United States Natality Models 2003-2014

DATA 621: Business Analytics and Data Mining

Daniel Dittenhafer & Justin Hink

April 24, 2016

1 Abstract

2 Keywords

natality, births, demographic theory

3 Literature Review

As a starting point for this study, Daniel Dittenhafer has done prior work analyzing births and unemployment rate in the United States. Dittenhafer found a negative relationship between births and unemployment during the time period studied of 2007 - 2012 (Dittenhafer, 2014). Dittenhafer's single predictor linear model using unemployment rate alone yielded an adjusted R^2 of 0.296 with a p-value approaching 0. Although the unemployment-based model showed some interesting analysis, its usefulness is still to be determined. On the other hand, the negative relationship finding appears to be a relavent output of the study, and one which our current research supports.

Morgan and Taylor published a paper in The Annual Review of Sociology regarding recent fertility trends, and specifically a shift to lower birth rates as compared to the second half of the twentieth century (Morgan and Taylor, 2006). Our research did not relate to this directly, but further reductions in births were observed beginning in 2008. This may constitute another change point or may be a continuation of the trend studied by Morgan and Taylor.

We include women's earnings as a possible predictor in this study based on research by Aliaksandr Amialchuk regarding wage related effects on fertility (Amialchuk, 2013). Across all women, women's education, men's education, men's earnings and metro area were all found to be significant in age-specific fertility regression. We were not able to include age specific earnings, but rather a single earnings measure for women of child bearing age (Bureau of Labor Statistics, 2015).

4 Methodology

4.1 Technology

We used the R runtime via RStudio as our primary environment for all data steps including data preparation, exploration, analysis, model deveopment, validation and selection. A GitHub repository was setup to facilitate collaboration amoungst the team, as well as to share our work in the spirit of reproducible research.

4.2 Data Preparation

Data sets from the Census Bureau, Centers for Disease Control, and Bureau of Labor Statistics identified and downloaded to our project GitHub repository (Dittenhafer and Hink, 2016). These data sets were subsequently joined together in order to provide a unified data set for analysis and modeling.

4.2.1 Natality Data

The natality data including birth counts per month were acquired from the Centers for Disease Control and Prevention in two data sets. The first data set contained data for the years 2003 - 2006 (Centers for Disease Control and Prevention, 2009). The second data set contained data for the years 2007 - 2014 (Centers for Disease Control and Prevention, 2016). The data sets were merged together and augmented with additional census, earning and unemployment data as described in the following sections.

4.2.2 Census Data

For the period of May 2010 - Decemeber 2015, the Census Bureau's census data was available as monthly population estimates broken down by age and gender (Census Bureau, 2015). The age data was in whole year granulatity and we created 10 year buckets for the female population by age: 15-24, 25-34, and 35-44.

For the period of 2000 - April 2010, monthly population estimates were only available for the total population (Census Bureau, 2010). We used annual age and gender estimates from the Census Bureau's 2000 - 2010 time period (converted to ratios) to divide the monthly total population into age and gender bins as shown in the following expressions:

Gender Bins

For each year, 2003 - 2010:

$$G_{year} = \frac{F_{year}}{P_{year}}$$

$$F_{month} = P_{month} * G_{year}$$

$$M_{month} = P_{month} - F_{month}$$

Where:

G Gender Ratio

F Total females, TOT_FEMALE

M Total males, TOT_MALE

P Total population, TOT_POP

Age Bins

Again, for each year, 2003 - 2010:

$$F_{year\_x\_y} = \sum_{i=x}^{y-1} F_{year\_i}$$

$$A_{year\_x\_y} = \frac{F_{year\_x\_y}}{F_{year}}$$

$$F_{month\_x\_y} = F_{month} * A_{year\_x\_y}$$

Where:

x Lower age bound of bin

y Upper age bound of bin

A Age bin's ratio

4.2.3 Earnings Data

The earnings data was acquired from the Bureau of Labor Statistics and specifically covers women's weekly earnings from 2003 - 2015 (Bureau of Labor Statistics, 2015). The acquired data was at a quarter year granularity and was transformed to a monthly granularity for use in this study by simply assigning a quarter's weekly earnings to each of the related 3 months in the 12 month annual period.

4.2.4 Unemployment Data

Unemployment data (U3) was acquired from the Bureau of Labor Statistics. The data was at a monthly granularity with no transformations applied before use in the study (Bureau of Labor Statistics, 2015).

4.3 Data Exploration

We conducted exploratory data analysis to better understand the relationships in the data including correlations, feature distributions and basic summary statistics.

4.4 Model Development

Ten models were developed and examined for significance using a subset (80%) of the original full data set. Gaussian, Poisson and Negative Binomial linear models were fit using a variety of predictor variables and their significance, VIFs, adjusted R^2 and Akaike information criteria (AIC) were examined.

4.5 Model Validation

A validation data set (VS) was created from a subset of the full data set (20%). This VS data set was used to test how well our candidate models generalize to unseen data. The validation metric for the multiple linear regression models is the mean squared error from the validation set.

5 Results

5.1 Data Exploration

5.1.1 Correlations

The following table shows the correlation coefficients associated with each variable and the dependent variable, *Births*.

Table 1: Pearson's r Correlation Coefficients

Births	1.0000000
FEMALE_35_44	0.3880661
Month	0.3646307
GenderRatio	0.2862173
FEMALE_15_24	-0.2307949
TOT_MALE	-0.3214851
TOT_POP	-0.3219328
TOT_FEMALE	-0.3223760
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666
FEMALE_25_34	-0.3879287

5.1.2 Seasonality

As one might expect, we saw seasonality in the birth data. As shown in the scatter plot, below, August is a very popular month for births. July and September are close behind. This suggests that many conceptions are occuring during the United States holiday season between Thanksgiving and New Years.

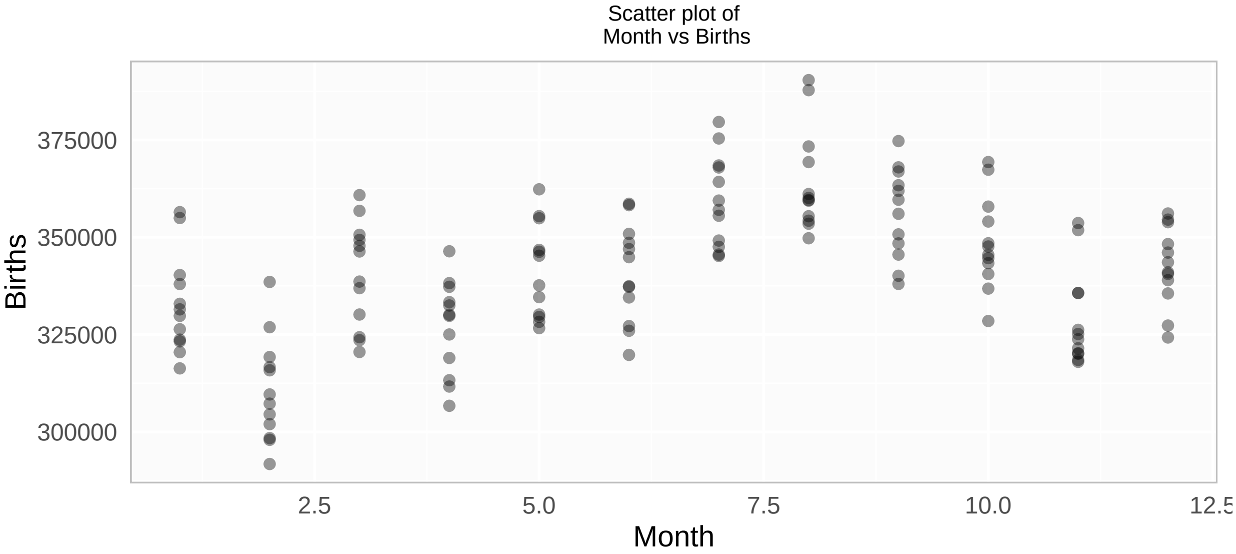


Figure 1: Month vs Births

5.1.3 Gender Ratio

The computed gender ratio which was used to enable the gender buckets for the period of 2003 - 2010 can be seen in the scatterplot below. For these years, the gender ratio is constant for all months of a given year while the birth counts fluctuate. Interestingly, the proportion of females has been dropping steadily, though only slightly during the time period being studied.

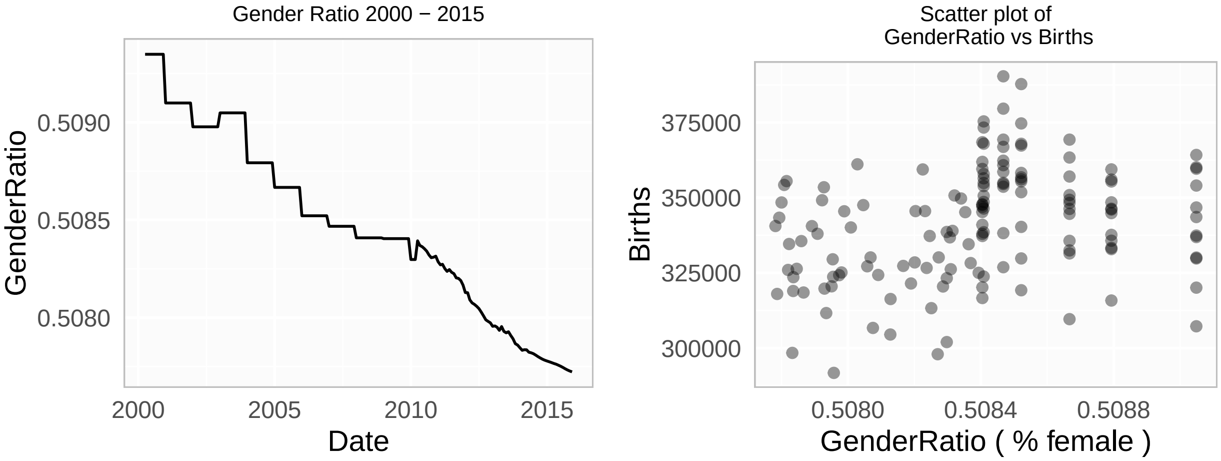


Figure 2: GenderRatio vs Date, GenderRatio vs Births

5.1.4 Earnings

Women's weekly earnings as a broad median value, as reported by the Current Population Survey via the Bureau of Labor Statistics, revealed a negative correlation with births, as previously shown.

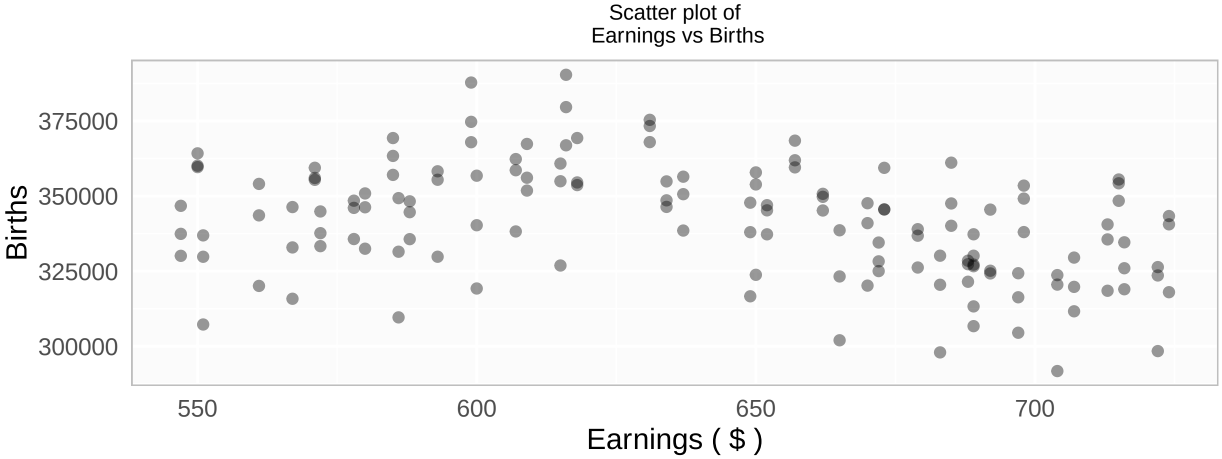


Figure 3: Earnings vs Births

As part of variance inflation factor (VIF) analysis, we found that the earnings measure we used was correlated with female population levels. As shown, in some cases this was a positive relationship (both 15-24 and 25-34), but for the 35-44 age range this was a negative relationship. In general, these relationship resulted in VIFs which significantly exceeded 10 when earnings and the female age values were included in a model. Further study on this relationship may be warranted inorder to better understand the drivers.

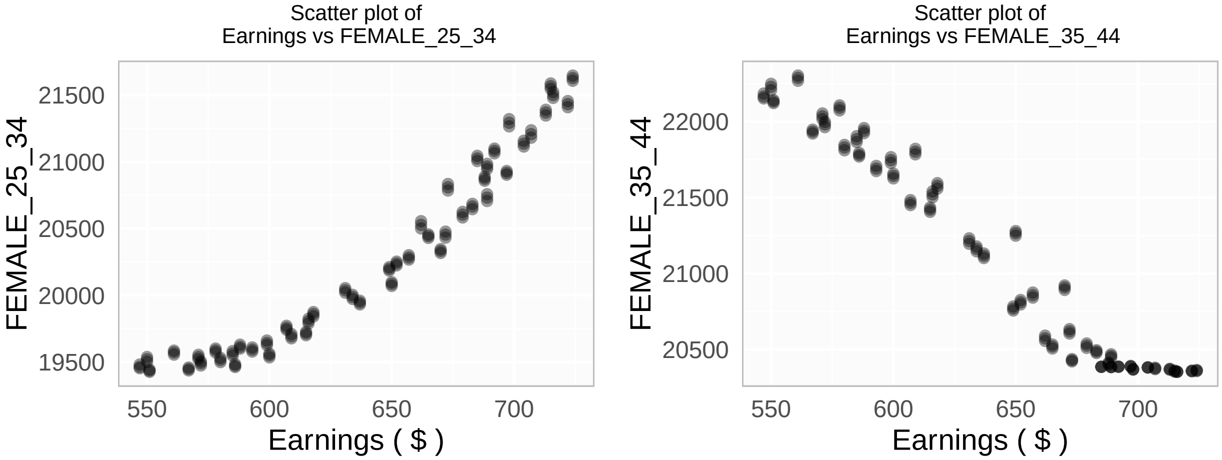


Figure 4: Earnings vs Female Population 25-34 and 35-44

5.2 Model Development

We started with all possible predictor variables from our data sets in a multiple linear regression model. This model yielded an adjusted R^2 of 0.61129, but had significant variance inflation factor issues (> 120,000,000).

Next, we took the significant predictors from the "All Variables" model and created a new, smaller model. Again, VIFs were large. Specifically, the population bins were highly related.

Table 2: Signficant Variables Linear Model VIFs

TOT_POP	97350400.65649
GenderRatio	47225.03676
TOT_FEMALE	93360042.80857
FEMALE_15_24	405.25307
FEMALE_25_34	481.02800
FEMALE_35_44	40.62177
Earnings	88.00442

A new model using guidance from the correlation analysis yielded our "High Correlation Variables" linear model. VIFs were better, but still well above our threshold of 10. Our fourth model was created using R's *step* function. The "Step" model had an impressive adjusted R^2 of 0.62001 for only 5 predictor variables (all significant at $\alpha < 0.05$). The variance inflation factors were improved, but unfortunately, still exceeded our threshold.

Next, we experimented with a revised "Significant Variables Minus" model which was based on the "Significant Variables" model but with three high VIF variables removed. Again, we saw high VIFs. Around this time, the relationship between the female age bins and earnings, and its impact on our data analysis became apparent.

A review of the variables and the introduction of a lag variable, $Month9Ago = Month - 9$, brought us to the next model, "Significant Limited". Again a multiple linear regression model, the "Significant Limited" model included 4 significant variables with an adjusted $R^2 = 0.52589$ and no VIF issues. This was promising, but we continued our investigation of possible models.

Table 3: Significant Variables Limited Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	468350.706460	0.0000000
Month *	2490.804603	0.0000000
Month9Ago *	2649.171132	0.0000000
FEMALE_25_34 *	-7.181586	0.0003757
UnemploymentRate *	-2190.844273	0.0030120

Table 4: Significant Variables Limited Linear Model VIFs

Month	1.062619
Month9Ago	1.028000
FEMALE_25_34	1.400205
UnemploymentRate	1.360376

A Poisson generalized linear model version of the "Significant Limited"" model was produced next. Like its sister model, the "Poisson Significant Ltd" looked good.

Table 5: Poisson Significant Limited Model Coefficient Estimates

	Estimate	Pr(>|z|)
Intercept *	13.1113316	0
Month *	0.0073299	0
Month9Ago *	0.0077043	0
FEMALE_25_34 *	-0.0000210	0
UnemploymentRate *	-0.0064324	0

A component we felt was missing was interaction between population and month. We again developed a model based on the “Signifcant Limited” model, but this time included an interaction term for *FEMALE_25_34* and *Month9Ago*. This did not perform as we anticipated and instead introduced VIF issues without any predictive benefit.

For our final two models we used R’s *stepAIC* to produce Poission and Negative Binomial generalized linear models, respectively. The “Negative Binomial Step” looked promising also with 6 predictor variables, all but one significant, and no VIF issues.

5.3 Model Validation

As mentioned earlier, a validation data set was reserved for use confirming the performance of each of the developed models. We ran each of the ten previously described models through the validation data and computed the mean squared error (MSE) of the resulting model output. We also captured the Akaike Information Criterion (AIC) for each of the models for reference.

The results of the multiple linear regression model validation are shown below.

Table 6: Linear Model Validation Error Results

Model	VS Error	Adj R^2	AIC	Variables	VIF
All Variables	161072343	0.6112935	2504.690	11	BAD
Neg Binomial Step	161290241	NA	2506.956	10	BAD
Poisson Step	161316049	NA	40569.475	10	BAD
Step	172024296	0.6200088	2497.456	5	BAD
Poisson Signif Ltd	176055186	NA	51551.445	4	OK
Significant Limited	176416016	0.5258888	2522.176	4	OK
Signif Ltd w/ Interaction	177767094	0.5218164	2524.118	5	BAD
High Cor	212269346	0.4788873	2535.031	6	BAD
Significant	227028994	0.4771385	2536.351	7	BAD
Significant Minus	231634851	0.3600735	2557.915	5	BAD

The variance inflation factor issues limited us to the “Significant Limited” and “Poisson Significant Ltd” models for further investigation. They have the least complex models and were in the median range in terms of validation performance although the AIC value for the “Poisson Significant Ltd” was strangely high.

5.3.1 In Depth: Significant Limited Linear Model

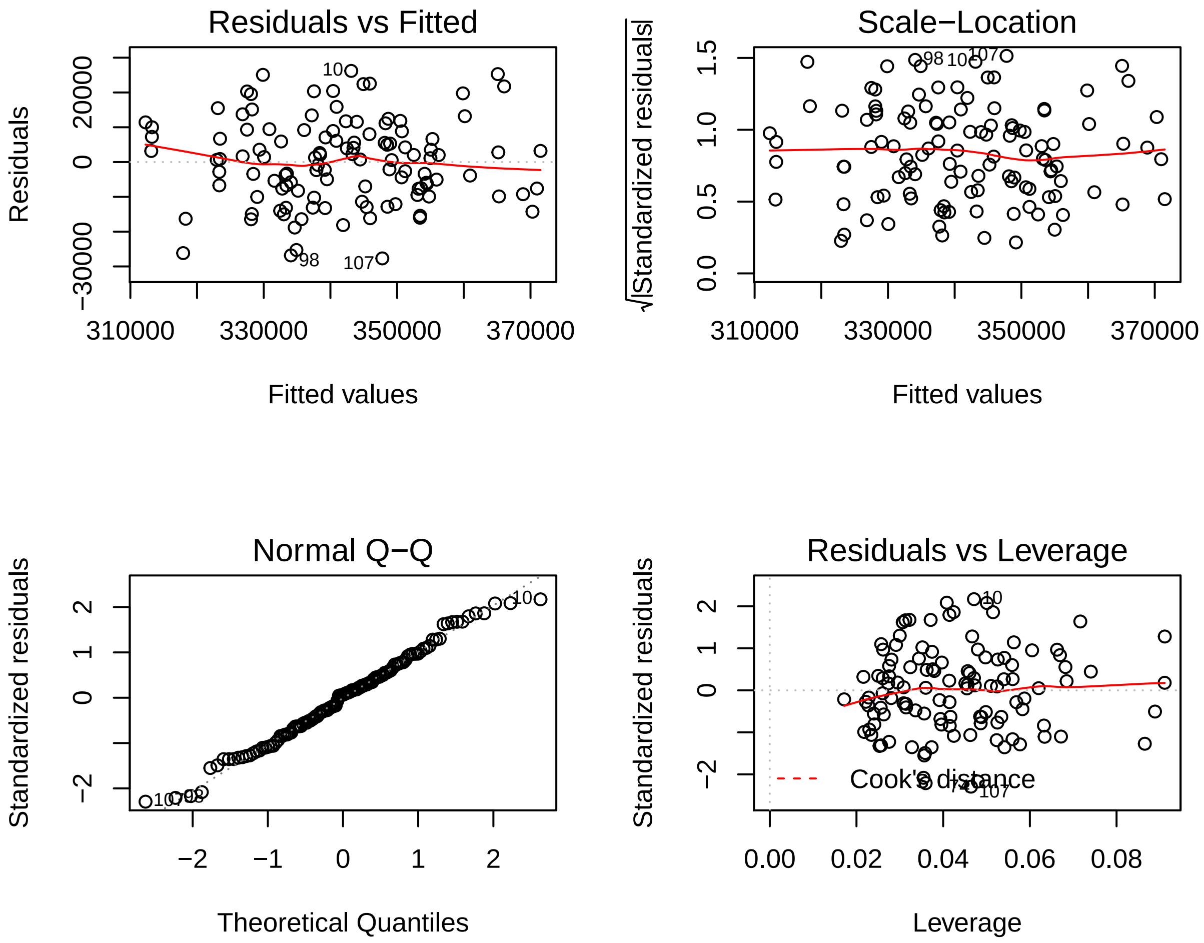
The Significant Limited model has an F-statistic of 32.89 and a mean squared error (MSE) of 146358133.79.

$$\begin{aligned} y_{births} = \quad & 468350.7064601 && +2490.8046026x_{Month} \\ &&& +2649.1711324x_{Month9Ago} \\ &&& -7.181586x_{FEMALE\_25\_34} \\ &&& -2190.8442727x_{UnemploymentRate} \end{aligned}$$

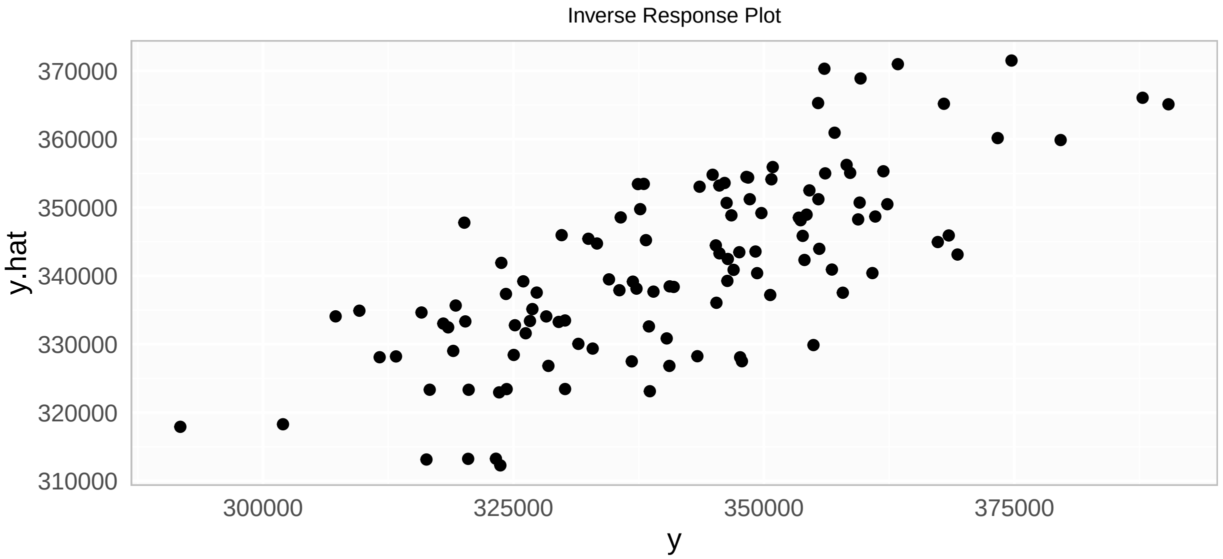
We can interpret the coefficients in the following manner. Holding all other predictors constant, for variable:

- *Month*, as the month of the year increased, a 2490.8 increase in births would occur.
- *Month9Ago*, as the 9 month lagged month of the year increased, a 2649.17 increase in births would occur.
- *FEMALE_25_34*, a unit increase in the population of females age 25-34 would yield a 7.18 decrease in births.
- *UnemploymentRate*, a unit increase in the *UnemploymentRate* related to a 2190.84 decrease in births.

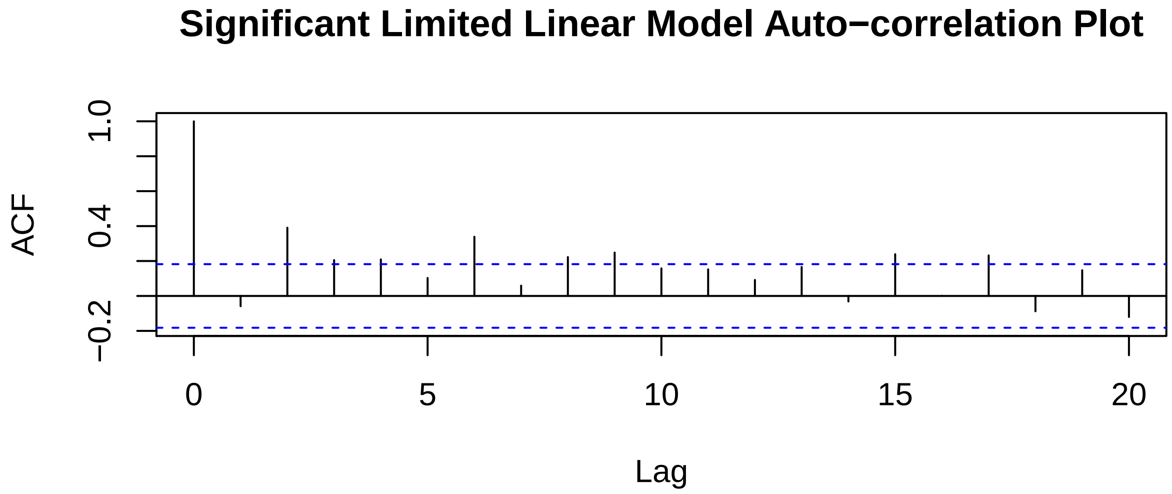
Linear regression diagnostic plots are shown below. Residuals appear to be normally distributed and variance seems to be fairly constant.



Looking at the inverse response plot, there does appear to be a good linear pattern to the predicted response versus actual.



Again, running an auto-correlation analysis with R's *acf* function shows a possible auto-correlation issue with lag 2 and 6.



5.3.2 In Depth: Poisson Significant Limited Model

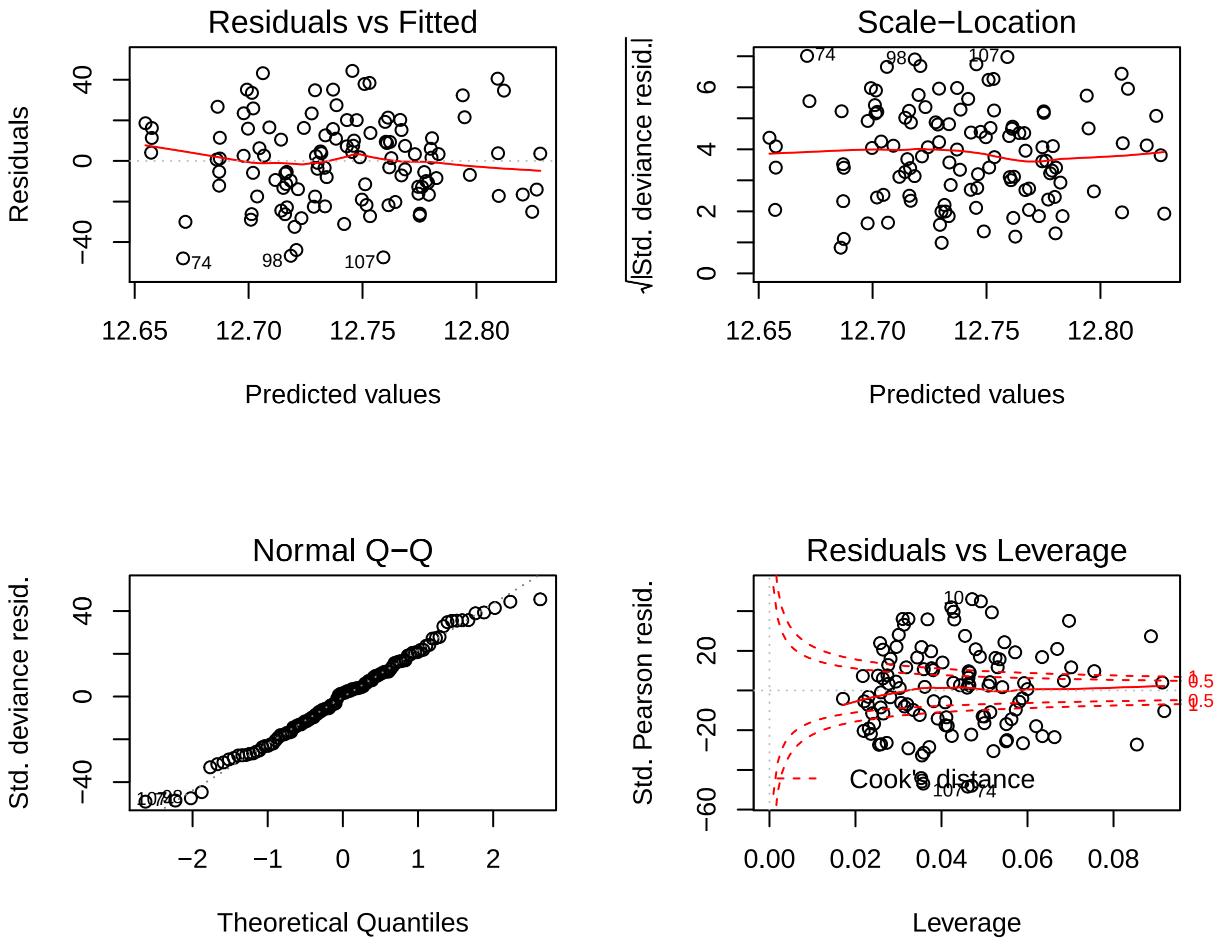
The mathematical form of the Poisson Significant Limited Model is as follows:

$$\begin{aligned} log\left(E(y_{births}|x)\right) = \quad 13.1113316 \quad & +0.0073299x_{Month} \\ & +0.0077043x_{Month9Ago} \\ & -0.000021x_{FEMALE\_25\_34} \\ & -0.0064324x_{UnemploymentRate} \end{aligned}$$

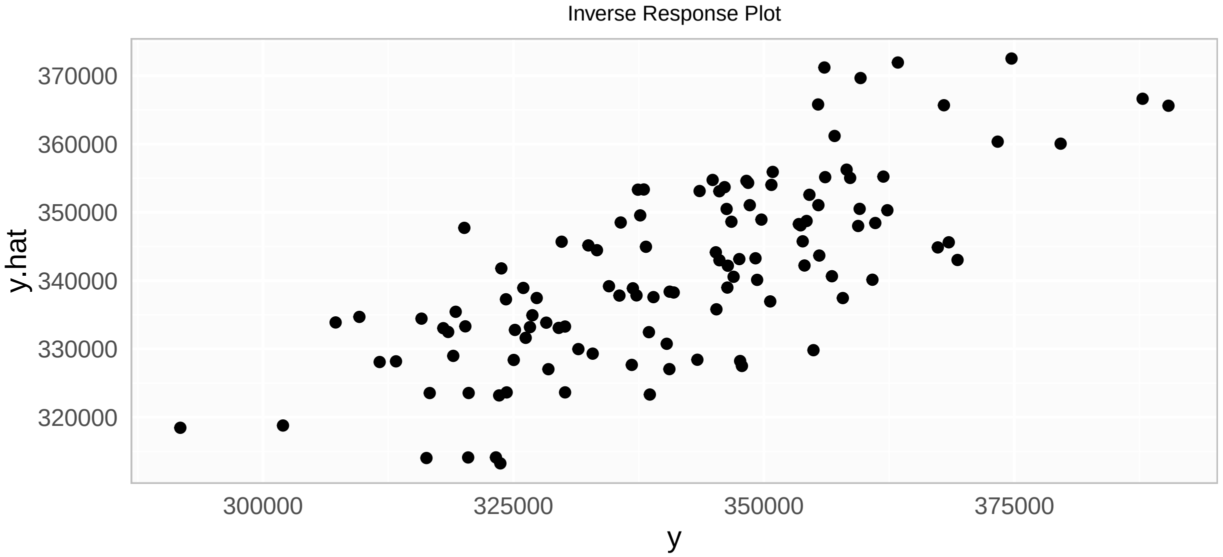
We can interpret the coefficients in the following manner. Holding all other predictors constant, for variable:

- *Month*, as the month of the year increased, a $e^{0.0073299} = 1.0073568$ times increase in births would occur.
- *Month9Ago*, as the 9 month lagged month of the year increased, a $e^{0.0077043} = 1.007734$ times increase in births would occur.
- *FEMALE_25_34*, a unit increase in the population of females age 25-34 would yield a $e^{-0.000021} = 0.999979$ times decrease in births.
- *UnemploymentRate*, a unit increase in the *UnemploymentRate* related to a $e^{-0.0064324} = 0.9935882$ times decrease in births.

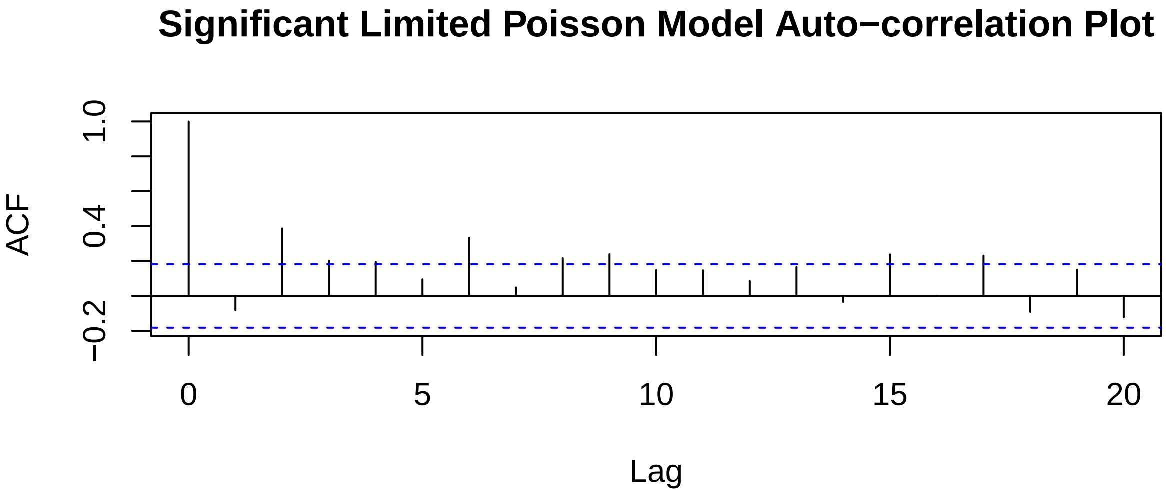
Regression diagnostic plots are shown below. Residuals appear to be normally distributed and variance seems to be fairly constant. The Leverage plot in the lower right shows many points which exceed Cook's distance and suggest points of high leverage.



Looking at the inverse response plot, there does appear to be a good linear pattern to the predicted response versus actual.



Again, running a more targeted auto-correlation analysis with R's *acf* function shows the same possible auto-correlation issue with lag 2 and 6.



6 Summary

7 Appendix: Supplemental Tables & Figures

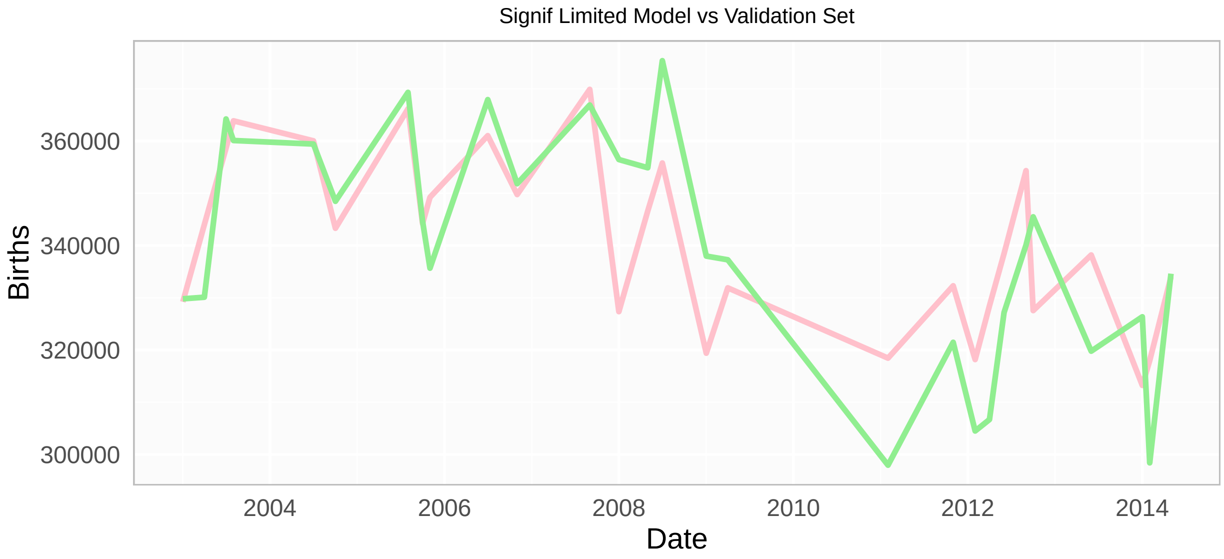


Figure 5: Significant Limited Model vs Validation Set

- Green = Validation Set
- Pink = Model Prediction

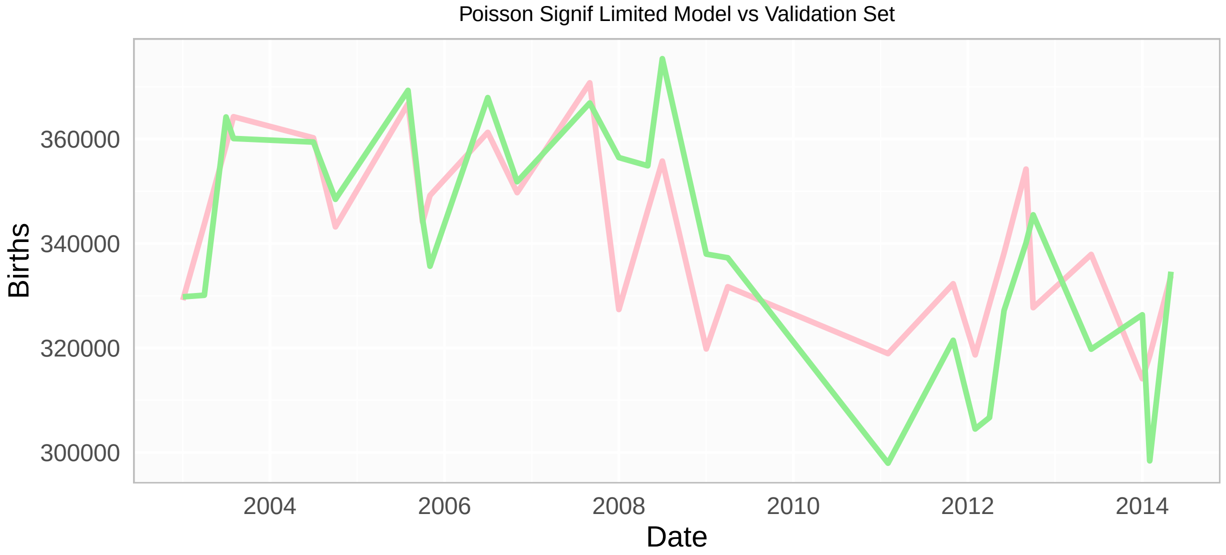


Figure 6: Poisson Significant Limited Model vs Validation Set

8 Appendix: Raw Code

9 References

Amialchuk, A. “Relative Wage Changes and Fertility in the US”. In: Eastern Economic Journal 39(2) (2013). DOI: 10.1057/eej.2013.2.

Bureau of Labor Statistics. Labor Force Statistics from the Current Population Survey 2003-2015 - LNS14000000. Accessed: April 24, 2016. 2015. URL: http://data.bls.gov/timeseries/LNS14000000.

— Median wkly earnings, Emp FT, Wage & sal wrkrs, Women - LEU0252882700. Accessed: March 10, 2016. 2015. URL: http://data.bls.gov/cgi-bin/surveymost?le.

Census Bureau. Monthly Postcensal Resident Population, by single year of age, sex, race, and Hispanic origin. Accessed: April 24, 2016. 2015. URL: http://www.census.gov/popest/data/national/asrh/2014/2014-nat-res.html.

— National Intercensal Estimates (2000-2010). Accessed: April 24, 2016. 2010. URL: http://www.census.gov/popest/data/intercensa

Centers for Disease Control and Prevention. Natality public-use data on CDC WONDER Online Database for years 2003-2006 available March 2009. Accessed: March 1, 2016. 2009. URL: http://wonder.cdc.gov/natality-v2006.html.

— Natality public-use data on CDC WONDER Online Database for years 2007-2014 available February 2016. Accessed: March 1, 2016. 2016. URL: http://wonder.cdc.gov/natality-current.html.

Dittenhafer, D. U.S. Births & Unemployment Rate 2007 - 2012. 2014. URL: https://github.com/dwdii/DataAcqMgmt/raw/master/
Dittenhafer-USBirthsAnalysis.pdf.

Dittenhafer, D. and J. Hink. NatalityModels. 2016. URL: https://github.com/dwdii/NatalityModels.

Morgan, S. and M. Taylor. Low Fertility at the Turn of the Twenty-First Century. Aug. 2006. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2849172/.