. We can write the generalized log odds as

$$\begin{aligned} z_t(\mathcal{J},\boldsymbol{\omega}) &= \sum_{j\in\mathcal{J}}\omega_j u_{w_j} \\ &= z_t^* \sum_{j\in\mathcal{J}}\omega_j\alpha_{1,w_j} + \sum_{j\in\mathcal{J}}\omega_j\xi_{w_j,t} \end{aligned}$$

The proof will follow the same strategy of the proof of Proposition 1. We need to show that $z_t(\mathcal{J}^*,\boldsymbol{\omega})$ satisfies the conditions to be an instrument: it is relevant, $\mathbb{E}[z_t(\mathcal{J}^*,\boldsymbol{\omega})Y_t] \neq 0$ and exogenous $\mathbb{E}[z_t(\mathcal{J}^*,\boldsymbol{\omega})e_t] = 0$, for $t = 1,\dots,T$.

Relevance: For $t = 1,\dots,T$

$$\begin{aligned} \mathbb{E}[z_t(\mathcal{J}^*,\boldsymbol{\omega})Y_t] &= \mathbb{E}\left[\left(z_t^*\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} + \sum_{j\in\mathcal{J}^*}\omega_j\xi_{w_j,t}\right)Y_t\right] \\ &= \mathbb{E}[z_t^*Y_t]\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} + \sum_{j\in\mathcal{J}^*}\omega_j\mathbb{E}[\xi_{w_j,t}Y_t] \\ &= \mathbb{E}[z_t^*Y_t]\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} \\ &\quad \text{(by assumption } \mathbb{E}[\xi_{w,t}Y_t] = 0 \text{ for all } w\text{)}. \end{aligned}$$

By assumption z_t^* can be used as an instrument for Y_t, hence $\mathbb{E}[z_t^*Y_t] \neq 0$. Also $\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} \neq 0$ by assumption. Together these imply $\mathbb{E}[z_t(\mathcal{J}^*,\boldsymbol{\omega})Y_t] \neq 0$.

Exogeneity: For $t = 1,\dots,T$

$$\begin{aligned} \mathbb{E}[z_t(\mathcal{J}^*,\boldsymbol{\omega})e_t] &= \mathbb{E}\left[\left(z_t^*\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} + \sum_{j\in\mathcal{J}^*}\omega_j\xi_{w_j,t}\right)e_t\right] \\ &= \mathbb{E}[z_t^*e_t]\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} + \sum_{j\in\mathcal{J}^*}\omega_j\mathbb{E}[\xi_{w_j,t}e_t] \\ &= \mathbb{E}[z_t^*e_t]\sum_{j\in\mathcal{J}^*}\omega_j\alpha_{1,w_j} \\ &\quad \text{(by assumption } \mathbb{E}[\xi_{w,t}Y_t] = 0 \text{ for all } w\text{)}. \end{aligned}$$

Once again by assumption z_t^* can be used as an instrument for Y_t and hence $\mathbb{E}[z_t^*e_t] =$

0. This implies $\mathbb{E}[z_t(\mathcal{J}^*, \boldsymbol{\omega})e_t] = 0$. $\square$

B Proofs in Section 4

Proof of Proposition 3

We will first establish a number of lemmas.

Lemma 1. Let $\mathbf{p}$ be a vector of multinational probabilities, such that $\sum_{j=1}^{V} p(w_j) = 1$ and let $\hat{\mathbf{p}}$ be the natural estimator. Then $P(\sup_j |\hat{p}(w_j) - p(w_j)| > \alpha) \leqslant 2\exp\left(-2N\alpha^2\right)$. This also holds for any subvector of $\mathbf{p}$.

Proof. $\sup_j |\hat{p}(w_j) - p(w_j)|$ has the bounded differences property. Recall that each $w_{t,n}$ is a one-hot vector of size V. The ith element of $w_{t,n}$ is either 0 or 1, and can change by at most 1 in absolute value. Therefore $|\hat{p_j} - p_j|$ changes by at most $c_n = 1/N$ for any j. Therefore using the bounded difference inequality

$$\begin{aligned} P\left(\sup_j |\hat{p}(w_j) - p(w_j)| > \alpha\right) &\leqslant \exp\left(\frac{2\alpha^2}{\sum_{i=n}^{N} c_n^2}\right) \\ &= \exp\left(\frac{2\alpha^2}{\sum_{i=n}^{N} (1/N)^2}\right) \\ &= 2\exp\left(-2N\alpha^2\right). \end{aligned}$$

□

Lemma 2. Let $\log(x)$ have domain $D = [L, \infty)$, then

$$|\log(x) - \log(y)| \leqslant \frac{1}{L}|x - y|,$$

for all $x, y \in D$.

Proof. Assume without loss of generality that $0 < L \leqslant x \leqslant y$. Then

$$\begin{aligned} |\log(x) - \log(y)| &= \log(y/x) \\ &= \log(1 + (y/x - 1) \\ &\leqslant y/x - 1 \\ &\text{(as } \log(1+u) \leqslant u \text{ for all } u > -1\text{)}, \end{aligned}$$

$$
\begin{aligned}
&= 1/x(y-x) \\
&\leqslant \frac{1}{L}(y-x) \\
&\text{(as } L \leqslant x \text{ by assumption)}, \\
&= \frac{1}{L}|y-x|.
\end{aligned}
$$

□

Lemma 3. Recall $z_t(\mathcal{J}, \boldsymbol{\omega}) = \sum_{j \in \mathcal{J}} \omega_j \log(\hat{p}_t(w_j))$. Then for each $t = 1, \ldots, T$

$$
P(|\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) - z_t(\mathcal{J}, \boldsymbol{\omega})| > \epsilon) \leqslant 2 \exp\left(-2N\left(\frac{\epsilon L}{J\omega_M}\right)^2\right).
$$

Proof.

$$
\begin{aligned}
P(|\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) - z_t(\mathcal{J}, \boldsymbol{\omega})| > \epsilon) &= P\left(|\sum_{j \in \mathcal{J}} \omega_j \log(\hat{p}_t(w_j)) - \sum_{j \in \mathcal{J}} \omega_j \log(p_t(w_j))| > \epsilon\right) \\
&= P\left(|\sum_{j \in \mathcal{J}} \omega_j (\log(\hat{p}_t(w_j)) - \log(p_t(w_j)))| > \epsilon\right) \\
&\leqslant P\left(\sum_{j \in \mathcal{J}} |\omega_j| |\log(\hat{p}_t(w_j)) - \log(p_t(w_j))| > \epsilon\right) \\
&\text{(by the triangle inequality and } |xy| = |x||y|), \\
&\leqslant P\left(J|\omega_M| \sup_j |\log(\hat{p}_t(w_j)) - \log(p_t(w_j))| > \epsilon\right) \\
&= P\left(\sup_j |\log(\hat{p}_t(w_j)) - \log(p_t(w_j))| > \frac{\epsilon}{J|\omega_M|}\right) \\
&\leqslant P\left(\frac{1}{L} \sup_j |\hat{p}_t(w_j) - p_t(w_j)| > \frac{\epsilon}{J|\omega_M|}\right) \\
&\text{(by Lemma 2)}, \\
&\leqslant 2 \exp\left(-2N\left(\frac{\epsilon L}{J|\omega_M|}\right)^2\right) \\
&\text{(by Lemma 1 with } \alpha = \frac{L\epsilon}{J|\omega_M|}).
\end{aligned}
$$

□

Proof of Proposition 3

Proof.

1.

$$
\begin{aligned}
P\left(\sup_t |\hat{z}_t(\mathcal{J},\boldsymbol{\omega}) - z_t(\mathcal{J},\boldsymbol{\omega})| > \epsilon\right) &= P\left(\bigcup_{t=1}^{T} |\hat{z}_t(\mathcal{J},\boldsymbol{\omega}) - z_t(\mathcal{J},\boldsymbol{\omega})| > \epsilon\right) \\
&\leqslant \sum_{t=1}^{T} P\left(|\hat{z}_t(\mathcal{J},\boldsymbol{\omega}) - z_t(\mathcal{J},\boldsymbol{\omega})| > \epsilon\right) \\
&\quad \text{(by Boole's inequality)}, \\
&\leqslant T2\exp\left(-2N\left(\frac{\epsilon L}{J|\omega_M|}\right)^2\right) \\
&\quad \text{(by Lemma 3)}.
\end{aligned}
$$

2. Follows if we choose $\epsilon = \frac{J|\omega_M|\sqrt{C^2+2\log(2T)}}{2\sqrt{N}L}$ for a sufficiently large constant C. We can see this by setting

$$
\exp(-C^2/2) = T2\exp\left(-2N\left(\frac{\epsilon L}{J|\omega_M|}\right)^2\right)
$$

and solving for ϵ.

□

Proof of Theorem 1

Lemma 4. If $\mathbb{E}[|(z_t^*)^q|^{r+\delta}] < \infty$ and $\mathbb{E}[|(\xi_{w,t})^q|^{r+\delta}] < \infty$ for all w, then $\mathbb{E}[|z_{j,t}^q|^{r+\delta}] < \infty$ for all $t = 1,\ldots,T$ and for each $j \in \{1,\ldots,J-1\}$.

Proof.

$$
\begin{aligned}
\mathbb{E}[|z_{j,t}^q|^{r+\delta}] &= \mathbb{E}\left[|z_t^*(\alpha_{1,w_j} - \alpha_{1,w_J}) + (\xi_{w_j,t} - \xi_{w_J,t})|^{q(r+\delta)}\right] \\
&\leqslant \mathbb{E}\left[2^{q(r+\delta)-1}|z_t^*(\alpha_{1,w_j} - \alpha_{1,w_J})|^{q(r+\delta)} + 2^{q(r+\delta)-1}|\xi_{w_j,t} - \xi_{w_J,t}|^{q(r+\delta)}\right] \\
&\quad (\text{as } |a+b|^q \leqslant 2^{q-1}(|a|^q + |b|^q) \text{ for } q \geqslant 1) \\
&= 2^{q(r+\delta)-1}|\alpha_{1,w_j} - \alpha_{1,w_J}|^{q(r+\delta)}\mathbb{E}[|z_t^*|^{q(r+\delta)}]
\end{aligned}
$$

$$
\begin{aligned}
&+ 2^{2q(r+\delta)-2}(\mathbb{E}[|\xi_{w_j,t}|^{q(r+\delta)}] + \mathbb{E}[|\xi_{w_J,t}|^{q(r+\delta)}]) \\
&\text{(by linearity of expectation, } |xy|^p = |x|^p|y|^p \text{ and} \\
&|a-b|^q \leqslant 2^{q-1}(|a|^q + |b|^q) \text{ for } q \geqslant 1)
\end{aligned}
$$

which is finite as $\mathbb{E}\left[|z_t^*|^{q(r+\delta)}\right] < \infty$ and $\mathbb{E}\left[|\xi_{w_j,t}|^{q(r+\delta)}\right] < \infty$ for each $j \in \{1, \ldots, J\}$. □

Proof. Consistency: Our outcome equation is

$$y_t = Y_t\theta_0 + e_t.$$

The feasible 2SLS estimator is

$$
\begin{aligned}
\hat{\theta}(\hat{z}_t) &= (\boldsymbol{Y}'\hat{\boldsymbol{Z}}(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{Y})^{-1}\boldsymbol{Y}'\hat{\boldsymbol{Z}}(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'y \\
&= (\boldsymbol{Y}'\hat{\boldsymbol{Z}}(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{Y})^{-1}\boldsymbol{Y}'\hat{\boldsymbol{Z}}(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}\hat{\boldsymbol{Z}}'(\boldsymbol{Y}'\theta_0 + e).
\end{aligned}
$$

We will show (at the end of the proof) at that $\frac{1}{T}\hat{\boldsymbol{Z}}'\boldsymbol{Y} \xrightarrow{p} \mathbf{Q}_{Yz}$ and $\frac{1}{T}\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}} \xrightarrow{p} \mathbf{Q}_{zz}$. Taking this as given,

$$\hat{\theta}(\hat{z}_t) - \theta_0 = (\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zY})^{-1}\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\frac{1}{T}\sum_{t=1}^{T}\hat{z}_t e_t + o_p(1). \tag{11}$$

So it will suffice to show that $\frac{1}{T}\sum_{t=1}^{T}\hat{z}_t e_t = o_p(1)$:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\hat{z}_t e_t &= \frac{1}{T}\sum_{t=1}^{T}(z_t + (\hat{z}_t - z_t))e_t \\
&= \frac{1}{T}\sum_{t=1}^{T} z_t e_t + \frac{1}{T}\sum_{t=1}^{T}(\hat{z}_t - z_t)e_t. \qquad (12)
\end{aligned}
$$

The first term in (12) is equal to $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[z_t e_t] + o_p(1)$, by the WLLN of McLeish (1975). The WLLN is justified as for each $j \in \{1, \ldots, J-1\}$ and each t,

$$
\begin{aligned}
\mathbb{E}[|z_{j,t}e_t|^{r+\delta}] &\leqslant (\mathbb{E}[|z_{j,t}^2|^{r+\delta}])^{1/2}(\mathbb{E}[|e_t^2|^{r+\delta}])^{1/2} \\
&\quad \text{(by Hölder's inequality)} \\
&\leqslant \infty
\end{aligned}
$$

(by Assumption 3.2 and Lemma 4).

The second term in (12) is, for each $j \in \{1, \ldots, J-1\}$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(\hat{z}_{j,t} - z_{j,t})e_t &\leqslant \sup_t|\hat{z}_{j,t} - z_{j,t}|\frac{1}{T}\sum_{t=1}^{T}|e_t| \\ &= O_p\left(\sqrt{\frac{\log T}{N}}\right)o_p(1) \\ &\text{(by Proposition 3)},\end{aligned}$$

where $\frac{1}{T}\sum_{t=1}^{T} e_t = o_p(1)$ by the WLLN of McLeish (1975) and justified by Assumption 3.2. Finally as $\sqrt{\log T/N} \to 0$ by assumption, $O_p\left(\sqrt{\frac{\log T}{N}}\right)o_p(1) = o_p(1)$.

It remains to show that $\frac{1}{T}\hat{\boldsymbol{Z}}'\boldsymbol{Y} \xrightarrow{p} \boldsymbol{Q}_{zY}$ and $\frac{1}{T}\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}} \xrightarrow{p} \boldsymbol{Q}_{zz}$. First $\frac{1}{T}\hat{\boldsymbol{Z}}'\boldsymbol{Y} \xrightarrow{p} \boldsymbol{Q}_{zY}$,

$$\begin{aligned}\frac{1}{T}\hat{\boldsymbol{Z}}'\boldsymbol{Y} &= \frac{1}{T}\sum_{t=1}^{T}\hat{z}_t Y_t \\ &= \frac{1}{T}\sum_{t=1}^{T}(z_t + (\hat{z}_t - z_t))Y_t \\ &= \frac{1}{T}\sum_{t=1}^{T}z_t Y_t + \frac{1}{T}\sum_{t=1}^{T}(\hat{z}_t - z_t)Y_t. \qquad (13)\end{aligned}$$

The first term in (13) is equal to $\mathbf{Q}_{zY} + o_p(1)$, by the WLLN of McLeish (1975). The WLLN is justified as for each $j \in \{1, \ldots, J-1\}$ and each t,

$$\begin{aligned}\mathbb{E}[|z_{j,t}Y_t|^{r+\delta}] &\leqslant (\mathbb{E}[|z_{j,t}^2|^{r+\delta}])^{1/2}(\mathbb{E}[|Y_t^2|^{r+\delta}])^{1/2} \\ &\text{(by Hölder's inequality)} \\ &\leqslant \infty \\ &\text{(by Assumption 3.2 and Lemma 4)}.\end{aligned}$$

The second term in (13) is, for each $j \in \{1, \ldots, J-1\}$,

$$\frac{1}{T}\sum_{t=1}^{T}(\hat{z}_{j,t} - z_{j,t})Y_t \leqslant \sup_t|\hat{z}_{j,t} - z_{j,t}|\frac{1}{T}\sum_{t=1}^{T}|Y_t|$$

$$= O_p\left(\sqrt{\frac{\log T}{N}}\right) O_p(1)$$

(by Proposition 3),

where $\frac{1}{T}\sum_{t=1}^{T} Y_t = O_p(1)$ by the WLLN of McLeish (1975) and justified Assumption 3.2. Finally as $\sqrt{\log T/N} \to 0$, $O_p\left(\sqrt{\frac{\log T}{N}}\right) O_p(1) = o_p(1)$.

Finally, $\frac{1}{T}\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}} \xrightarrow{p} \boldsymbol{Q}_{zz}$,

$$\begin{aligned}
\frac{1}{T}\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}} &= \frac{1}{T}\sum_{t=1}^{T} \hat{z}_t \hat{z}_t' \\
&= \frac{1}{T}\sum_{t=1}^{T} (z_t + (\hat{z}_t - z_t))(z_t + (\hat{z}_t - z_t))' \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T} z_t z_t'}_{(1)} + 2\underbrace{\frac{1}{T}\sum_{t=1}^{T} z_t(\hat{z}_t - z_t)'}_{(2)} + \underbrace{\sum_{t=1}^{T} \hat{z}_t \hat{z}_t'}_{(3)}. \qquad (14)
\end{aligned}$$

(1) in (14) is equal to $\mathbf{Q}_{zz} + o_p(1)$, by the WLLN of McLeish (1975) and justified by Assumption 3.2. (2) in (14) is, for each $i, j \in \{1, \ldots, J-1\}$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} (\hat{z}_{i,t} - z_{i,t}) z_{j,t} &\leqslant \sup_t |\hat{z}_{i,t} - z_{i,t}| \frac{1}{T}\sum_{t=1}^{T} |z_{j,t}| \\
&= O_p\left(\sqrt{\frac{\log T}{N}}\right) O_p(1)
\end{aligned}$$

(by Proposition 3),

where $\frac{1}{T}\sum_{t=1}^{T} |z_{j,t}| = O_p(1)$ by the WLLN of McLeish (1975) and justified Assumption 3.2. (3) in (14) is, for each $i, j \in \{1, \ldots, J-1\}$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} (\hat{z}_{i,t} - z_{i,t})(\hat{z}_{j,t} - z_{j,t}) &\leqslant \sup_t |(\hat{z}_{i,t} - z_{i,t})(\hat{z}_{j,t} - z_{j,t})| \\
&\leqslant \sup_t |\hat{z}_{i,t} - z_{i,t}| \sup_t |\hat{z}_{j,t} - z_{j,t}| \\
&= O_p\left(\sqrt{\frac{\log T}{N}}\right) O_p\left(\sqrt{\frac{\log T}{N}}\right)
\end{aligned}$$

$$
\begin{aligned}
&\text{(by Proposition 3)},\\
&= O_p\left(\frac{\log T}{N}\right)\\
&= o_p(1)\\
&\text{(as } \sqrt{\log T/N} \to 0\text{)}.
\end{aligned}
$$

By Assumption 3.2, $\mathbf{Q}_{zz}$ is nonsingular for large T and $\mathbf{Q}_{zz}^{-1}$ is bounded by Assumption 3.3.

Asymptotic normality:

Multiplying (11) by $\sqrt{T}$ yields

$$
\sqrt{T}(\hat{\theta}(\hat{z}_t) - \theta_0) = (\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zY})^{-1}\mathbf{Q}_{Yz}\mathbf{Q}_{zz}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{z}_t e_t + o_p(1).
$$

The term $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{z}_t e_t$ can be written as

$$
\begin{aligned}
\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{z}_t e_t &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(z_t + (\hat{z}_t - z_t))e_t\\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t e_t + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\hat{z}_t - z_t)e_t. \qquad (15)
\end{aligned}
$$

The second term in (15) is, for each $j \in \{1, \ldots, J-1\}$,

$$
\begin{aligned}
\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\hat{z}_{j,t} - z_{j,t})e_t &\leqslant \sup_t|(\hat{z}_{j,t} - z_{j,t})|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}|e_t|\\
&= O_p\left(\sqrt{\frac{\log T}{N}}\right)O_p(1)\\
&\text{(by Proposition 3)},
\end{aligned}
$$

where $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}|e_t| = O_p(1)$ follows from Theorem 2.6 of Domowitz and White (1982), justified by Assumption 3.2. Finally as $\sqrt{\log T/N} \to 0$, $O_p\left(\sqrt{\frac{\log T}{N}}\right)O_p(1) = o_p(1)$. The first term in (15) converges in distribution to $N(0, \mathbf{\Omega})$. Recall that the average

covariance matrix is

$$\mathbf{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[e_t^2 z_t z_t'] + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[e_t e_{t-\tau}(z_t z_{t-\tau}' + z_{t-\tau}z_t')].$$

$\mathbf{\Omega}$ is positive for T large by Assumption 3.5, and hence $\mathbf{\Omega}^{-1/2}$ is well defined. By the CLT of Theorem 2.6 of Domowitz and White (1982), justified by Assumption 3.4, $\frac{1}{\sqrt{T}}\mathbf{\Omega}^{-1/2}\sum_{t=1}^{T}\hat{z}_t e_t \xrightarrow{d} N(0, I)$.

Asymptotic Variance:

Our estimator for the asymptotic variance $\boldsymbol{V}_\theta$ is

$$\hat{V}_\theta(\hat{z}_t) = (\boldsymbol{Y}'P_{\hat{\boldsymbol{Z}}}\boldsymbol{Y})^{-1}\boldsymbol{Y}'P_{\hat{\boldsymbol{Z}}}\hat{\mathbf{\Omega}}(\hat{z}_t)P_{\hat{\boldsymbol{Z}}}\boldsymbol{Y}(\boldsymbol{Y}'P_{\hat{\boldsymbol{Z}}}\boldsymbol{Y})^{-1}$$

We have shown above that $\frac{1}{T}\hat{\boldsymbol{Z}}'\boldsymbol{Y} \xrightarrow{p} \mathbf{Q}_{\boldsymbol{zY}}$ and $\frac{1}{T}\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}} \xrightarrow{p} \boldsymbol{Q}_{zz}$, so it suffices to show that $\hat{\mathbf{\Omega}}(\hat{z}_t) \xrightarrow{p} \mathbf{\Omega}$.

Recall that

$$\hat{\mathbf{\Omega}}(\hat{z}_t) = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2\hat{z}_t\hat{z}_t' + \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}\hat{e}_t\hat{e}_{t-\tau}(\hat{z}_t\hat{z}_{t-\tau}' + \hat{z}_{t-\tau}\hat{z}_t')$$

and

$$\mathbf{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[e_t^2 z_t z_t'] + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[e_t e_{t-\tau}(z_t z_{t-\tau}' + z_{t-\tau}z_t')].$$

Additionally define

$$\mathbf{\Omega}(\theta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\theta)^2 z_t' z_t + \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\theta)(y_{t-\tau} - Y_{t-\tau}\theta)(z_t z_{t-\tau}' + z_{t-\tau}z_t')$$

and

$$\begin{aligned}\tilde{\mathbf{\Omega}}(\theta) = &\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(y_t - Y_t\theta)^2 z_t' z_t] \\ &+ \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}\mathbb{E}[(y_t - Y_t\theta)(y_{t-\tau} - Y_{t-\tau}\theta)(z_t z_{t-\tau}' + z_{t-\tau}z_t')].\end{aligned}$$

The proof will proceed as follows. We want to show that $(\hat{z}_t) \xrightarrow{p} \mathbf{\Omega}$. We will do so by showing

1. $(\hat{z}_t) \xrightarrow{p} \mathbf{\Omega}(\hat{\theta})$,
2. $\mathbf{\Omega}(\theta) \xrightarrow{a.s.} \tilde{\mathbf{\Omega}}(\theta)$ uniformly in θ,
3. $\tilde{\mathbf{\Omega}}(\hat{\theta}) \xrightarrow{p} \tilde{\mathbf{\Omega}}(\theta_0)$,
4. $\tilde{\mathbf{\Omega}}(\theta_0) \xrightarrow{p} \mathbf{\Omega}$.

1. $\hat{\mathbf{\Omega}}(\hat{z}_t) \xrightarrow{p} \mathbf{\Omega}(\hat{\theta})$

$\hat{\mathbf{\Omega}}(\hat{z}_t)$ is comprised on 2 terms,

$$\hat{\mathbf{\Omega}} = \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2\hat{z}_t\hat{z}_t'}_{(1)} + \underbrace{\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(\hat{z}_t\hat{z}_{t-\tau}' + \hat{z}_{t-\tau}\hat{z}_t')}_{(2)},$$

as is $\mathbf{\Omega}(\hat{\theta})$

$$\hat{\mathbf{\Omega}} = \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2 z_t z_t'}_{(a)} \tag{16}$$

$$+ \underbrace{\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(z_t z_{t-\tau}' + z_{t-\tau}\hat{z}_t')}_{(b)}. \tag{17}$$

We can rewrite (1) as

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2\hat{z}_t\hat{z}_t' &= \frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2(z_t + (\hat{z}_t - z_t))(z_t + (\hat{z}_t - z_t))' \\ &= \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2 z_t z_t}_{(1.1)} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2(\hat{z}_t - z_t)z_t'}_{(1.2)}\end{aligned}$$

$$+ \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2 z_t(\hat{z}_t - z_t)'}_{(1.2)}$$

$$+ \underbrace{\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2(\hat{z}_t - z_t)(\hat{z}_t - z_t)'}_{(1.3)}. \tag{18}$$

(1.1) is identical to (a) in (17), so next we will show that (1.2) and (1.3) are $o_p(1)$. Note that $y_t - Y_t\hat{\theta} = e_t + Y_t(\theta_0 - \hat{\theta})$. (1.2), for each i and j is

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2(\hat{z}_{i,t} - z_{i,t})z_{j,t} &\leqslant \sup_t|\hat{z}_{i,t} - z_{i,t}|\frac{1}{T}\sum_{t=1}^{T}|(e_t + Y_t(\theta_0 - \hat{\theta}))^2 z_{j,t}| \\
&= \sup_t|\hat{z}_{i,t} - z_{i,t}|\frac{1}{T}\sum_{t=1}^{T}|e_t^2 z_{j,t} + 2e_t Y_t(\theta_0 - \hat{\theta})z_{j,t} + Y_t^2(\theta_0 - \hat{\theta})^2 z_{j,t}| \\
&\leqslant \underbrace{\sup_t|\hat{z}_{j,t} - z_{j,t}|\frac{1}{T}\sum_{t=1}^{T}|e_t^2 z_{j,t}|}_{(1.2.1)} + \underbrace{2\sup_t|\hat{z}_{i,t} - z_{i,t}||\theta_0 - \hat{\theta}|\frac{1}{T}\sum_{t=1}^{T}|e_t Y_t z_{j,t}|}_{(1.2.2)} \\
&\quad + \underbrace{\sup_t|\hat{z}_{i,t} - z_{i,t}||\theta_0 - \hat{\theta}|^2\frac{1}{T}\sum_{t=1}^{T}|Y_t^2 z_{j,t}|}_{(1.2.3)}
\end{aligned}$$

(1.2.1) is $O_p\left(\sqrt{\frac{\log T}{N}}\right)O_p(1)$. The $O_p(1)$ term is by the WLLN, which is justified by $\mathbb{E}[|e_t^2 z_{j,t}|^{r+\delta}] \leqslant \mathbb{E}[|e_t^3|^{r+\delta}]^{2/3}\mathbb{E}[|z_{j,t}^3|^{r+\delta}]^{1/3}$ for each j, by Hölder's inequality and is finite by Assumption 3.2.

(1.2.2) is $O_p\left(\sqrt{\frac{\log T}{N}}\right)o_p(1)O_p(1)$. The $o_p(1)$ follows from the consistency of $\hat{\theta}$. The $O_p(1)$ is by the WLLN, which is justified by

$$\mathbb{E}[|e_t Y_t z_{j,t}|^{r+\delta}] \leqslant \mathbb{E}[|e_t^3|^{r+\delta}]^{1/3}\mathbb{E}[|Y_t^3|^{r+\delta}]^{1/3}\mathbb{E}[|z_{j,t}^3|^{r+\delta}]^{1/3}$$

for each j by Hölder's inequality and is finite by Assumption 3.2.

(1.2.3) is $O_p\left(\sqrt{\frac{\log T}{N}}\right)o_p(1)^2O_p(1)$. The $O_p(1)$ is by the WLLN, which is justified by $\mathbb{E}[|Y_t^2 z_{j,t}|^{r+\delta}] \leqslant \mathbb{E}[|Y_t^3|^{r+\delta}]^{2/3}\mathbb{E}[|z_{j,t}^3|^{r+\delta}]^{1/3}$ for each j by Hölder's inequality and

is finite by Assumption 3.2.

(1.3) in Equation (18) is for each i and j

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}(y_t - Y_t\hat{\theta})^2(\hat{z}_{i,t} - z_{i,t})(\hat{z}_{j,t} - z_{j,t}) &= \frac{1}{T}\sum_{t=1}^{T}(e_t + Y_t(\theta_0 - \hat{\theta}))^2(\hat{z}_{i,t} - z_{i,t})(\hat{z}_{j,t} - z_{j,t}) \\
&\leqslant \sup_t|\hat{z}_{i,t} - z_{i,t}|\sup_t|\hat{z}_{j,t} - z_{j,t}| \\
&\times \frac{1}{T}\sum_{t=1}^{T}|e_t^2 + 2e_tY_t(\theta_0 - \hat{\theta}) + Y_t^2(\theta_0 - \hat{\theta})^2| \\
&\leqslant \underbrace{\sup_t|\hat{z}_{i,t} - z_{i,t}|\sup_t|\hat{z}_{j,t} - z_{j,t}|\frac{1}{T}\sum_{t=1}^{T}|e_t^2|}_{(1.3.1)} \\
&+ \underbrace{2\sup_t|\hat{z}_{i,t} - z_{i,t}|\sup_t|\hat{z}_{j,t} - z_{j,t}||\theta_0 - \hat{\theta}|\frac{1}{T}\sum_{t=1}^{T}|e_tY_t|}_{(1.3.2)} \\
&+ \underbrace{\sup_t|\hat{z}_{i,t} - z_{i,t}|\sup_t|\hat{z}_{j,t} - z_{j,t}||\theta_0 - \hat{\theta}|^2\frac{1}{T}\sum_{t=1}^{T}|Y_t^2|}_{(1.3.3)}
\end{aligned}$$

(1.3.1) is $O_p\left(\frac{\log T}{N}\right)o_p(1)$. The $O_p(1)$ is by the WLLN, which is justified by Assumption 3.2.

(1.3.2) is $O_p\left(\frac{\log T}{N}\right)o_p(1)O_p(1)$. The $O_p(1)$ is by the WLLN, which is justified by $\mathbb{E}[|e_tY_t|^{r+\delta}] \leqslant \mathbb{E}[|e_t^2|^{r+\delta}]^{1/2}\mathbb{E}[|Y_t^2|^{r+\delta}]^{1/2}$ by Hölder's inequality and is finite by Assumption 3.2.

(1.3.3) is $O_p\left(\frac{\log T}{N}\right)o_p(1)^2O_p(1)$. The $O_p(1)$ is by the WLLN, which is justified by Assumption 3.2.

Now we turn our attention to (2),

$$\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(\hat{z}_t\hat{z}'_{t-\tau} + \hat{z}_{t-\tau}\hat{z}'_{t-\tau}).$$

It suffices to only consider

$$\begin{aligned}
\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})\hat{z}_t\hat{z}'_{t-\tau} &= \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})\hat{z}_t\hat{z}'_{t-\tau} \\
&= \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(z_t + (\hat{z}_t - z_t))(z_{t-\tau} + (\hat{z}_{t-\tau} - z_{t-\tau}))' \\
&= \underbrace{\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})z_t z'_{t-\tau}}_{(2.1)} \\
&+ \underbrace{\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})z_t(\hat{z}_{t-\tau} - z_{t-\tau})'}_{(2.2)} \\
&+ \underbrace{\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(\hat{z}_t - z_t)z'_{t-\tau}}_{(2.2)} \\
&+ \underbrace{\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(\hat{z}_t - z_t)(\hat{z}_{t-\tau} - z_{t-\tau})'}_{(2.3)}.
\end{aligned}$$

(2.1) is identical to (b) in (17), so next we will show that (2.2) and (2.3) are $o_p(1)$. (2.2) is for each i and j

$$\begin{aligned}
&\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t - Y_t\hat{\theta})(y_{t-\tau} - Y_{t-\tau}\hat{\theta})(\hat{z}_{i,t-\tau} - z_{i,t-\tau})z_{j,t-\tau} \\
&= \frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(e_t + Y_t(\theta_0 - \hat{\theta}))(e_{t-\tau} + Y_{t-\tau}(\theta_0 - \hat{\theta}))(\hat{z}_{i,t-\tau} - z_{i,t-\tau})z_{j,t-\tau} \\
&\leqslant \underbrace{\sup_t|\hat{z}_{i,t} - z_{i,t}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|e_t e_{t-\tau}z_{j,t-\tau}|}_{(2.2.1)} \\
&+ \underbrace{\sup_t|\hat{z}_{i,t} - z_{i,t}||\theta_0 - \hat{\theta}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|e_t Y_{t-\tau}z_{j,t-\tau}|}_{(2.2.2)}
\end{aligned}$$

$$+\underbrace{\sup_t|\hat{z}_{i,t}-z_{i,t}||\theta_0-\hat{\theta}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|Y_t e_{t-\tau}z_{j,t-\tau}|}_{(2.2.2)}$$

$$+\underbrace{\sup_t|\hat{z}_{i,t}-z_{i,t}||\theta_0-\hat{\theta}|^2\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|Y_t Y_{t-\tau}z_{j,t-\tau}|}_{(2.2.3)}.$$

Consider first (2.2.1). We want to show that $\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|e_t e_{t-\tau}z_{j,t-\tau}| = O_p(1)$. First note that the expectation of each term is bounded uniformly by Assumption 3.2, and hence bounded by functions that are uniformly $r+\delta$ integrable. Then using Theorem 2.5 of Domowitz and White (1982),

$$|(T-\tau)^{\gamma-1}\sum_{t=1}^{T}|e_t e_{t-\tau}z_{j,t-\tau}| - \mathbb{E}[|e_t e_{t-\tau}z_{j,t-\tau}|]| \overset{a.s.}{\to} 0, \tag{19}$$

for each $\tau = 1,\ldots,l$ and $0<\gamma<\delta/(r+\delta)$. Thus $\frac{1}{T}\sum_{t=\tau+1}^{T}|e_t e_{t-\tau}z_{j,t-\tau}| < T^{-\gamma}\kappa$ for some $\kappa<\infty$, and therefore $\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|e_t e_{t-\tau}z_{j,t-\tau}| \leqslant lT^{-\gamma}\kappa$. Given that $l = O(T^{\gamma})$ for $0<\gamma<\delta/(r+\delta)$, we have our desired result. Thus (2.2.1) is $O_p\left(\sqrt{\frac{\log T}{N}}\right)O_p(1)$ by Proposition 3 and is $o_p(1)$ as $\sqrt{\frac{\log T}{N}}\to 0$ by assumption. Similar results hold for (2.2.2) and (2.2.3), noting that the expectations of these terms are bounded by Assumption 3.2 and hence bounded by functions that are uniformly $r+\delta$ integrable.

(2.3) is, for each i and j

$$\begin{aligned}
&\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(y_t-Y_t\hat{\theta})(y_{t-\tau}-Y_{t-\tau}\hat{\theta})(\hat{z}_{i,t}-z_{i,t})(\hat{z}_{j,t-\tau}-z_{j,t-\tau})\\
&\quad=\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}(e_t+Y_t(\theta_0-\hat{\theta}))(e_{t-\tau}+Y_{t-\tau}(\theta_0-\hat{\theta}))(\hat{z}_{i,t}-z_{i,t})(\hat{z}_{j,t-\tau}-z_{j,t})\\
&\quad\leqslant\underbrace{\sup_t|\hat{z}_{i,t}-z_{i,t}|\sup_t|\hat{z}_{j,t}-z_{j,t}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|e_t e_{t-\tau}|}_{(2.3.1)}
\end{aligned}$$

$$+\underbrace{\sup_t|\hat{z}_{i,t}-z_{i,t}|\sup_t|\hat{z}_{j,t}-z_{j,t}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|e_tY_{t-\tau}|}_{(2.3.2)}$$

$$+\underbrace{\sup_t|\hat{z}_{i,t}-z_{i,t}|\sup_t|\hat{z}_{j,t}-z_{j,t}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|Y_te_{t-\tau}|}_{(2.3.2)}$$

$$+\underbrace{\sup_t|\hat{z}_{i,t}-z_{i,t}|\sup_t|\hat{z}_{j,t}-z_{j,t}|\frac{1}{T}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T}|Y_tY_{t-\tau}|}_{(2.3.3)}.$$

Using identical arguments to those above, all these terms are $o_p(1)$ and hence (2.3) is $o_p(1)$ as required.

2. $\boldsymbol{\Omega}(\theta) \xrightarrow{a.s.} \tilde{\boldsymbol{\Omega}}(\theta)$ **uniformly in** θ

Let $d_{t,\tau} = e_te_{t-\tau}z_{i,t}z_{j,t-\tau} - \mathbb{E}[e_te_{t-\tau}z_{i,t}z_{j,t-\tau}]$, noting that we have suppressed the dependence of i, j and θ. We need to show that

$$|\frac{1}{T}\sum_{t=1}^{T}d_{t,0}| \xrightarrow{a.s.} 0, \tag{20}$$

$$\sum_{\tau=1}^{l}|\frac{1}{T}\sum_{t=1}^{T}d_{t,\tau}| \xrightarrow{a.s.} 0, \tag{21}$$

uniformly in θ, for all τ, where $l = O(T^\gamma)$, $0 < \gamma < \delta/(r+\delta)$. To show (21), note that

$$\sup_{\theta\in\Theta}\sum_{\tau=1}^{l}|\frac{1}{T}\sum_{t=1}^{T}d_{t,\tau}| \leqslant \sum_{\tau=1}^{l}\sup_{\theta\in\Theta}|\frac{1}{T}\sum_{t=1}^{T}d_{t,\tau}|.$$

The $d_{t,\tau}$ are continuous uniformly in t and τ and dominated by $r+\delta$ integrable functions for all τ. Then by Theorem 2.5 of Domowitz and White (1982)

$$\sup_{\theta\in\Theta}\sum_{\tau=1}^{l}|\frac{1}{T}\sum_{t=1}^{T}d_{t,\tau}| < lT^{-\gamma}\kappa$$

for some $\kappa < \infty$, for large T and almost every sequence $\{z_t, e_t\}$. The sets F_l of sequences $\{z_t, e_t\}$ such that $\sup_{\theta\in\Theta}|T^{\gamma-1}\sum_{t=1}^{T} d_{t,\tau}|$ do not converge to 0 for $\tau = 1, \ldots, l$ given $l = O(T^{\gamma})$, $0 < \gamma < \delta/(r+\delta)$, constitute an increasing sequence of sets of measure zero, such that $P(\bigcup_{l=1}^{\infty} F_l) = 0$. Equation (20) holds using identical arguments to those used for (19).

3. $\tilde{\mathbf{\Omega}}(\hat{\theta}) \xrightarrow{p} \tilde{\mathbf{\Omega}}(\theta_0)$

From the mean value theorem for $i, j = 1, \ldots, p$

$$|\tilde{\mathbf{\Omega}}_{i,j}(\hat{\theta}) - \tilde{\mathbf{\Omega}}_{i,j}(\theta_0)| \leqslant |\tilde{\mathbf{\Omega}}_{i,j}(\bar{\theta})(\hat{\theta} - \theta_0)| \leqslant \sum_{i=1}^{p} \left| \frac{\partial \tilde{\mathbf{\Omega}}_{i,j}(\bar{\theta})}{\partial \theta_i} \right| |\hat{\theta}_i - \theta_{i,0}|, \tag{22}$$

where $\bar{\theta}$ lies in the segment between $\hat{\theta}$ and θ_0 and the last inequality follows from the triangle inequality. By Assumption 3.2, for all i

$$\left| \frac{\partial \tilde{\mathbf{\Omega}}_{i,j}(\bar{\theta})}{\partial \theta_i} \right| \leqslant T^{\gamma}\Delta + \Delta,$$

and hence (22) is bounded above by

$$\Delta T^{\gamma-1/2} \sum_{i=1}^{p} |T^{1/2}(\hat{\theta}_i - \theta_{i,0})| + \Delta \sum_{i=1}^{p} |\hat{\theta}_i - \theta_{i,0}|. \tag{23}$$

$\hat{\theta}_i \xrightarrow{a.s.} \theta_{i,0}$ by our consistency result and $T^{1/2}(\hat{\theta}_i - \theta_{i,0}) = O_p(1)$ by our asymptotic normality result. As $\gamma < 1/2$, the right hand side of (23) is $o_p(1)$.

4. $\tilde{\mathbf{\Omega}}(\theta_0) \to \mathbf{\Omega}$

$$\tilde{\mathbf{\Omega}}(\theta_0) - \mathbf{\Omega} = \frac{1}{T} \sum_{\tau=l+1}^{T-1} \sum_{t=\tau+1}^{T} \mathbb{E}[e_t e_{t-\tau}(z_t z'_{t-\tau} + z_{t-\tau} z'_t)].$$

We only need to show

$$\frac{1}{T}\sum_{\tau=l+1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[e_t e_{t-\tau} z_t z'_{t-\tau}] \to 0.$$

As $\mathbb{E}[|z_t e_t|^{2+2\eta}] < \infty$ by Assumption 3.6, where $\eta = r + \delta - 1$, by Lemma 2.2 of White and Domowitz (1984)

$$\mathbb{E}[e_t e_{t-\tau}(z_t z'_{t-\tau} + z_{t-\tau} z'_t)] \leqslant c\alpha(\tau)^{\eta/(2+2\eta)},$$

where c is a finite constant. Therefore

$$\frac{1}{T}\sum_{\tau=l+1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[e_t e_{t-\tau} z_t z'_{t-\tau}] \leqslant c\sum_{\tau=l+1}^{T-1}\alpha(\tau)^{\eta/(2+2\eta)}. \tag{24}$$

(24) goes to zero as

$$\lim_{T\to\infty}\sum_{\tau=l+1}^{T-1}\alpha(\tau)^{\eta/(2+2\eta)} = \lim_{T\to\infty}\sum_{\tau=0}^{T-1}\alpha(\tau)^{\eta/(2+2\eta)} - \lim_{T\to\infty}\sum_{\tau=0}^{l}\alpha(\tau)^{\eta/(2+2\eta)}$$

and $\sum_{\tau=0}^{\infty}\alpha(\tau)^{\eta/(2+2\eta)} < \infty$ by Assumption 3.1.

□

C Additional Figures and Tables

Word	Score	Word	Score	Word	Score
undervalue	0.596	beat	0.527	unanimously	0.519
repurchase	0.573	tender	0.526	buoy	0.518
surpass	0.554	top	0.525	bake	0.518
upgrade	0.551	visible	0.524	get	0.518
rally	0.548	soar	0.524	fragment	0.518
surge	0.547	horizon	0.523	activist	0.518
treasury	0.543	tanker	0.523	cardiology	0.518
customary	0.539	deepwater	0.522	oversold	0.517
imbalance	0.538	reconnaissance	0.522	bidder	0.517
jump	0.538	tag	0.521	cheer	0.517
declare	0.535	deter	0.521	exceed	0.517
unsolicited	0.535	valve	0.519	terrain	0.517
up	0.534	foray	0.519	terrific	0.516
discretion	0.531	clip	0.519	upbeat	0.516
buy	0.531	fastener	0.519	gratify	0.515
climb	0.528	bracket	0.519	armor	0.515
bullish	0.527	potent	0.519		

Table 4: Top 50 highest positive sentiment terms and their weights, from KKX Table A2.

Word	Score	Word	Score	Word	Score
shortfall	0.323	fall	0.446	unfavorable	0.462
downgrade	0.382	covenant	0.451	regain	0.462
disappointing	0.392	woe	0.452	deficit	0.462
tumble	0.402	slash	0.453	irregularity	0.463
blame	0.414	resign	0.454	erosion	0.464
hurt	0.414	delay	0.454	bondholder	0.464
plummet	0.423	subpoena	0.454	weak	0.465
auditor	0.424	lackluster	0.455	hamper	0.465
plunge	0.429	soften	0.456	overrun	0.467
waiver	0.429	default	0.46	inefficiency	0.467
miss	0.43	soft	0.46	persistent	0.468
slowdown	0.433	widen	0.46	notify	0.468
halt	0.435	postpone	0.46	allotment	0.469
sluggish	0.439	unfortunately	0.46	worse	0.469
lower	0.441	insufficient	0.462	setback	0.471
downward	0.443	unlawful	0.462	grace	0.472
warn	0.444	issuable	0.462		

Table 5: Top 50 negative sentiment-charged terms and their weights, from KKX Table A2.

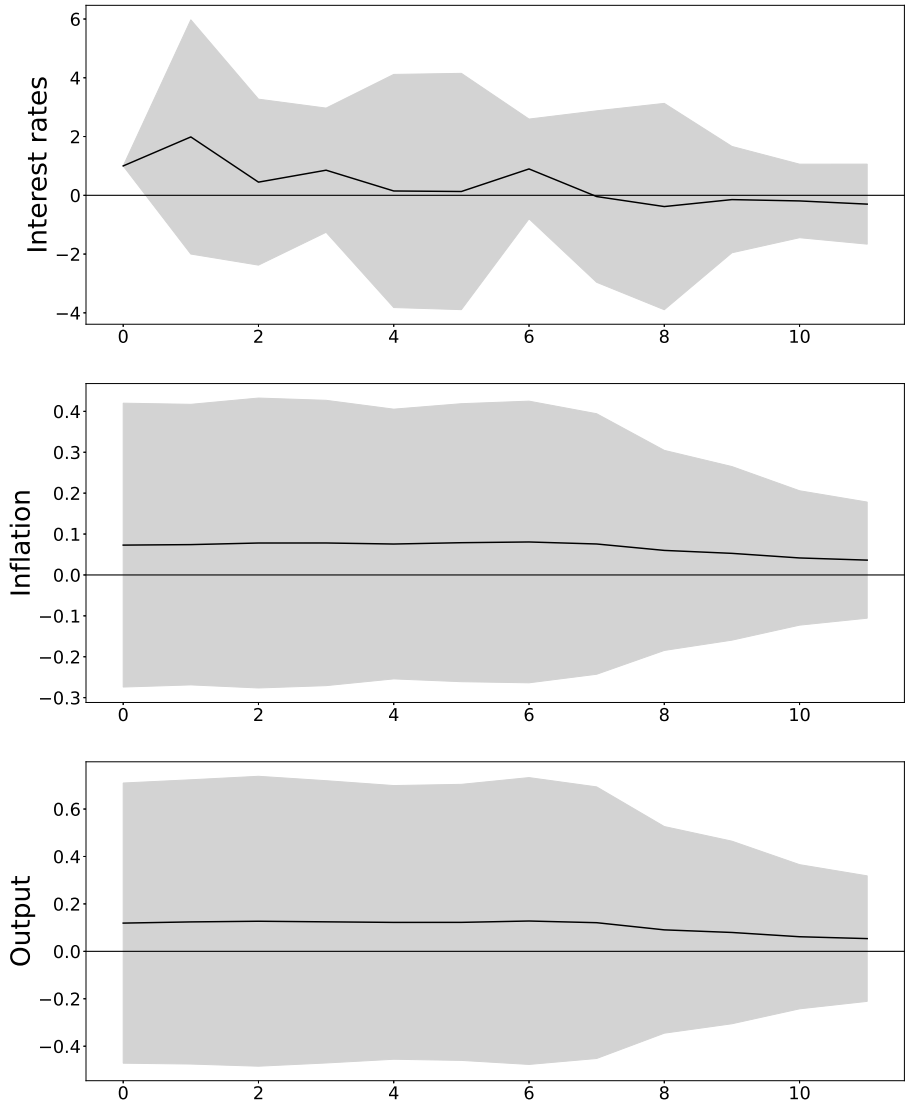

Figure 3: Impulse responses of interest rates (top), inflation (middle) and output (bottom) to a one unit contractionary monetary policy shock identified using the Romer and Romer 2004 monetary policy shock series. Horizon is quarters. Shaded area represents the 95% confidence interval.

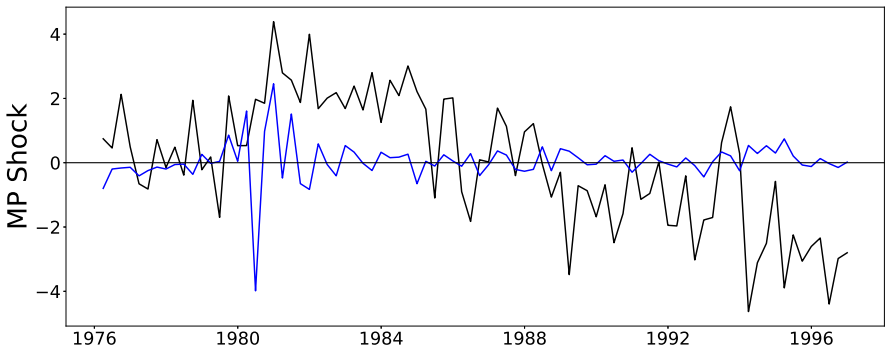

Figure 4: Text instruments (black) and Romer and Romer (2004) monetary policy shocks (blue). Monetary policy shock using text instruments is estimated using the fitted values for the first stage estimated on the full sample, demeaned.

D Word Mover's Distance and Cosine Similarity

The cosine similarity is computed by first computing the $D \times V$ document term frequency matrix $\mathbf{F}$ of the appropriately preprocessed definitions of each term. Entry $F_{d,v}$ is the frequency of term v in document d. Denoting each row of the document term matrix as a 'document vector', the cosine similarity between document i and j is simply the inner product of the document vectors for i and j.

WMD relies on word embeddings, a method to transform each term in a vocabulary to an element in a low dimensional vector space (much lower dimensional than the size of the vocabulary V), which we will denote $\boldsymbol{v}_i$. The word embeddings that we use are the *word2vec* (Mikolov et al. 2013a; Mikolov et al. 2013b) Google News corpus pre-trained vectors.[27] Other popular choices for pre-trained word vectors are GloVe (Pennington, Socher, and Manning 2014), fastText (Bojanowski et al. 2017) and BERT (Devlin et al. 2019).

WMD is computed as the cost to transform a document vector $\boldsymbol{d}$ into a document vector $\boldsymbol{d}'$. The 'travelling cost' between terms i and j is denoted $c(\boldsymbol{v}_i, \boldsymbol{v}_j) = \|\boldsymbol{v}_i - \boldsymbol{v}_j\|^2$. Let $T_{ij} \in \mathbb{R}^{V \times V}$ be a sparse flow matrix where $T_{ij} \geqslant 0$ denotes how much of word i in $\boldsymbol{d}$ moves to word j in $\boldsymbol{d}'$.

To transform document $\boldsymbol{d}$ entirely into document $\boldsymbol{d}'$, we ensure that the entire outgoing flow for each term i equals d_i, that is $\sum_j T_{ij} = d_i$, and the entire incoming flow to term j equals d'_j, $\sum_i T_{ij} = d'_j$. The WMD between two documents d and d' is the solution to the following optimal transport problem

$$\min_{\boldsymbol{T} \geqslant 0} \sum_{i,j=1}^{V} T_{ij}\, c(i,j),$$
$$\text{s.t.} \sum_{j=1}^{V} T_{ij} = d_i, \quad \forall i \in \{1, \cdots, V\}, \quad \sum_{i=1}^{V} T_{ij} = d'_j, \quad \forall j \in \{1, \cdots, V\}.$$

27. These 300-dimensional word vectors are trained on 100 billion words, and cover a vocabulary of around 3 million words. They can be obtained from the word2vec website.

D.1 Synonyms: Word Mover's Distance

50 top synonyms for '*monetary policy*', '*inflation*' and '*output*', computed using WMD. Terms are ranked in ascending WMD from target term, with the target term listed first.

Monetary Policy

monetary policy, international monetary system, open market operations, talk down, macroeconomic policy, monetary system, zero interest rate, floating exchange rate, central bank, easy monetary policy

Inflation

inflation, imported inflation, non-accelerating inflation rate of unemployment, inflationary gap, wage-price spiral, creeping inflation, anticipated inflation, Phillips curve, inflationary spiral, price stability, real balances, price level, real interest rate, sacrifice ratio, real balance effect, underlying rate of inflation, price index, strengthening of a currency, capital gain, kinked demand curve, escalator clause, natural rate of interest, real GDP, real GNP, relative price, nominal anchor, current prices, overfull employment, fiscal drag, monetarism, price volatility, over-stimulation, real wages, capital appreciation, real exchange rate, natural rate of unemployment, sound money, real income, real terms, oil price, devaluation, structural break, store of value, cuts in expenditure, under-valued currency, cost(s), internal balance, soft landing, revaluation, depression

Output

output, economic profit, capacity, inputs, potential output, process innovation, efficiency, inefficiency, innovation, knowledge economy, ceiling, factor cost, infrastructure, free enterprise, outsourcing, constant prices, capital goods, initial conditions, mark-to-market, demand, economic indicators, mass production, development, double counting, producer good, invention, supply, real costs, corporate sector, economic efficiency, production function, fiscal stimulus, turnover, total factor productivity, redeployment, economic planning, economic development, GNP, parameter, technology, floor, simulation, aggregate demand, enterprise, service industry, prior, foreign trade, conspicuous consumption, break-even, eurodollars

E Smoothing Multinomial Frequencies

One practical issue the generalized log odds estimator faces is that it is undefined when the sample frequency of a term is 0, due to the log transformation. This occurs only in finite samples as by definition of $p_t(w, c) \in (L, 1)$, (this L is used in the construction of Proposition 3).

A common way to address this issue is by smoothing the multinomial frequencies.[28] We will use additive or Laplace smoothing, which can be motivated as the posterior mean of the multinomial frequencies with a Dirichlet prior.

Define $x_{j,t} = \frac{1}{N_t} \sum_{w=1}^{N_t} \mathbf{1}_{w_{t,n}=w_j}$, for $j = 1, \ldots, V$. Then

$$(x_{1,t}, \ldots, x_{V,t}) \sim \text{Multi}(p_t(w_1), \ldots, p_t(w_V)).$$

Assume that $p_t(w_j)$ follows a Dirichlet prior with parameter $\eta_{j,t}$, that is

$$(p_t(w_1), \ldots, p_t(w_V)) \sim \text{Dir}(\eta_{1,t}, \ldots, \eta_{V,t}).$$

The Dirichlet distribution is a conjugate prior for the multinomial distribution, and hence the posterior has closed form

$$(p_t(w_1), \ldots, p_t(w_V))|(x_{1,t}, \ldots, x_{V,t}) \sim \text{Dir}(\eta_{1,t} + x_{1,t}, \ldots, \eta_{V,t} + x_{V,t}).$$

We will use the posterior mean as our estimator of $p_t(w_j, c)$

$$\begin{aligned} \hat{p}_t(w) &= \frac{\eta_{j,t} + x_{j,t}}{\sum_{j=1}^{V} \eta_{j,t} + x_{j,t}} \\ &= \frac{\eta_{j,t} + x_{j,t}}{N_t + \sum_{j=1}^{V} \eta_{j,t}}, \end{aligned}$$

where the second equality follows from $\sum_{j=1}^{V} x_{j,t} = N_t$. If one were to assume a uniform Dirichlet prior, which sets $\eta_{j,t} = 1$, for $j = 1, \ldots, V$, our estimator becomes

$$\hat{p}_t(w_j, c) = \frac{1 + x_{j,t}}{N_t + V}.$$

28. For a comprehensive list of smoothing techniques in Natural Language Processing see Chen and Goodman (1999).

In our applications, as well as our Monte Carlo simulations, we set $\eta_{j,t} = 0.00001$ for all j and t.

F Generalized log odds

Proposition 4. For any $\mathcal{J}$, context c, weights $\boldsymbol{\omega}$, for each t as $N \to \infty$:

1. $\hat{z}_t(\mathcal{J}, \boldsymbol{\omega})$ is consistent,

$$\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) \xrightarrow{p} z_t(\mathcal{J}, \boldsymbol{\omega}).$$

2. $\hat{z}_t(\mathcal{J}, \boldsymbol{\omega})$ is asymptotically normal

$$\sqrt{N}(\hat{z}_t(\mathcal{J}, \boldsymbol{\omega}) - z_t(\mathcal{J}, \boldsymbol{\omega})) \xrightarrow{d} N(0, G_t),$$

where the asymptotic variance is

$$G_t = \sum_{j \in \mathcal{J}} \frac{\omega_j^2}{p_t(w_j)} - \sum_{i,j \in \mathcal{J}} \omega_j \omega_i.$$

Note that the variance of the generalized log odds estimator is $G_t = \sum_{j \in \mathcal{J}} \frac{\omega_j^2}{p_t(w_j)}$ if $\sum_{j \in \mathcal{J}} \omega_j = 0$, which we will impose when we use this estimator as an instrument.

Proof of Proposition 4

Proof. **Consistency:**

As the $w_{t,n}$ are i.i.d., we can apply the Law of Large Numbers for each $w_j \in \mathcal{J}$ and for each $t = 1, \ldots, T$,

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{w_{t,n}=w_j} &\xrightarrow{p} \mathbb{E}[\mathbf{1}_{w_{t,n}=w_j}] \\ &= p_t(w_{t,n} = w_j). \end{aligned}$$

The result then follows from the Continuous Mapping Theorem, noting $\sum_{j \in \mathcal{J}} \omega_j \log(x_j)$ is continuous in each x_j.

Asymptotic Normality: For notational convenience denote $\hat{p}_t(w_j) := \frac{1}{N_t} \sum_{w=1}^{N_t} \mathbf{1}_{w_{t,n}=w_j}$.

Denote $\mathbf{p}_t = [p_t(w_1), \ldots, p_t(w_V)]'$, the vector of conditional probabilities, and $\hat{\mathbf{p}}_t = [\hat{p}_t(w_1), \ldots, \hat{p}_t(w_V)]'$, the vector of estimates of the conditional probabilities.

We will show asymptotic normality using the delta method.

Using the Central Limit Theorem we have, $\sqrt{N}(\hat{\mathbf{p}}_t - \mathbf{p}_t) \xrightarrow{d} N(0, M_t)$, where $M_t = P_t - \mathbf{p}_t\mathbf{p}_t'$, and P_t is a diagonal matrix with entries $\mathbf{p}_t$.

The generalized log odds is a function of $\mathbf{p}_t$, with derivative $\nabla(\mathbf{p}_t)_j = \omega_j / p_t(w_j)$ for $j \in \mathcal{J}$ and 0 otherwise. By the delta method,

$$\sqrt{N}(z_t(\hat{\mathbf{p}}_t) - z_t(\mathbf{p}_t) \xrightarrow{d} N(0, G_t),$$

where

$$G_t = \nabla\mathbf{p}_t{}' M_t \nabla\mathbf{p}_t = \sum_{j\in\mathcal{J}} \frac{\omega_j^2}{p_t(w_j)} - \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{J}} \omega_j\omega_i.$$

□

G Conditional Probabilities

The results of Proposition 2 demonstrate that under the weak conditions of Assumption 1, we can find valid instruments using the log odds ratio of two special words in a shared context. In this section, we show that under those same assumptions (adapted for a single word), conditional probabilities are not exogenous, and hence are not instruments. We later show that under a stronger set of assumptions we can use conditional probabilities as instruments.

Consider the conditional probability of w^* under the assumption $\alpha_{1,w^*} \neq 0$ and $\alpha_{2,w^*} = 0$. It is

$$p_t(w^*) = \frac{\exp\{x_t'\gamma + \alpha_{1,w^*} z_t^* + \xi_{w^*,t}\}}{\sum_{w'=1}^{V} \exp\{x_t'\gamma + \alpha_{1,w'} z_t^* + \alpha_{2,w'} e_t + \xi_{w',t}\}}. \tag{25}$$

This conditional probability are a not valid instrument, because it is not exogenous (although it does satisfy relevance), due to the presence of e_t in the denominator of (25). This is not merely as artifact of the logit form of conditional probability, this represents a tradeoff for speakers (see Section 3).[29]

Proposition 5. Suppose w^* satisfies $\alpha_{1,w^*} \neq 0$ and $\alpha_{2,w^*} = 0$. Then $p_t(w^*)$ is an not instrument for Y_t.

If we place stronger assumptions on all terms in the vocabulary, we will be able to use conditional probabilities as instruments.

Assumption 4. There exists a term w^* such that

1. $|\alpha_{1,w^*}| > |\alpha_{1,w'}|$, for all $w' \neq w^*$.
2. $\alpha_{2,w} = 0$ for all w.
3. $\mathbb{E}[Y_t|e_t] = 0$.

Assumption 4.1 states that the likelihood of w^* is shifted by z_t^* more than any other word. This is sufficient to show $p_t(w^*)$ is strictly increasing in z_t^*, which is used to show relevance. The need for this stronger condition (relative to Assumption 1.1)

29. The endogeneity problem of conditional probabilities depends on the extent to which the denominator in (25) varies with e_t. Under certain conditions, the denominator (known as the *log partition function*) can be well approximated by a constant (known as *self-normalization*, see Andreas and Klein (2015)).

is that $p_t(w^*)$ is nonlinear in z_t^*. Assumption 4.2 states that the likelihood of all words do not depend on the confounding unobservables e_t. This is what will allow us to remove the effect e_t of the denominator term of $p_t(w^*)$. Assumption 4.3 is a strengthening of the uncorrelated assumption $\mathbb{E}[Y_t e_t] = 0$. This is required to show exogeneity as $p_t(w^*)$ is a nonlinear function of e_t.

Proposition 6. Let Assumption 4 hold. Then $p_t(w^*)$ is an instrument for Y_t.

G.1 Proofs in Appendix G

Conditional probabilities are nonlinear transformations of the data $(x_t, z_t^*, e_t, \xi_{w,t}))$. In order to show the covariance between the conditional probabilities and z_t^* and e_t, under general distributions, we will need the following lemma.

Lemma 5. For a random variable X, not constant almost surely, if $g(x)$ is strictly decreasing, and $\text{Cov}(g(x), x) < \infty$, then $\text{Cov}(g(x), x) < 0$. If $g(x)$ is strictly increasing, and $\text{Cov}(g(x), x) < \infty$, then $\text{Cov}(g(x), x) > 0$.

Proof.

$$\begin{aligned}\text{Cov}(g(x), x) &= \mathbb{E}[Xg(x)] - \mathbb{E}[X]\mathbb{E}[g(x)] \\ &= \mathbb{E}[(X - \mathbb{E}[X])g(X)]\end{aligned}$$

Note that $\mathbb{E}[X - E[X]g(\mathbb{E}[x])] = \mathbb{E}[X - E[X]]g(\mathbb{E}[X]) = 0$, and hence

$$= \mathbb{E}[(X - \mathbb{E}[X])(g(X) - g(\mathbb{E}[X])].$$

As g is strictly decreasing $(X - \mathbb{E}[X])(g(X) - g(\mathbb{E}[X]) \leqslant 0$, with equality when $X = \mathbb{E}[X]$. As X is not constant almost surely, $\text{Cov}(g(x), x) < 0$.

The proof for when $g(x)$ is strictly increasing is identical. □

Proof of Proposition 5

Proof. It will suffice to show that for certain parameter values that we fail to satisfy exogeneity. Suppose $|\alpha_{2,w*}| > |\alpha_{2,w'}|$ for all $w' \neq w^*$. The conditional probability is

$$p_t(w^*) = \frac{\exp(x_t'\gamma + \alpha_{1,w*} z_t^* + \xi_{w*,t})}{\sum_{w'=1}^{V} \exp(x_t'\gamma + \alpha_{1,w'} z_t^* + \alpha_{2,w'} e_t + \xi_{w',t})}.$$

If we take the derivative w.r.t. e_t

$$\frac{\partial p_t(w^*)}{\partial e_t} = \frac{\exp(x_t'\gamma + \alpha_{1,w*} z_t^* + \xi_{w*,t})}{(\sum_{w'=1}^{V} \exp(x_t'\gamma + \alpha_{1,w'} z_t^* + \alpha_{2,w'} e_t + \xi_{w',t}))^2}$$
$$\times \sum_{w'=1}^{V} (\alpha_{2,w*} - \alpha_{2,w'}) \exp(x_t'\gamma + \alpha_{1,w'} z_t^* + \alpha_{2,w'} e_t + \xi_{w',t}).$$

As $|\alpha_{2,w*}| > |\alpha_{2,w}|$, for all $w \neq w^*$, this derivative is either strictly positive ($\alpha_{2,w*} > 0$) or strictly negative (if $\alpha_{2,w*} < 0$). Assume w.l.o.g. that $\alpha_{2,w*} > 0$ and hence $p_t(w^*)$ is strictly increasing in e_t. Therefore

$$\mathbb{E}[p_t(w^*) e_t] > 0,$$

by Lemma 5. $p_t(w^*)$ is not exogenous. □

Proof of Proposition 6

Proof. Under Assumption 4, we have

$$p_t(w^*) = \frac{\exp(x_t'\gamma + \alpha_{1,w*} z_t^* + \xi_{w*,t})}{\sum_{w'=1}^{V} \exp(x_t'\gamma + \alpha_{1,w'} z_t^* + \xi_{w',t})}.$$

Relevance:

If we take the derivative of $p_t(w^*)$ w.r.t. z_t^*

$$\frac{\partial p_t(w^*)}{\partial z_t^*} = \frac{\exp(x_t'\gamma + \alpha_{1,w*} z_t^* + \xi_{w*,t})}{(\sum_{w'=1}^{V} \exp(x_t'\gamma + \alpha_{1,w'} z_t^* + \xi_{w',t}))^2}$$
$$\times \sum_{w'=1}^{V} (\alpha_{1,w*} - \alpha_{1,w'}) \exp(x_t'\gamma + \alpha_{1,w'} z_t^* + \xi_{w',t}).$$

By Assumption 4.2, $|\alpha_{1,w^*}| > |\alpha_{1,w}|$, for all $w \neq w^*$, and hence this derivative is either strictly positive ($\alpha_{1,w^*} > 0$) or strictly negative (if $\alpha_{1,w^*} < 0$). Assume w.l.o.g. that $\alpha_{1,w^*} > 0$ and hence $p_t(w^*)$ is strictly increasing in z_t^*. Therefore

$$\mathbb{E}[p_t(w^*)Y_t] = \mathbb{E}[p_t(w^*)(\delta z_t^* + \nu_t)] > 0,$$

by Lemma 5, and where the last line follows from the reduced form $Y_t = \delta z_t^* + \nu_t$. Hence $p_t(w^*)$ is relevant.

Exogeneity:

$$\mathbb{E}[p_t(w^*)e_t] = \mathbb{E}[\mathbb{E}[p_t(w^*)e_t|z_t^*] = \mathbb{E}[p_t(w^*)\mathbb{E}[e_t|z_t^*]] = 0,$$

where the last equality follows from the assumption $\mathbb{E}[e_t|z_t^*] = 0$. □