Natality Models

DATA621 Business Analytics & Data Mining - CUNY

Daniel Dittenhafer & Justin Hink

May 9, 2016

Agenda

- Research Question
- Data Sets
- Methodology
- Models
- Results

Research Question

Using national aggregate data, how can the number of births be forecast? What factors are significant predictors of births in the United States?

Why is this important?

-Having a glimpse into a society's future population size and demographics can give governments (and private companies) insight into things that need to be done to keep things running efficiently

- In the case of a private company it may illuminate new business areas and profit opportunities

Data Sets

- Natality, 2007-2014
- Natality, 2003-2006
- Women's Earnings 2003-2014 - Current Population Survey
- Unemployment Rates 2003-2015
- Census Estimates 2000-2010
- Census Estimates 2010-2015

Methodology

- Data Tidying
- Joining Data
- Exploration
- Building Models
- Selecting Model(s)

Methodology - Data Tidying

- Census Data
- 2000-2010: Only Total Population
- 2010-2015: Broken down by Age and Gender

Methodology - Gender Bins

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

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

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

Where:

G	Gender Ratio
F	Total females, TOT_FEMALE
M	Total males, TOT_MALE
P	Total population, TOT_POP

Methodology: Age Bins

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

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

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

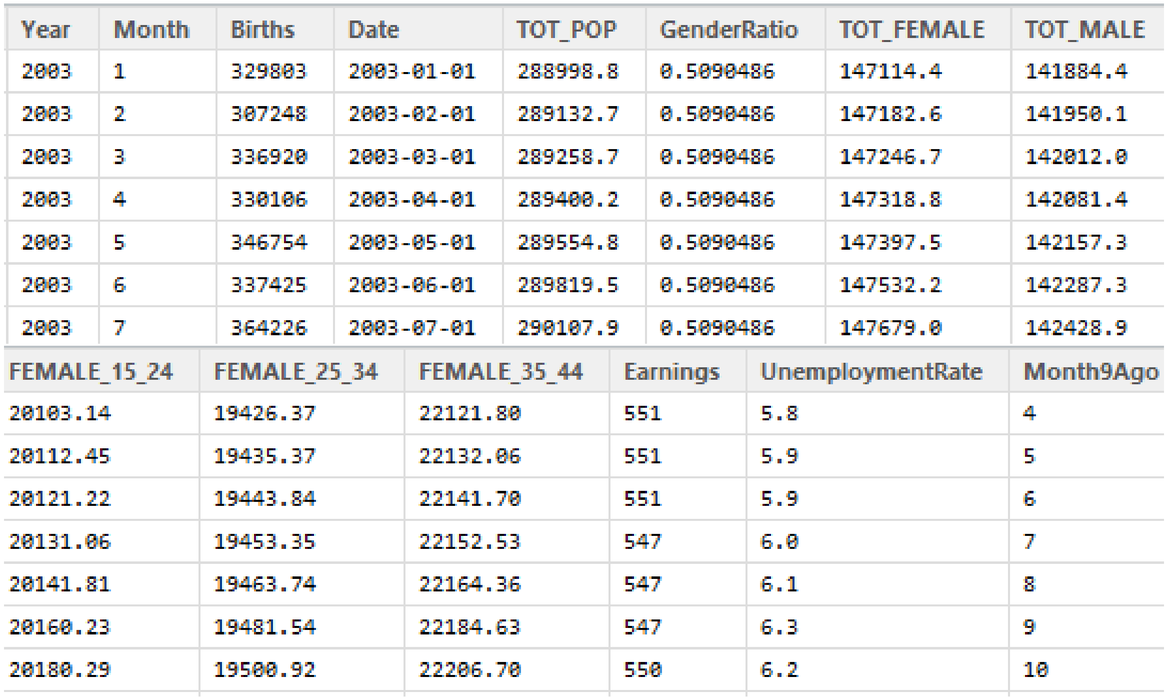
Where:

x Lower age bound of bin

y Upper age bound of bin

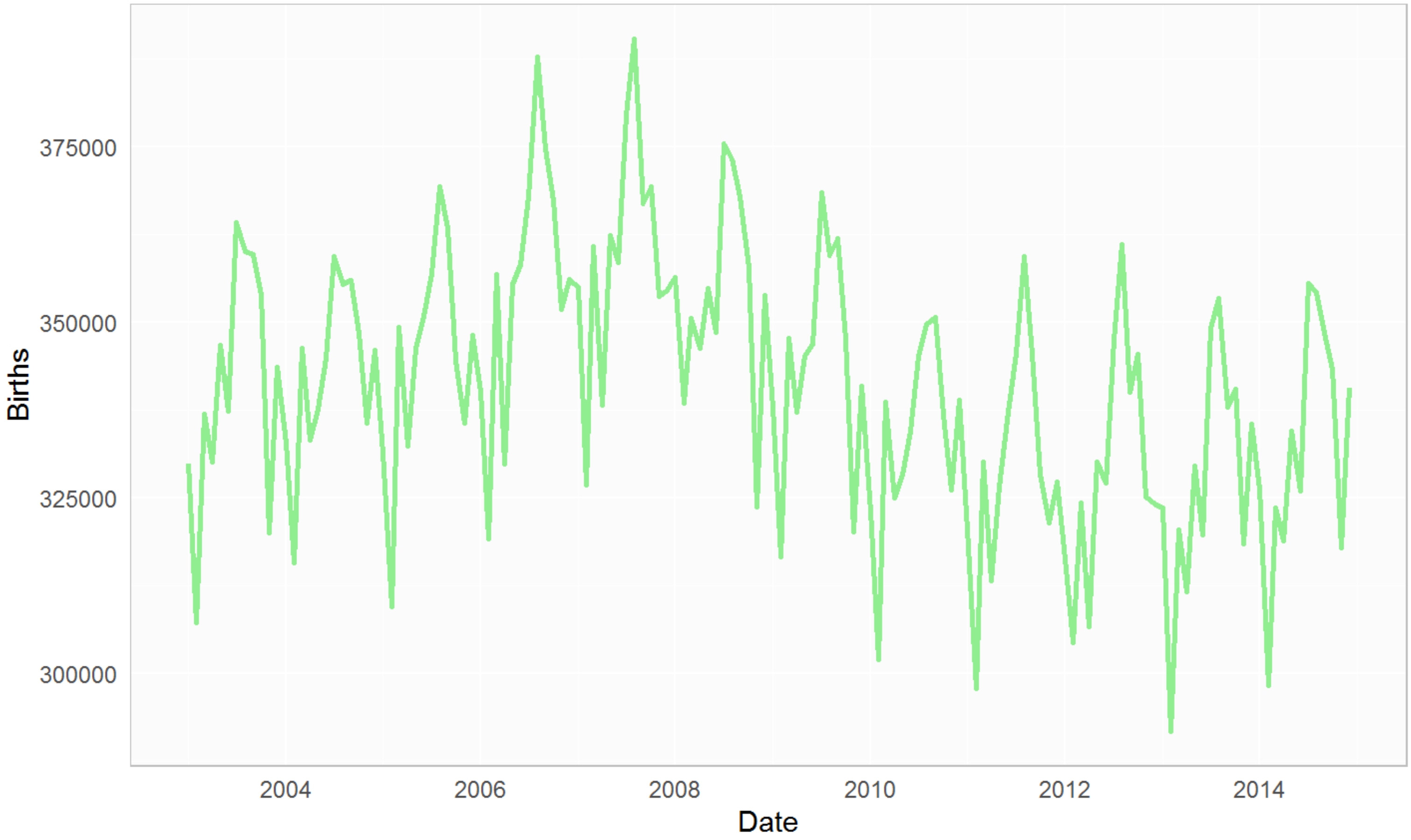
A Age bin's ratio

Model Data

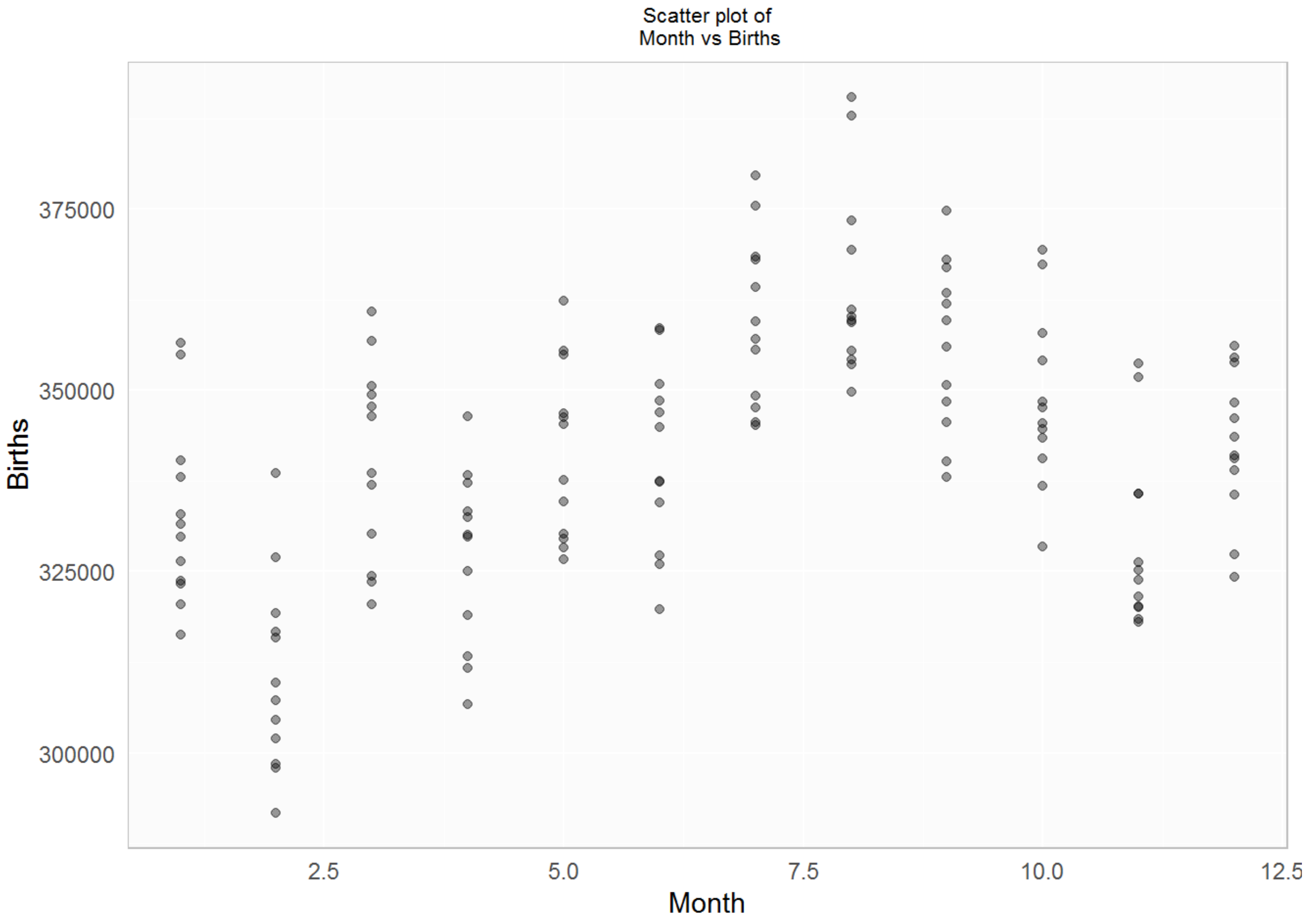


Year	Month	Births	Date	TOT_POP	GenderRatio	TOT_FEMALE	TOT_MALE	FEMALE_15_24	FEMALE_25_34	FEMALE_35_44	Earnings	UnemploymentRate	Month9Ago
2003	1	329803	2003-01-01	288998.8	0.5090486	147114.4	141884.4	20103.14	19426.37	22121.80	551	5.8	4
2003	2	307248	2003-02-01	289132.7	0.5090486	147182.6	141950.1	20112.45	19435.37	22132.06	551	5.9	5
2003	3	336920	2003-03-01	289258.7	0.5090486	147246.7	142012.0	20121.22	19443.84	22141.70	551	5.9	6
2003	4	330106	2003-04-01	289400.2	0.5090486	147318.8	142081.4	20131.06	19453.35	22152.53	547	6.0	7
2003	5	346754	2003-05-01	289554.8	0.5090486	147397.5	142157.3	20141.81	19463.74	22164.36	547	6.1	8
2003	6	337425	2003-06-01	289819.5	0.5090486	147532.2	142287.3	20160.23	19481.54	22184.63	547	6.3	9
2003	7	364226	2003-07-01	290107.9	0.5090486	147679.0	142428.9	20180.29	19500.92	22206.70	550	6.2	10

Methodology: Year over Year Birth Rates



Methodology: Birth Rates By Month



Models

- 11 models investigated
  - 7 Multiple Linear Regression models
  - 3 Possion Generalized Linear models
  - 1 Negative Binomial Generalized Linear model
- 80% training data / 20% validation

Models

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

Models: Significant Limited

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

Models

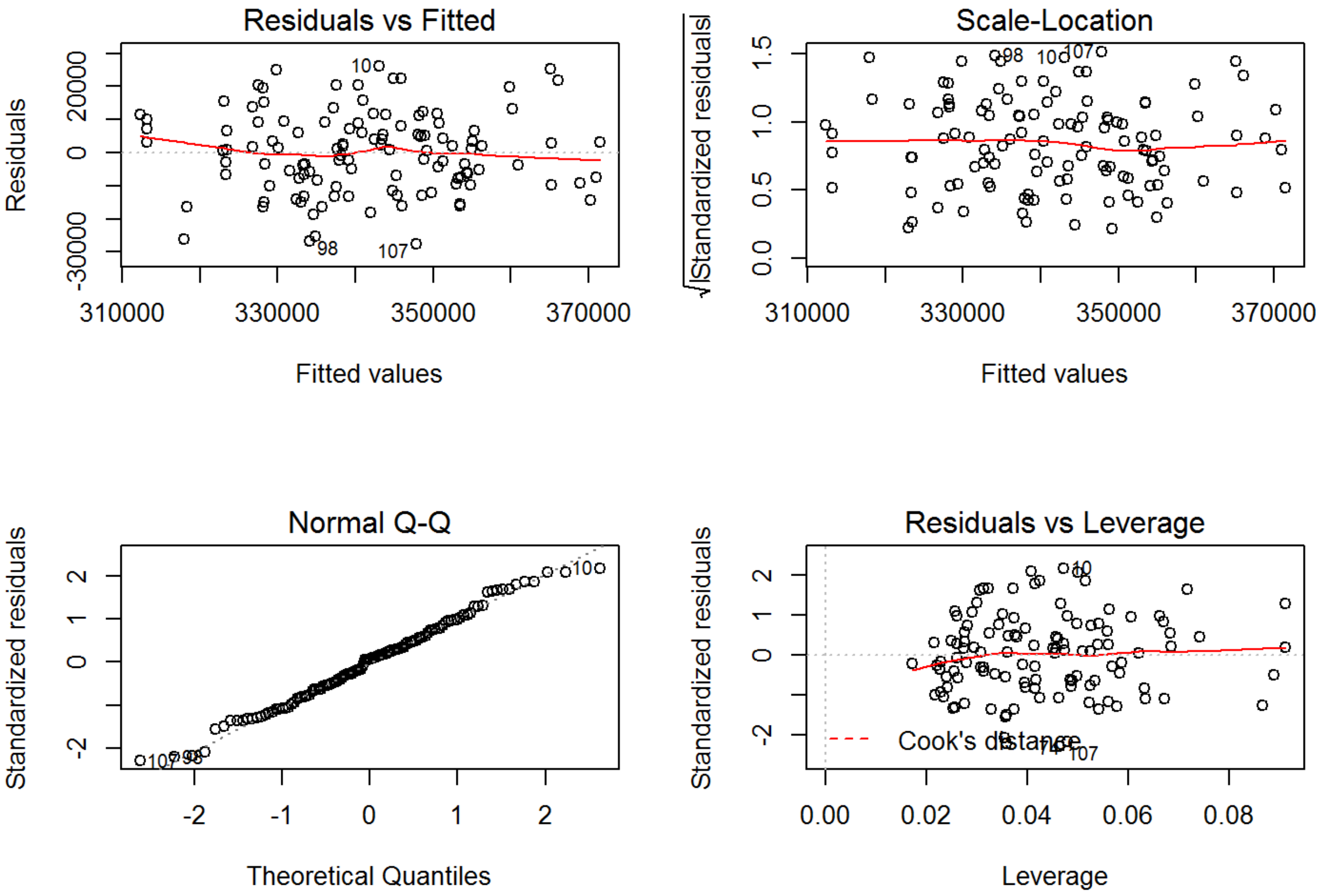
The Significant Limited model has an F-statistic of 32.89 and a mean squared error (MSE) of 146358133.79.

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

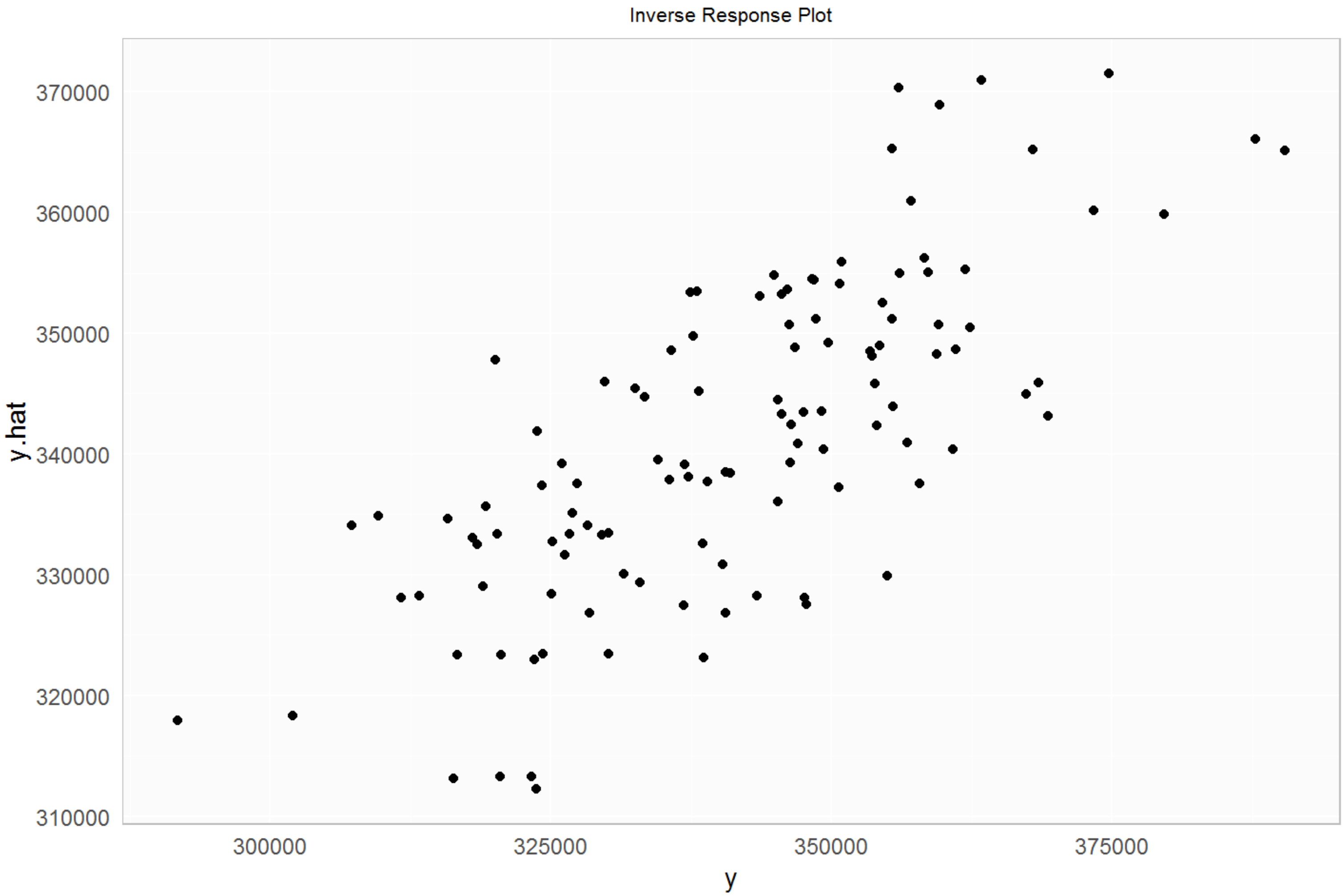
We can interpret the coefficients in the following manner. Holding all other predictors constant, for variable:

- *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.

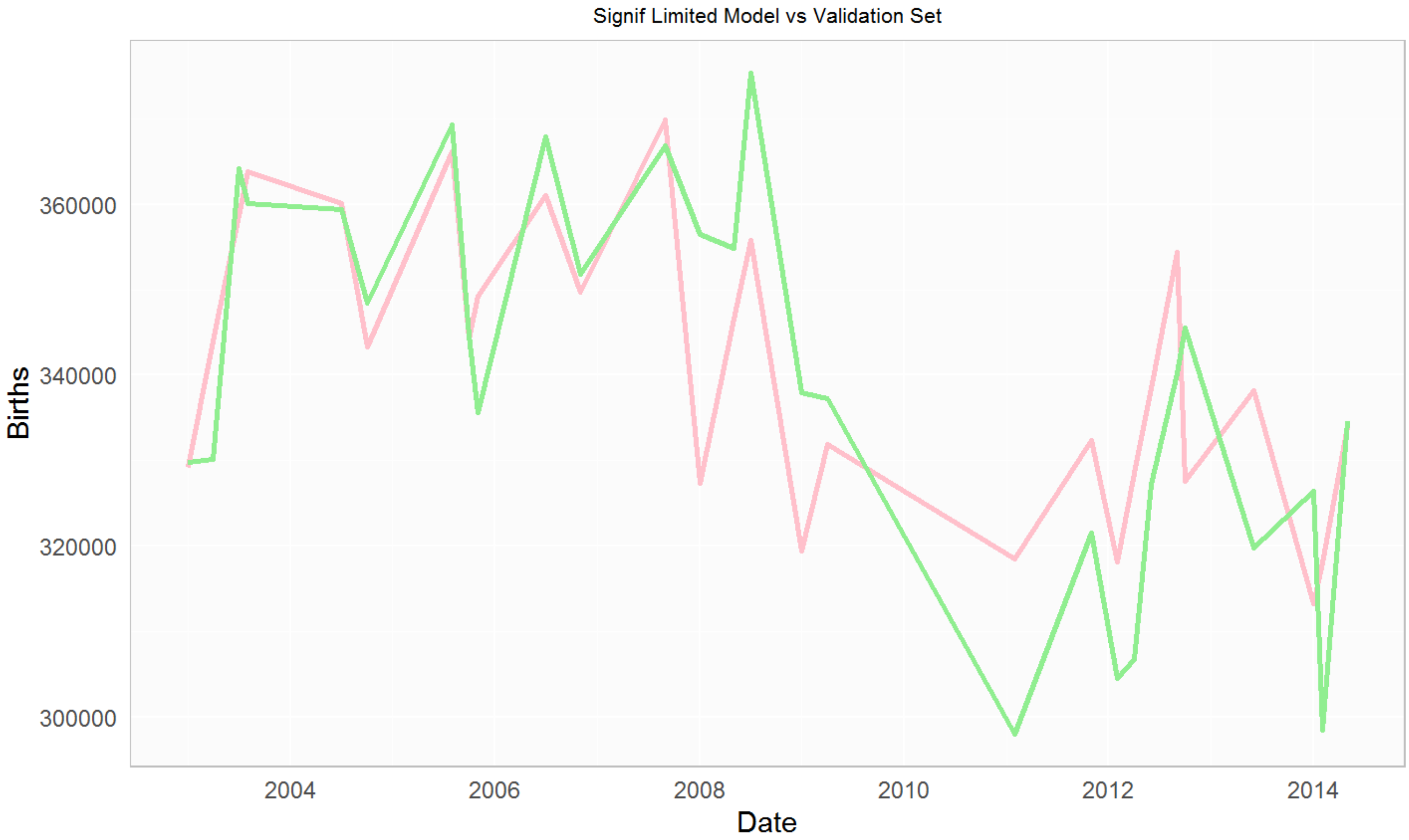
Results: Diagnostics



Results: Inverse Response Plot



Results



- Red = Model
- Green = Actual

Conclusions

- Reasonable for the time period studied
- Changepoint during recession?
- Other independent variables we aren't including?
- Further validation against historical data, 1980s, 1990s?

Thank you!

- Questions?