ECON_1 TASK

ANDREW BLOHM

1. Abstract

In this paper we estimate improved parameter estimates for determining the market share of individual electricity production technologies using a multinomial logit regression model. The contributions of this paper are two fold: first, the data cleaning process required for the subsequent analysis is significant and not likely to have been accomplished before. We make this data available online at `https://github.com/andymd26/fuzzy-waffle`. Second, GCAM is an integrated assessment model, which presently uses historic data to parameterize the electricity production technology market shares, which are then nonvarying. This work incorporates important determinants in the relationship between preferences, market factors, and market share into the GCAM framework and thus strengthens its output.

2. Theory: Discrete choice models

An agent (n) faces a choice of technology alternatives (j). We assume that each agent is a utility maximizer (U_{nj}) in the selection of a technology alternative. However, U_{nj} is unobservable and consequently unknown to the researcher. Instead, we can observe information about the available choice set (x_{nj}), as well as the agent (s_n). Using this information we can specify a representative utility function, which is $V_{nj} = V(x_{nj}, s_n)$.[1]

We can write each agents utility function as a function of observed and unobserved utility, $U_{nj} = V_{nj} + \epsilon_{nj}$, where the distribution of the error term (ϵ_{nj}) depends on the researchers specification of the observed utility (V_{nj}). Given that we cannot observe ϵ_{nj} directly we instead derive the probability of a technology choice. The probability that the agent chooses technology j (i.e., the expected value of the indicator function) is $Pr(j|x) = Pr(I[h(x,\epsilon) = j] = 1) = \int I[h(x,\epsilon) = j] \cdot f(\epsilon) d\epsilon$.[2]

The probability of an individual choosing the alternative that they actually chose is: $\prod_i (P_{ni})^{y_{ni}}$, where y_{ni} is either one for the selected alternative or zero for the remaining unselected options. Given that the exponents are either zero or one the result reduces to the term P_{ni}, which is just the probability of the chosen alternative. The likelihood function then is $\mathcal{L}(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}}$, which can be rewritten as $ln(\mathcal{L}(\beta)) = \sum_n \sum_i y_{ni} ln\,(P_{ni})$, where the specification of P_{ni} depends on the specification of the model.

Date: September 19, 2016.

[1] $U_{nj} \neq U_{nj}$ given the unobservable aspects of the utility function.

[2] The unobserved parameters ϵ_{nj} follow a probability density distribution $f(\epsilon)$, which in the logistic framework we assume to be distributed extreme value with variance of $\frac{\pi^2}{6}$.

3. Data

[Discussion of EIA Form 860 and AEO supplemental tables]. The dependent variable of our analysis is the estimated annual capacity addition by prime mover and primary fuel type weighted by its capacity factor. We generate the variable using information from the EIA Form 860, which is a survey of electric utilities and contains detailed information down to the generator level. The variable is calculated as the difference in summer capacity across all unique pairings of primary fuel and prime mover technology. One issue that remains in the data, having accounted for retirements and primary fuel changes, is the continued existence of negative values (a negative value in this situation would imply a negative capacity expansion, which should be impossible given the removal of retired and fuel switching plants). A survey of the data appears to show that this could be an issue of plant derating.[3]

As a result of these issues, we instead use the in service date for each unit. Only a small number (50) of the approximately 375,000 entries in the data are missing the in service date. One potential issue with using this date, as opposed to backing out the capacity addition in the process above is that it will not pick up repowering or derating in the data. However, given the purpose of this exercise this should not be a problem.

We weight the dependent variable by its capacity factor (EIA 2016; EIA 2014).

Fixed operations and maintenance costs are expenses that don't vary with production such as routine preventive and predictive maintenance, general and administrative expenses, fees required to participate in NERC and other regulatory bodies, etc. (EIA 2010). Variable operation and maintenance costs are expenses that by definition vary with production such as, water, disposal expenses, power purchases, consumable materials and supplies, etc. (EIA 2010).

We assume the predominant operation cost for a fossil fuel generator is the fuel cost. For renewable sources we assume .... Fuel price data for the period (1/1986-5/2016) is from the Energy Information Administration (EIA) sources. We use the investment cost information (i.e., overnight cost) by electricity production technology from the Global Change Assessment Model (GCAM) to inform our model.

We assume that fuel costs are the predominant operating cost for fossil fuel generation plants. The Energy Information Administration (EIA) in the Electric Power Monthly report publishes electric utility receipts of and average cost for fossil fuels used in the power generation sector. We generate a monthly time-series dataset for the period 1/1986-5/2016 using the Electric Power Monthly reports from June 1996 (Table 26), January 2010 (Table 4.2), and May 2016 (Table 4.2). The data includes price information for coal, petroleum liquids, and natural gas inputs. The prices listed are averages and in nominal units (i.e., unadjusted for inflation). For continuity purposes we use the total petroleum price from the June 1996 report instead of the heavy oil price.

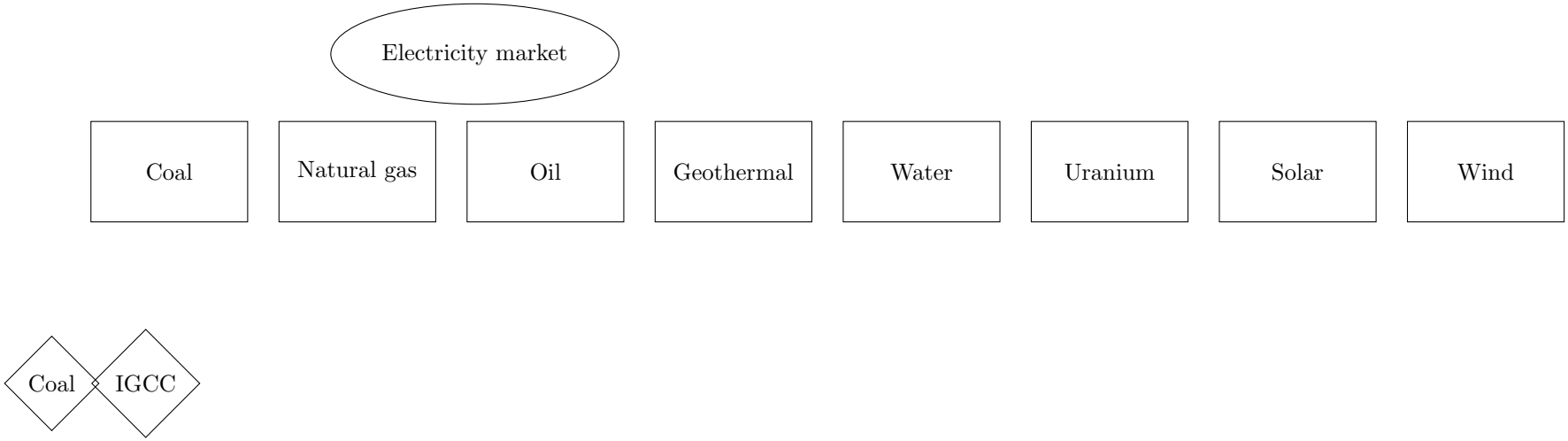
The decision faced by decision-makers in our analysis is the choice amongst power generation input fuels (not yet focusing on the technology choice). We use the monthly installed net summer capacity by fuel type as the dependent variable of our analysis. Existing nameplate and net summer capacity by energy source, producer type, and state is provided by the EIA. Net summer capacity represents the maximum capacity as determined by testing during the summer peak demand period

[3]For example, Utility 13781 operates Plant 3982, Generator 4, which has an installed summer capacity of 22 MW and a steam turbine to convert biomass to electricity. The unit exists in the database for the period 1990 to 1995 but in 1992 the installed capacity drops to 20 MW. By 1995 the capacity returns to 22 MW before the unit then drops out of the survey for the period 1996 to 2004. When it returns in 2004 it has a capacity of 22.6, which declines to 22 MW in 2005. By 2010, the installed summer capacity is 15 MW, which remains its capacity to present. During this period there is no change in the operating status of the unit nor was it retired and brought back online.

(i.e., June 1st to September 30th) and is lower than the nameplate capacity since it includes electricity consumption by auxiliaries and the station itself ()(EIA Glossary, 2016). We consider the capacity change in the electric power industry (i.e., we ignore commercial and industrial power). Also, we are only considering the positive change in net summer capacity in our discrete choice model (i.e., we ignore net summer capacity reductions).

Investment cost data is from the GCAM model input data files. The model has overnight cost estimates in 2010 USD per kW for the time periods 1975, 1990, 2005, 2010, and 2015. We use the following simple linear regression model to interpolate between these data points for each technology (k).

$$ln(IC_{kt}) = \beta_{0k} + \beta_{1kt} \cdot Year_t + \beta_{2kt} \cdot tech \cdot year\epsilon \tag{1}$$


Electricity market
Coal
Natural gas
Oil
Geothermal
Water
Uranium
Solar
Wind
Coal
IGCC

4. Model

These types of models are usually specified as a set of individuals (n) choosing between some alternatives (j). We propose that the total net change in capacity, which varies year by year, is the result of an exogenous process that exists outside the model. The purpose of this model is to determine the factors that effect the technology choice, which is a discrete choice problem.

The determinants of the choice made by each individual amongst the entire choice set can be classified according to three indexes: the properties of each alternative, the choice situation, and properties of the individual (Croissant mlogit package). While we do face a repeated choice situation we don't have any individual decision makers in our analysis (i.e., we don't have his/her income information). We use the mlogit package in R for the data analysis (Croissant 2013).

At least three types of variables are possible in the multinomial logit framework: alternative specific variables (x_{ij}) with generic effects (β); individual specific variables (z_i) with an alternative specific coefficient (γ_j); and alternative specific variables (w_j) with alternative specific coefficients (λ_j)(Croissant 2013). Are they properties of the choice set or the individual decision maker? In our case,

Our dataset is a repeated cross section, whereby we have capacity additions over time but we don't follow the individuals making the decisions. The dependent variable of the analysis is the annual new capacity, weighted by the capacity factor, for each production technology, primary fuel pairing. Given that our dependent variable is the annual new capacity, which is allocated across several different production technologies and fuel pairings, it is not mutually exclusive, which is an underlying assumption of the discrete choice modeling framework. To address this issue we could model the decision process for each

4.1. Logistic implementation.

If a random variable Y_i can take one of several discrete values $1, 2, \ldots, J$ then the probability that the i-th response is in the j-th category is λ_{ij} (i.e., $\lambda_{ij} = \Pr\{Y_i = j\}$). Assuming the categories are both mutually exclusive and exhaustive then $\sum_{j=1}^{J} \lambda_{ij} = 1$.[4] The log-odds as estimated by the multinomial logit model then is $log\frac{\lambda_{ij}}{\lambda_{iJ}} = x_i^T \beta_j$. The multinomial logit model estimates $J-1$ equations that compare each of the $J-1$ categories to the baseline category J (i.e., the multinomial logit model reduces to the standard logistic regression in the case of $J = 2$.

In the multinomial case (i.e., more than two choices) the ...

The dependent variable of interest is the market share of each technology (or fuel type) $S_1, S_2, \ldots, S_B$ with each share defined as $S_b = \frac{Q_b}{\sum_B Q_B}$. We assume that Q_b (i.e., total sales, production, etc.) follows some Cobb-Douglas type production function, $Q_B = e^{\alpha_b} p_b^{\beta_b}$, which can be rewritten as $Q_B = exp\left(\alpha_b + \beta_b \cdot ln(p_b)\right)$. We can rewrite the market share (S_{bt}) as a function of the factors of production (see Equation 7), where S_{bt} is interpreted as the likelihood that a technology b is selected at time period t.

For the logit specification the log likelihood function is $ln(\mathcal{L}(\beta)) = \sum_n \sum_i y_{ni} \left(\beta^T x_{ni}\right) - \sum_n \sum_i y_{ni} ln\left(\sum_j e^{\beta^T x_{nj}}\right)$. The maximum likelihood estimates of the β parameter is then $\frac{\partial ln(\mathcal{L}(\beta))}{\partial \beta} = 0$, which is $\sum_n \sum_i (y_{ni} - P_{ni}) \cdot x_{ni}$, where $P_{ni} = \frac{e^{\beta^T x_{ni}}}{\sum_j e^{\beta^T x_{nj}}}$.

The multinomial and Poisson models have an equivalence that we can exploit in order to model the market shares using a generalized linear model (rather than the multinomial logit framework).

[4] The i in this case would represent each time period so it should actually be λ_{tj}.

One advantage of this is our own familiarity with the GLM package in R, as compared to the mlogit package.

If we have a set of Poisson distributed random variables, such as the annual capacity additions in the power sector, $X_1 \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$, ..., $X_k \sim P(\lambda_k)$ where $\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_k$ (at least not necessarily equal) then the conditional distribution (i.e., probability distribution for a sub-population) of the random variables $X = (X_1, X_2, \ldots, x_k)$ is multinomial, $X \sim Mult(n, \pi)$, where $n = \sum_k X_k$ and $\pi_j = \frac{\lambda_j}{\sum_k \lambda_k}$.

The relationship between the Poisson and multinomial model allows us to estimate a Poisson model using a generalized linear model and then calculate the log-odds from the resulting coefficients. If we treat the random counts (Y_{ij}) (i.e., capacity additions by production technology and primary fuel) as Poisson random variables with mean (μ_{ij}) then we can estimate the following log-linear model.

$$log(\mu_{ij}) = \nu + \theta_i + \alpha_j + x_{ij}^T \beta_j \tag{2}$$

The model includes an indicator variable (θ_i) for each group (i.e., a multinomial observation), which ensures that the denominator of the multinomial (n_i) is equivalent () (Guzmans lecture notes).[5] These variables ensure that the equivalence of the multinomial and Poisson models through the denominator of the multinomial model (i.e., $\sum_k \lambda_k$). The denominator of π_j (i.e., $\sum_k \lambda_k$) is equal to $\sum_i p_i$ and the numerator for each π_j (i.e., λ_j) is equal to $\sum_i c_j$. In our model, we include θ_i as 15 dummy variables representing the years of the data, 9 dummy variables for the power production technology and fuel combinations chosen, and each alternative selected at the decision opportunity (i.e,. each response category). The x_{ij} terms are all interaction terms with the grouping variable.

Then the log odds that observation i falls in power production technology/fuel pairing j relative to another pairing can be calculated by equation **??**. The intercept, ν, and group dummies θ_i, cancel out.

$$log\Big(\frac{u_{ij}}{u_{ik}}\Big) = (\alpha_j - \alpha_k) + x_{ij}^T(\beta_j - \beta_k) \tag{3}$$

This is the odds that the observation i would be in category j relative to the category k. The parameters of the multinomial model then are $a_j = \alpha_j - \alpha_k$ and $b_j = \beta_j - \beta_k$. This is equivalent to the multinomial logit model, which assumes a linear model for the log-odds of each response (see equation 4).

$$\gamma_{ij} = log\Big(\frac{u_{ij}}{u_{ik}}\Big) = \alpha_j + x_i^T \beta_j \tag{4}$$

4.2. **Logit model specification assumptions.** In the case of the logistic model the probability of choosing technology choice i is equal to $Pr_{ni} = Pr(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj} \quad \forall \quad j \neq i)$. This can be rewritten as the following $Pr_{ni} = (\epsilon_{nj} < \epsilon_{ni} + V_{ni} - V_{nj} \quad \forall \quad j \neq i)$. Given

[5]Guzman points out that in the multinomial context the denominator (n_i) (i.e., the number of observations in the group) is known but in the Poisson model they are random).

that the density of the unobserved utility is $f(\epsilon_{nj}) = e^{-\epsilon_{nj}} e^{-e^{-\epsilon_{nj}}}$ and the cumulative distribution function is $F(\epsilon_{nj}) = e^{-e^{-\epsilon_{nj}}}$ then the probability of choosing technology i is $Pr_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\epsilon_{nj} + V_{ni} - V_{nj})}} \right) e^{\epsilon_{ni}} e^{-e^{\epsilon_{ni}}} \, d\epsilon_{ni}$. [6] The integral simplifies to $Pr_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$.[7]

We can normalize the model for the scale of the utility by setting the variance equal to $\frac{\pi^2}{6}$. the scale of the utility is irrelevant to the alternative chosen needs to be normalized as the alternative with the highest utility is the same no matter the scale (i.e., the alternative chosen is the same in these two formulations, $U^0_{nj} = V_{nj} + \epsilon_{nj}$ and $U^1_{nj} = \lambda V_{nj} + \lambda \epsilon_{nj}$). Researchers usually do this through the normalization of the variance of the error terms. If we assume the error terms are iid then we can implement this by dividing the model coefficients by the standard deviation of the unobserved portion of utility. The coefficients now represent the 'effect of the observed variables relative to the standard deviation of the unobserved factors' (Train 2009, p. 24). The error variances in a traditional logit are normalized to $\frac{\pi^2}{6}$ (Train 2009). If we feel that we have heteroskedastic errors at any point then we need to review the discussion on p25 (basically just normalize one of the subpopulations and then the variance estimates for the other groups will be relative to that baseline).

The absolute value of U_{nj} does not matter only the difference in utility between alternatives since $Pr(U_{ni} - U_{nj} > 0, \quad \forall \quad j \neq i)$. This fact has implications on variable selection, in that we only include parameters that capture differences across the alternatives. The absolute value of the constant term (k_j) does not matter only the difference. To do this we need to normalize the absolute value of one constant by setting it equal to zero. The remaining $J - 1$ constants are then the difference in constant values for each alternative, as compared to the baseline.[8] The same holds true for any variable that doesn't vary for an individual between alternatives (i.e., individual income). This issue can be addressed through the normalization of the variable or by interacting it with attributes of the alternatives.

The logit model assumes that $\epsilon_{ni} \quad \forall i$ (i.e., technology choice selected) are iid extreme value (i.e., unobserved portion of the utility is uncorrelated over the choice set and have the same variance).[9] The dependence of decisions on prior decision making can also violate the independence assumption inherent to the logit model. The logistic implementation makes the strong assumption that the unobserved portion of the utility function are independent across alternatives. Though it should be noted that through proper model specification we could reduce the importance of the independence assumption. If we specify the utility model well (i.e., $U_{nj} = V_{nj} + \epsilon_{nj}$), accounting for the important differences between alternatives, then the error term should be white noise (i.e., the error for one alternative provides little to no information about the error for another alternative). Should we believe correlation exists across alternatives then we either respecify the model to account for the correlation structure; explore other alternatives later that do not have this restrictive assumption (i.e., GEV, probit, and mixed logit); or continue with the logit specification as an approximation. The generalized extreme value model (GEV) allows for correlation in the unobserved factors over alternatives (if the correlation amongst these factors is zero then the GEV is equivalent to the logit model).

---

[6] The Pr_{ni} can be interpreted as the joint probability of not choosing alternative i multiplied by the probability of choosing alternative i.

[7] V_{ni} is usually specified to be linear in parameters (i.e., not multiplicative).

[8] This can be implemented in the mlogit R package through the 'reflevel' command.

[9] The assumption of independence can be quite limiting.

We implement the model using the mlogit R package. We implement this using the weights command in the mlogit R package.

(1) set reflevel
(2) the logit model is a misspecification if taste variation is at least partly random (a probit or mixed logit is a better alternative)
(3) an implication of independence from irrelevant alternatives (IIA) is that the logit model substitutes proportionally across alternatives (nested logits, probits and mixed logit offer solutions to this problem); tests are available to test the IIA assumption
(4) How to treat the capacity factors? Net summer capacity for wind equals nameplate capacity
(5) as cross sectional or panel data?
(6) For a discrete choice model the choice set must be mutually exclusive, exhaustive, and finite from the decision makers point of view. In our particular circumstance, we can redefine the choice set, as it is possible to use more than one technology, to be any combination of the technology choices such that the set is then mutually exclusive (i.e., A only, B only, or both A and B). To ensure that the list is exhaustive might require that we include a none of the above option as well.
(7) net change in capacity can be negative
(8) if a technology has no chance of being selected it can be removed from the logit specification
(9) the relationship between the logit probability and representative utility is S-shaped which means that if representative utility is very low (relatively), a small increase in the utility of the alternative has little effect on the probability of its being chosen (from a policy perspective this would be like choosing between investing in an area with poor bus service or an area with sufficient bus service: more impact in the area with sufficient service on bus ridership)

$$logit(p) = \beta_0 + \beta_1 \cdot time_i \tag{5}$$

$$logit(p) = \beta_0 + \beta_1 \cdot time_i + \beta \tag{6}$$

$$S_{bt} = \frac{exp\left(\alpha_b + \gamma_t + \beta_b \cdot ln(p_{bt})\right)}{\sum_{k=1}^{B} exp\left(\alpha_k + \gamma_t + \beta_k \cdot ln(p_{kt})\right)} \tag{7}$$

In the literature this model is referred to as the multiplicative competitive interaction (MCI) model ()(Cooper 1988, 1993). The MCI model assumes that a proportional increase in all prices (i.e., all fuel types in this case) would cause no change in the distribution of market shares (assuming equality in the β_b parameters). An alternative model formulation is the multinomial logit (MNL) model, which includes the actual prices (instead of the logged prices). The MNL model assumes that an absolute price increase for all market participants (i.e., all fuel types) would result in no change to the distribution of market shares (assuming all β_b parameters are equal). Of the two assumptions the MCI seems the more realistic one.

We can expand the previous model (see Equation 7) to include cross-price elasticities.

$$Q_B = exp\left(\alpha_b + \gamma_t + \beta_b \cdot ln(p_{bt}) + \sum_{k \neq b} \eta_{bk} \cdot ln(p_k t)\right) \tag{8}$$

5. Next steps

We might consider the nested logit models to incorporate the specific technology choice once the fuel choice has been made. glm implementation with percentage dependent variable and a weight equal to the change in installed MW because the percentages are not equivalent between the years (i.e., the absolute value is different than the relative percentage).