Naveed Sharif
Data Source: Kaiser Permanente
Title: Forecasting Demand for Kaiser's National Account Book of Business
Class: Research Methods, Spring 2018

Introduction:

Kaiser Permanente is a mammoth of a health care provider. If you live in Northern California, the odds of you receiving health care coverage from Kaiser is very high. In fact, one out of four individuals eligible for health care coverage in Northern California are enrolled in a Kaiser Permanente health care plan. That ratio approximates to just under four million lives. Outside of California, such as Georgia or Virginia, that ratio is much smaller. The bulk of the reason is due to brand perception, consumer preferences, and most importantly, Kaiser Permanente's competitive Rate Position[1]. For Kaiser Permanente to successfully generate larger market share outside of California, a robust methodology for forecasting the following year's market share within its existing group accounts is necessary.

Developing this forecast is imperative because currently no robust econometric forecast exists within the membership department. The availability of such a forecast will allow Account Managers to allocate resources to existing accounts with poor market share performance, such as the ability to provide attractive consumer products. Additionally, the Actuarial Committee can set competitive pricing tactics to combat its Rate Position (RP) against competitors.

I developed an econometric forecast that can be used during Kaiser Permanente's (KP) existing account's open enrollment period. The econometric forecast method I used is an Autoregressive Distributed Lag Model (ADL(1,1)). The data is structured as an unbalanced panel format and since I believe there are leftover unobserved effects within the serial correlation in the composite error term, I used the Random Effects (RE) estimator to control for the leftover heterogeneity in the model. The ADL(1,1) RE model has an Adjusted $R^2$ of 87% and a Root Mean Squared Forecast Error (RMSFE) of 0.3702. In other words, the model explains 87% of the variation of change in an existing accounts market share, and the RMSFE makes up only 9% of the errors in the fitted values[2].

Background:

A prior forecast model exists in the membership department. However, it has major flaws. First, it was built over 5 years ago. Ever since the Affordable Health Care Act, a great deal of the health care market has changed. Another issue with the old forecast is that it was built as a cross-sectional model, using only one-year period as a simple OLS regression

$$KPMembership_i = \beta_0 + \beta_1 RP_i + \varepsilon_i$$

where $i$, is for each existing account within KP's book of business. This simple regression has three obvious flaws: (1) RP is endogenous to account characteristics such as industry and product mix offering. RP is also endogenous to the risk profile of the employees within the account. Therefore, the independence in the error term is violated

$$E(\varepsilon \mid X) \neq 0$$

[1] Rate Position = (Competitor Rate / KP Rate) - 1
[2] Residuals / Fitted Values

To illustrate, let us consider using regression notation on observation data to clearly identify where the endogeneity can occur. Suppose,

*[0.1]* $$Y = \beta_0 + \beta_1 X + \varepsilon$$

where *X* is the predictor variable we are interested in estimating (e.g. KP's RP) and *Y* is the outcome variable (e.g. KP Membership) we would like to measure *X's* relationship too. Note, we do not need to do a regression to obtain $\beta_1$, we can calculate it;

*[0.2]* $$\beta_1^{OLS} = \frac{\sigma_{yx}}{\sigma_x^2}$$

However, let's examine how the presence of endogeneity in the question predictor *X* results with a bias in the OLS estimator. If we manipulate equation 0.1 by using covariance algebra (take covariance *X* throughout the equation), we get the following;

*[0.3]* $$\begin{aligned} cov(Y, X) &= cov(\beta_0 + \beta_1 X + \varepsilon, X) \\ . &= \beta_1 cov(X, X) + cov(\varepsilon, X) \\ . &= \beta_1 var(X) + cov(\varepsilon, X) \\ \sigma_{yx} &= \beta_1 \sigma_X^2 + \sigma_{\varepsilon x} \end{aligned}$$

Now dividing throughout equation 0.3 by the population variance of the question predictor *X;*

*[0.4]* $$\frac{\sigma_{yx}}{\sigma_x^2} = \beta_1 + \frac{\sigma_{\varepsilon x}}{\sigma_x^2}$$

The $\frac{\sigma_{\varepsilon x}}{\sigma_x^2}$ ratio in equation 0.4 will give us the magnitude and direction of the bias in $\beta_1$. The larger the ratio, the larger the bias.

Note that in equation 0.4, the population covariance of *Y* with *X*, $\sigma_{yx}$, divided by the population variance of *X*, $\sigma_x^2$, can only be equal to the impact of $\beta_1$, when the second term on the right-hand side of the equation is zero. This in turn can only happen if the population covariance of the predictor and the residual, $\sigma_{\varepsilon x}$, equal to zero.

In other words,

*[0.5]* $$\frac{\sigma_{yx}}{\sigma_x^2} = \beta_1, \quad when\ \sigma_{\varepsilon x} = 0$$

Now that we have a good understanding of where bias enters in our setup, let's move on to the second obvious flaw.

(2) The model is not specified as a time series model, instead it is specified as a cross-sectional dataset. Which is partially explaining the low Adjusted $R^2$ (9%). Finally, (3) the current model does not include any tests or accuracy measurements, such as the Durbin-Watson test for serial correlation and RMSFE for accuracy.

The ADL(1,1) RE model I built also has its flaws, but for different reasons. The length of time points for each existing National account, *i*, is short. The data points only span six years. This restricts the type of forecast model that could be built, such as one with longer lags specified in the model.

Data Description Part 1:

Since I work at KP, there is no shortage in internal KP data. The data I collected is stored in several different databases. I used Microsoft SQL Suite to extract the dataset from tables existing in the databases. The SQL query I wrote extracted (1) competitor and account characteristics such as RP, product offerings from the Sales Connect database, (2) demographic characteristics such as average household income or proportion of owned homes from the GEMS database, (3) member risk characteristics such as claims experience from the Risk Score database, and (4) member characteristics such as a member's age or gender from the SMP database. I merged all datasets together by a unique member ID, account ID, and a member's zip code.

The dataset used for the model is an unbalanced panel dataset. 1,028 existing National accounts, *i*, have been followed (or attempted to follow) between years 2012 through 2018. Making it a total of 4,095 observations. Ideally, for the ADL(1,1) model to forecast properly, I need the dataset to be stationary. In case the time series is nonstationary, I will not have enough time points (since $T$ is small, 6 years) to include both a first difference and an AR(1) term in the model. However, the RE estimator should account for the serial correlation.

The key outcome of interest is estimating demand based on the Berry Inversion[3] method. Suppose a consumer's indirect utility function is assumed to have the form

$$u_{itj} = \beta X_{jt} - \alpha p_{jt} + \xi_{jt} + \varepsilon_{itj}$$

where $j$, is the choice of a health care plan, $p$ is the price of each health care plan, $X$ are observed characteristics, $\xi$ are the unobserved characteristics, and $\varepsilon$ is the idiosyncratic error term. Then the market share for plan $j$ at time $t$ is given by

$$s_{jt}(x, \beta, \alpha, \xi) = \frac{e^{\beta x_{jt} - \alpha p_{jt} + \xi_{jt}}}{\sum_{k=0}^{J} e^{\beta x_{kt} - \alpha p_{kt} + \xi_{kt}}}$$

Normalize the mean utility associated with the outside option to 0

$$u_{i0t} = \varepsilon_{i0t}$$

The probability of choosing the outside plan is then

$$s_{0t}(x, \beta, \alpha, \xi) = \frac{1}{\sum_{k=0}^{J} e^{\beta x_{kt} - \alpha p_{kt} + \xi_{kt}}}$$

[3] Berry, Steven, James Levinsohn, and Ariel Pakes. "Automobile prices in market equilibrium." *Econometrica: Journal of the Econometric Society* (1995): 841-890

The probability of choosing any other plan *j* is then

$$s_{jt}(x, \beta, \alpha, \xi) = \frac{e^{\beta x_{jt} - \alpha p_{jt} + \xi_{jt}}}{\sum_{k=0}^{J} e^{\beta x_{kt} - \alpha p_{kt} + \xi_{kt}}}$$

By taking the log odds ratio, I can get the following linear equation

*[1.0]* $$Log(S_{jt}) - Log(S_{0t}) = \beta X_{jt} - \alpha p_{jt} + \xi_{jt} + \varepsilon_{itj}$$

(Note I still have $\xi$, the unobserved characteristics. I will be tackling the unobserved characteristics using a RE estimator in the estimation procedure). The value in using the Berry Inversion method in forecasting KP's demand is that it essentially treats the dependent variable as market share for KP. Most importantly it mirrors the shape of market share in which the boundaries are between 0% and 100%. Therefore, using the Berry Inversion to estimate demand for KP Health Care enrollees for account *i* is

$$[Log(\frac{adds+subscribers}{eligibles}) - Log(1 - \frac{adds+subscribers}{eligibles})]_{it} = Log[\frac{(\frac{adds+subscribers}{eligibles})}{(1 - \frac{adds+subscribers}{eligibles})}]_{it} = Demand_{it}$$

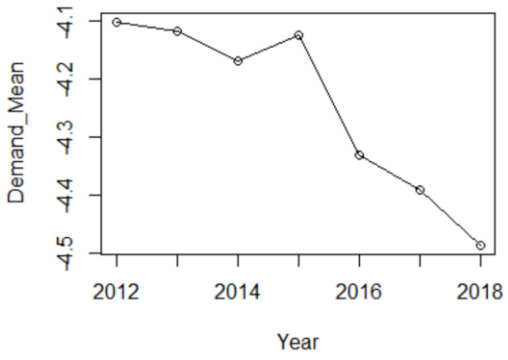
where the coefficients on the right-hand side of the Berry Inversion equation (1.0) are calculated as logit coefficients and can be used to predict market share demand

$$\frac{e^{\beta X_{jt} - \alpha p_{jt} + \xi_{jt} + \varepsilon_{itj}}}{1 + e^{\beta X_{jt} - \alpha p_{jt} + \xi_{jt} + \varepsilon_{itj}}}$$

<u>Data Description Part 2:</u>

The empirical motivation for my paper is driven by the fact that since 2012, KPs change in demand within the National line of business has been decreasing over time

**Chart 1: Change in Demand Over Time**




The National line of business is made up of companies where KP members must be located within three of the eight KP regions. For example, Walmart is a National Account, and its KP members are located within Northern California, Hawaii, and Georgia. Bank of America is also a National Account and its KP members are located within Northern California, Sothern California, Virginia, Oregon, and Hawaii. The National Account business line is one of KP's most competitive business lines. Competitive rates are hard

to come by, but most importantly the business line is driven by the demand for consumer products such as High Deductible products (HDHP). Designing HDHP products has not been one of KP's strengths. Table 1, provides summary statistics of the key variables used in the estimation of KP's demand

Table 1: Summary Statistics

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| Demand | 4,095 | | 1.75 | -10.19 | -5.29 | -2.98 | -1.31 |
| IndustryDescription_RetailTrade | 4,095 | | 0.33 | 0 | 0 | 0 | 1 |
| IndustryDescription_HealthCareandSocialAssistance | 4,095 | | 0.25 | 0 | 0 | 0 | 1 |
| L_MembAveAge | 4,095 | | 0.26 | 2.56 | 3.30 | 3.53 | 4.41 |
| FCR | 4,095 | | 0.46 | 1.00 | 1.70 | 2.31 | 5.50 |
| propOwnedHomes | 4,095 | | 0.06 | 0.25 | 0.54 | 0.61 | 0.74 |
| propBlackHisp | 4,095 | | 0.15 | 0.06 | 0.23 | 0.47 | 0.82 |
| propCollEducated | 4,095 | | 0.07 | 0.36 | 0.61 | 0.70 | 0.86 |
| NumOfCarriers | 4,095 | | 0.94 | 1 | 2 | 3 | 5 |
| KPNumOfUniqueProducts | 4,095 | | 1.00 | 1 | 1 | 2 | 7 |
| CompoffersHSAorHRA | 4,095 | | 0.48 | 0 | 0 | 1 | 1 |
| CompRateMean | 4,095 | | | | | | |
| KPRateMean | 4,095 | | | | | | |

National Accounts within the Retail Trade industry, such as Walmart, have a higher out of pocket expense then members in the Health Care and Social Assistance industry. This makes them more price sensitive and susceptible to the richer health benefit products that KP primarily offers. Exponentiating the members log average age (L_MembAveAge) reveals a very young population. The average age between all groups is thirty-one years old. Young members rarely go to the hospital compared to older members. Therefore, they prefer more consumer driven products, such as an HDHP. The Family Content Ratio (FCR), which is the ratio of the number of dependents that are covered through a subscriber's plan is also low. Having a low FCR is driven by members less likely to be married and to have children, or accounts not offering dependent coverage. Other key variables are number of unique KP product choices offered and if competitors are offering HDHP products (CompOffersHSAorHRA). Finally, both KP and Competitor rates charged to an employer were incorporated in the forecast. KP's rate is $18 cheaper than its Competitor. However, this rate is what is charged to the employer, not the employee. The employee rate, is the out of pocket expense the member sees. This data point was not available.

<u>Estimation Procedure:</u>

I used an Autoregressive Distributed Lag Model, ADL($p$,$q$), to forecast KP's demand for the next year

$$[1.1] \qquad Y_{it} = \beta_0 + \beta_1 Y_{it-1} + \ldots + \beta_p Y_{it-p} + \delta_1 X_{it-1} + \ldots + \delta_q X_{it-q} + \varepsilon_{it}$$

where $Y_{it}$ is estimated using the Berry Inversion technique while $p$ and $q$ are the number of lag dependent and independent variables, respectively

$$[1.2] \qquad [Log(S_{jt}) - Log(S_{0t})]_{it} = \beta_0 + \beta_1 Y_{it-1} + \ldots + \beta_p Y_{it-p} + \delta_1 X_{it-1} + \ldots + \delta_q X_{it-q} + \varepsilon_{it}$$

The ADL method was used over the AR($p$) method

$$[1.3] \qquad [Log(S_{jt}) - Log(S_{0t})]_{it} = \beta_0 + \beta_1 Y_{it-1} + \ldots + \beta_p Y_{it-p} + \delta_1 X_{i,t} + \varepsilon_{it}$$

because of a $q$ variable that significantly improved forecasting KP's demand. I believe there are also left over unobserved effects, $\xi_i$ , within the serial correlation in the composite error term

$$[1.4] \qquad [Log(S_{jt}) - Log(S_{0t})]_{it} = \beta_0 + \beta_1 Y_{it-1} + \ldots + \beta_p Y_{it-p} + \delta_1 X_{it-1} + \ldots + \delta_q X_{it-q} + \xi_i + \varepsilon_{it}$$

Since I believe only exogenous variables are included in the forecast, then I am assuming that $\xi_i$ is uncorrelated with each explanatory variable in all time periods in the model

$$Cov(X_{it}, \xi_i) = 0; \; t = 1, 2, \ldots, T$$

Therefore equation (1.4) becomes a Random Effects model[4]. If I define the composite error as

$$v_{it} = \xi_i + \varepsilon_{it}$$

then the ADL($p$,$q$) equation with a composite error term can be written as

[1.5] $$[Log(S_{jt}) - Log(S_{0t})]_{it} = \beta_0 + \beta_1 Y_{it-1} + \ldots + \beta_p Y_{it-p} + \delta_1 X_{it-1} + \ldots + \delta_q X_{it-q} + v_{it}$$

Because $\xi_i$ is in the composite error in each time period, the $v_{it}$ are serially correlated across time. Under the Random Effects assumption

$$Corr(v_{it}, \xi_{is}) = \sigma_\xi^2 / (\sigma_\xi^2 + \sigma_\varepsilon^2), \; t \neq s$$

This serial correlation can potentially be substantial, and because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as will the usual test statistics. I can use General Least Squares (GLS) to solve the serial correlation problem here. For the procedure to have good properties, I need large $N$ and relatively small $T$. In which case, both I have in the panel dataset.

The GLS transformation that eliminates the serial correlation in the errors can be defined as

[1.6] $$\Theta = 1 - [\sigma_\varepsilon^2 / (\sigma_\varepsilon^2 + T\sigma_\xi^2)]^{1/2}$$

Then the transformed ADL($p$,$q$) equation [1.5] turns out to be

[1.7] $$\{[Log(S_{jt}) - Log(S_{0t})]_{it} - \Theta\overline{[Log(S_{jt}) - Log(S_{0t})]_{it}}\} = \beta_0(1 - \Theta) + \ldots$$

$$+ \beta_1(Y_{it-1} - \Theta\bar{Y}_{it-1}) + \ldots + \beta_p(Y_{it-p} - \Theta\bar{Y}_{it-p}) + \delta_1(X_{it-1} - \Theta\bar{X}_{it-1}) + \ldots + \delta_q(X_{it-q} - \Theta\bar{X}_{it-q}) + (v_{it} - \Theta\bar{v}_{it})$$

where the overbar denotes the time averages and involves quasi-demeaned data on each variable. The RE transformation subtracts a fraction of that time average, where the fraction depends on $\sigma_\xi^2$, $\sigma_\varepsilon^2$ and the number of time periods, $T$.

I can then exponentiate the right-hand side of equation (1.7) and predict the market share for each of KP's National Account, $i$

[1.8] $$\{[Log(S_{jt}) - Log(S_{0t})]_{it} - \Theta\overline{[Log(S_{jt}) - Log(S_{0t})]_{it}}\} =$$

$$\frac{e^{\beta_0(1-\Theta) + \beta_1(Y_{it-1} - \Theta\bar{Y}_{it-1}) + \ldots + \beta_p(Y_{it-p} - \Theta\bar{Y}_{it-p}) + \delta 1(X_{it-1} - \Theta\bar{X}_{it-1}) + \ldots + \delta q(X_{it-q} - \Theta\bar{X}_{it-q}) + (v_{it} - \Theta\bar{v}_{it})}}{1 + e^{\beta_0(1-\Theta) + \beta_1(Y_{it-1} - \Theta\bar{Y}_{it-1}) + \ldots + \beta_p(Y_{it-p} - \Theta\bar{Y}_{it-p}) + \delta 1(X_{it-1} - \Theta\bar{X}_{it-1}) + \ldots + \delta q(X_{it-q} - \Theta\bar{X}_{it-q}) + (v_{it} - \Theta\bar{v}_{it})}}$$

[4] Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. MIT press, 2010.

I used equation (1.8) with an ADL(1,1) model, tested the use of a Fixed Effects (FE) estimator versus a Random Effects estimator using the Hausman Test, and concluded the RE estimator to be most favorable in forecasting next year's demand for KP

*[1.9]*

$$(Demand_{it} - \hat{\Theta}Demand_{it}) =$$

$$\hat{\beta}_0(1 - \hat{\Theta}) + \hat{\beta}_1(Demand_{it-1} - \hat{\Theta}\overline{Demand}_{it-1}) + \hat{\beta}_2(Trend_{it} - \hat{\Theta}\overline{Trend}_{it}) + \hat{\beta}_3(Retail_{it} - \hat{\Theta}\overline{Retail}_{it}) + \dots$$

$$+ \hat{\beta}_4(HealthCare_{it} - \hat{\Theta}\overline{HealthCare}_{it}) + \hat{\beta}_5(Log(MemberAge)_{it} - \hat{\Theta}\overline{Log(MemberAge}_{it}) + \dots$$

$$+ \hat{\beta}_6(FCR_{it} - \hat{\Theta}\overline{FCR}_{it}) + \hat{\beta}_7(PropOwnedHomes_{it} - \hat{\Theta}\overline{PropOwnedHomes}_{it}) + \dots$$

$$+ \hat{\beta}_8(PropBlackHisp_{it} - \hat{\Theta}\overline{PropBlackHisp}_{it}) + \hat{\beta}_9(PropCollEducated_{it} - \hat{\Theta}\overline{PropCollEducated}_{it}) + \dots$$

$$+ \hat{\beta}_{10}(NumofCarriers_{it} - \hat{\Theta}\overline{NumofCarriers}_{it}) + \hat{\beta}_{11}(KPNumofProducts_{it} - \hat{\Theta}\overline{KPNumofProducts}_{it}) + \dots$$

$$+ \hat{\beta}_{12}(CompOffersHSA_{it} - \hat{\Theta}\overline{CompOffersHSA}_{it}) + \hat{\beta}_{13}(CompRate_{it} - \hat{\Theta}\overline{CompRate}_{it}) + \dots$$

$$+ \hat{\beta}_{14}(CompRate_{it-1} - \hat{\Theta}\overline{CompRate}_{it-1}) + (v_{it} - \hat{\Theta}v_{it})$$

I used the Adjusted $R^2$ for the measurement of fit for equation (1.9), the Durbin Watson test to evaluate serial correlation in the idiosyncratic error term, and the RMSFE to measure the forecast accuracy

$$RMSFE = ([E(Y_{t+1} - \hat{Y}_{t+1|j})^2]^{1/2})$$

I would have preferred to hold out a portion of the data in estimating equation (1.9) and evaluate how well the model would forecast the hold out dataset. But, I was left with a small count of time periods in the estimation of the model. Therefore, the hold out criteria was not feasible. Additionally, good practice would have been to evaluate the selection of the length of lags by using the AIC or BIC criterion

$$BIC(p) = Log(\frac{SSR(p)}{T}) + (p + 1)\frac{Log(T)}{T}$$

$$AIC(p) = Log(\frac{SSR(p)}{T}) + (p + 1)\frac{2}{T}$$

Since I only had enough time data points for one lag, I ignored the calculation.

To evaluate the predictive importance of the included lag of a Competitors Average Rate in equation (1.9), I applied an F-Statistic

*[2.0]*

$$F = \frac{(SSR_r - SSR_{ur})/q}{(SSR_{ur})/n - k - 1}$$

where $SSR_r$ is the sum of squared residuals from the estimated forecast with the lag component dropped in the estimation of equation (1.9), $SSR_{ur}$ is the sum of squared residuals from the estimated forecast with the lag included in the estimation of equation (1.9), and $q$ is the number of restrictions imposed in the moving from the unrestricted to the restricted model

$$q = numerator\ degrees\ of\ freedom = df_r - df_{ur}$$

The *SSR* in the denominator of *F* is divided by the degrees of freedom in the unrestricted model, $SSR_{ur}$

$$n - k - 1 \;=\; denominator\ degrees\ of\ freedom \;=\; df_{ur}$$

Therefore, using equation (1.9), the hypothesis was

$$H_0: \beta_{13} = 0, \beta_{14} = 0,$$

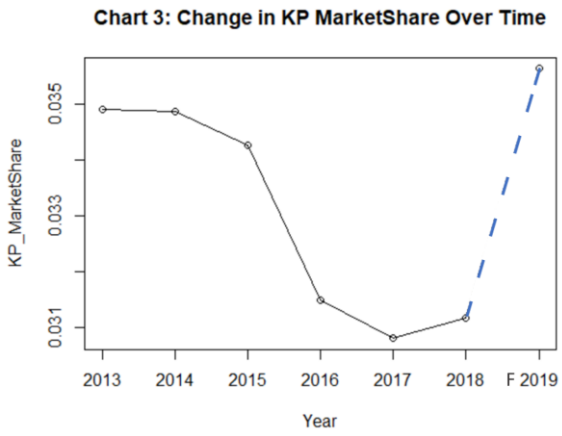
$$H_A: H_0\ is\ not\ true$$

and was rejected in favor of the alternative hypothesis. This led to the ADL(1,1) RE equation (1.9)

Results Part 1:

Table 2: Regression Results

| | Dependent variable: | | | |
|---|---|---|---|---|
| | Demand for Kaiser Permanente | | | |
| | AR(0) Pooled (1) | AR(0) RE (2) | AR(1) RE (3) | ADL(1,1) RE (4) |
| Demand Lag1 | | | | |
| Trend | | | | |
| Industry: Retail | | | | |
| Industry: Health Care | | | | |
| Log(Member Average Age) | | | | |
| Family Content Ratio | | | | |
| Proportion Owned Homes | | | | |
| Proportion Black/Hispanic | | | | |
| Proportion Colledge Educ | | | | |
| # of Carriers | | | | |
| # of KP Unique Products | | | | |
| Competitor Offers HSA/HRA | | | | |
| Avg Competitor Rate | | | | |
| Avg Competitor Rate Lag1 | | | | |
| Constant | | | | |
| Observations | 4,095 | 4,095 | 3,067 | 1,874 |
| R2 | 0.15 | 0.26 | 0.81 | 0.87 |
| Adjusted R2 | 0.15 | 0.26 | 0.81 | 0.87 |
| F Statistic | 61.43*** (df = 12; 4082) | 118.60*** (df = 12; 4082) | 1,012.90*** (df = 13; 3053) | 886.35*** (df = 14; 1859) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 | | | |
| | RMSFE: AR(0) Pooled = 1.61 \| AR(0) RE = .55 \| AR(1) RE = .44 \| ADL(1,1) RE = .37 | | | |




Results Part 2:

I developed four different forecast models, AR(0) Pooled, AR(0) RE, AR(1) RE, and ADL(1,1) RE. The ADL(1,1) RE forecast had the best results. The Adjusted $R^2$ was 87%, and the RMSFE was 0.37 (see notes section in Table 2). Additionally, in the ADL (1,1) RE forecast, all variables used in the model were statistically significant, except for Proportion Black/Hispanic, and Proportion College Educated within an account. However, using the F test from equation (2.0), they were jointly statistically significant

$$H_0: \text{Proportion Black/Hispanic} = 0, \text{Proportion College Educated} = 0; \; H_A: H_0 \text{ is not true}$$

$$F = \frac{(258 - 257)/2}{(257)/1{,}874 - 2 - 1} = 3.64$$

with a *5%* significance level, $q = 2$, and $n - k - 1 = 1{,}871$, the critical value is *3.00*. Therefore, $3.64 > 3.00$. I rejected the null hypothesis, $H_0$, in favor the alternative hypothesis $H_A$, where both Proportion Black/Hispanic and Proportion College Educated was jointly statistically significant.

The estimated $\widehat{\Theta}$, for the RE GLS estimator in equation (1.6), was *0.44*. Since $\widehat{\Theta}$ was far from *0*, a larger fraction of the unobserved characteristics in the error term had been parsed out or quasi-demeaned on each variable. If $\widehat{\Theta}$ had been closer to *1*, it would have made the RE estimator and the FE estimator similar. But $\widehat{\Theta}$ was not close to *1*. In fact, it was reassuring to see that after applying the Hausman Test, the RE estimator was chosen as the better method.

Another test I applied on the forecast ADL(1,1) RE model was the Durbin Watson test for serial correlation. After controlling for Trend in the model, and using robust standard errors, the forecast did not suffer from serial correlation.

Since this is a forecast, the coefficients in the model do not have a casual interpretation. The model was set up using the Berry Inversion technique, therefore the coefficients of the model have a logistic regression interpretation. This did not stop me from applying counterfactuals to the forecast. The counterfactuals I applied was by (1) adding two new products to the "# of KP Products Offered", (2) multiplying "Competitor Offers HSA/HRA" by zero, where I accounted for one of the new products KP offered, as an HSA/HRA product. Therefore, KP and competitor HSA/HRA products are now at parity. Finally, (3) increased the "Average Competitor Rate" by one hundred dollars. In Chart 3, you can spot the marginal significance the three counterfactuals have on KP's forecasted 2019 market share.

Limitations of the Results:

Some limitations to consider were the count of lags I can apply to the forecast. Since $T$ was small, I couldn't test for the length of lags. This led to avoiding a more powerful statistical test like the AIC or BIC statistic. Consequently, the ADL(1,1) RE forecast could have had better results with the addition of lags. I was also limited with differencing the time series, in case of nonstationary, because of a trended series. I did include a Trend regressor, however the time series could have a $2^{nd}$ degree trend that was not factored into the forecast.

If I had access to additional years of data, for each account, the ADL(1,1) forecast could have been tested more thoroughly. Also, Competitor Rates might be suffering from endogeneity issues. The idiosyncratic term could have Competitor data, such as the out of pocket expense the employee sees, or the product mix the Competitor is offering. Potentially both are correlated with Competitor Rates. Thus leading to having bias and inconsistent estimates. However, since the model is about forecasting and less about causal interpretation, the endogeneity issues should be less of a concern.

Conclusion:

In this research paper, I applied three econometric methods that enhanced the accuracy and methodology in forecasting Kaiser Permanente's market share within the National Account business line. The Berry Inversion was used to treat the dependent variable as market share for KP. An Auto Regressive Distributed Lag Model with one dependent and one independent lag variable resulted in the best model fit and accuracy. Finally, a Random Effects estimator was applied to control for any leftover heterogeneity in the model. The value-add in applying econometric methods resulted in a significant lift in the model fit. The forecast Adjusted $R^2$ increased nine times compared to the previous forecast. Which led to the forecast of counterfactuals that would shape strategies in improving KP's market share. Additionally, proper tests and specifications to the estimation procedure paved way for defining limitations and goals for next iterations in improving the forecast in the future.