Natality Models Data Exploration

DATA 621: Business Analytics and Data Mining

Daniel Dittenhafer & Justin Hink

April 24, 2016

```
## Start:  AIC=2504.96
## Births ~ Month + TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate +
##     Month9Ago
##
##                     Df    AIC
## - TOT_FEMALE         1 2503.0
## - TOT_POP            1 2503.0
## - FEMALE_35_44       1 2503.0
## - GenderRatio        1 2503.0
## - Earnings           1 2503.1
## - FEMALE_15_24       1 2503.3
## <none>                 2505.0
## - UnemploymentRate   1 2505.0
## - FEMALE_25_34       1 2507.0
## - Month              1 2516.2
## - Month9Ago          1 2538.4
##
## Step:  AIC=2503.01
## Births ~ Month + TOT_POP + GenderRatio + FEMALE_15_24 + FEMALE_25_34 +
##     FEMALE_35_44 + Earnings + UnemploymentRate + Month9Ago
##
##                     Df    AIC
## - Earnings           1 2501.2
## - FEMALE_35_44       1 2501.2
## - FEMALE_15_24       1 2501.8
## - GenderRatio        1 2502.4
## <none>                 2503.0
## - UnemploymentRate   1 2503.0
## - TOT_POP            1 2504.5
## - FEMALE_25_34       1 2506.1
## - Month              1 2516.3
## - Month9Ago          1 2537.1
##
## Step:  AIC=2501.17
## Births ~ Month + TOT_POP + GenderRatio + FEMALE_15_24 + FEMALE_25_34 +
##     FEMALE_35_44 + UnemploymentRate + Month9Ago
##
##                     Df    AIC
## - FEMALE_35_44       1 2499.3
## - FEMALE_15_24       1 2500.1
## - GenderRatio        1 2500.6
## <none>                 2501.2
## - UnemploymentRate   1 2501.9
## - TOT_POP            1 2502.6
## - FEMALE_25_34       1 2504.3
## - Month              1 2519.2
## - Month9Ago          1 2559.5
```

```
##
## Step:  AIC=2499.34
## Births ~ Month + TOT_POP + GenderRatio + FEMALE_15_24 + FEMALE_25_34 +
##     UnemploymentRate + Month9Ago
##
##                     Df    AIC
## - FEMALE_15_24       1 2498.2
## - GenderRatio        1 2498.8
## <none>                 2499.3
## - TOT_POP            1 2500.6
## - FEMALE_25_34       1 2502.3
## - UnemploymentRate   1 2503.6
## - Month              1 2535.5
## - Month9Ago          1 2557.6
##
## Step:  AIC=2498.18
## Births ~ Month + TOT_POP + GenderRatio + FEMALE_25_34 + UnemploymentRate +
##     Month9Ago
##
##                     Df    AIC
## - GenderRatio        1 2497.5
## <none>                 2498.2
## - TOT_POP            1 2505.7
## - UnemploymentRate   1 2506.6
## - FEMALE_25_34       1 2531.5
## - Month              1 2534.4
## - Month9Ago          1 2555.7
##
## Step:  AIC=2497.51
## Births ~ Month + TOT_POP + FEMALE_25_34 + UnemploymentRate +
##     Month9Ago
##
##                     Df    AIC
## <none>                 2497.5
## - UnemploymentRate   1 2514.0
## - TOT_POP            1 2520.9
## - FEMALE_25_34       1 2529.6
## - Month              1 2551.1
## - Month9Ago          1 2554.6

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Births ~ Month + TOT_POP + GenderRatio + TOT_FEMALE + FEMALE_15_24 +
##     FEMALE_25_34 + FEMALE_35_44 + Earnings + UnemploymentRate +
##     Month9Ago
##
## Final Model:
## Births ~ Month + TOT_POP + FEMALE_25_34 + UnemploymentRate +
##     Month9Ago
##
##
##              Step Df       Deviance Resid. Df Resid. Dev      AIC
## 1                                         105   116.0240 2504.956
## 2   - TOT_FEMALE  1 0.000077861566       106   116.0239 2503.009
## 3     - Earnings  1 0.000098352098       107   116.0240 2501.171
```

```
## 4 - FEMALE_35_44  1 0.000023449090       108   116.0240 2499.336
## 5 - FEMALE_15_24  1 0.000110356610       109   116.0241 2498.185
## 6  - GenderRatio  1 0.000002090639       110   116.0241 2497.513
```

1 Data Exploration

The unified data set for this project contains 144 rows of data with 1 response variable and 13 predictor variables. An exploration of this data follows.

1.1 Missing Values

An analysis of missing values in the data set revealed 0 variables with incomplete data.

1.2 Correlations

The following table shows Pearson's r correlation coefficients between the numeric independent variables and the response variable *Births*.

Table 1: Pearson's r Correlation Coefficients

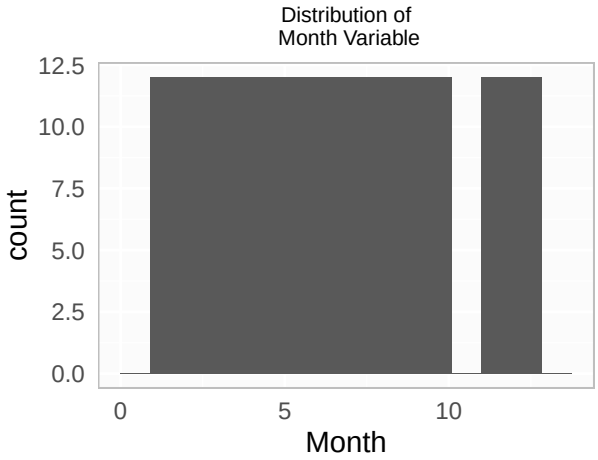
Births	1.0000000
FEMALE_35_44	0.3880661
Month	0.3646307
GenderRatio	0.2862173
FEMALE_15_24	-0.2307949
TOT_MALE	-0.3214851
TOT_POP	-0.3219328
TOT_FEMALE	-0.3223760
Year	-0.3593053
Earnings	-0.3697992
UnemploymentRate	-0.3862666
FEMALE_25_34	-0.3879287

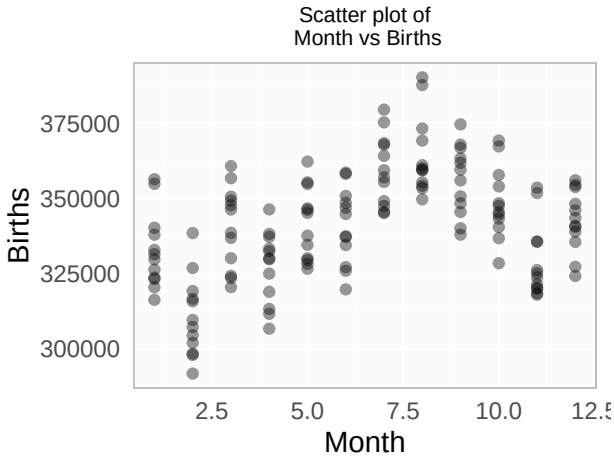
1.3 Variable Month

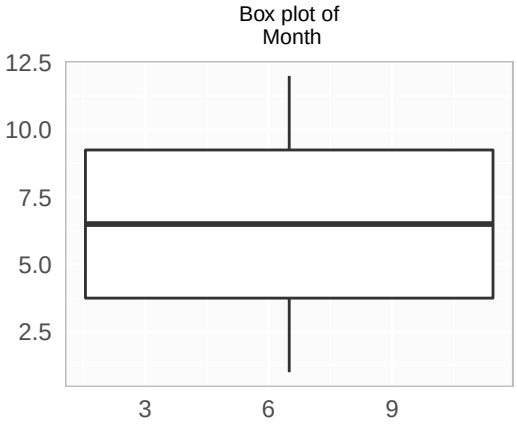
The *Month* variable is the month of birth. As one should expect, the distribution is uniform, but we can see some seasonality to the relationship between *Births* and *Month* with July and August being high frequency birth months.

Table 2: Month Variable Statistics

min	mean	stdev	median	max
1	6.5	3.464102	6.5	12





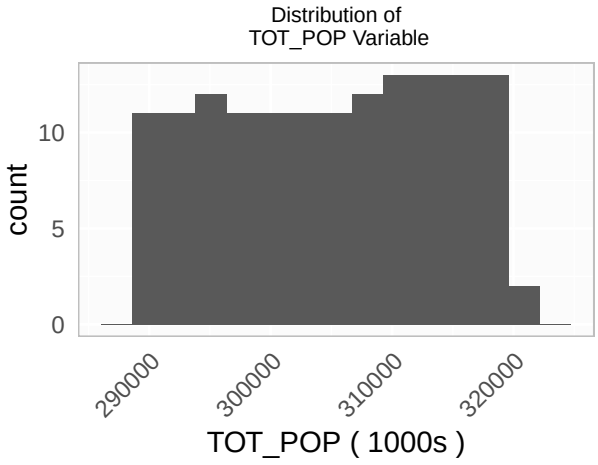


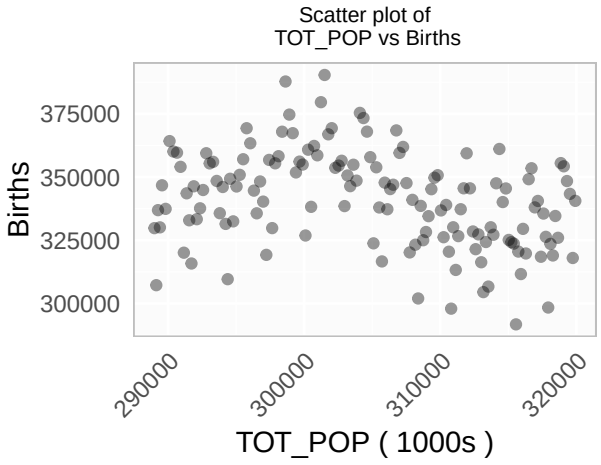
1.4 Variable TOT_POP

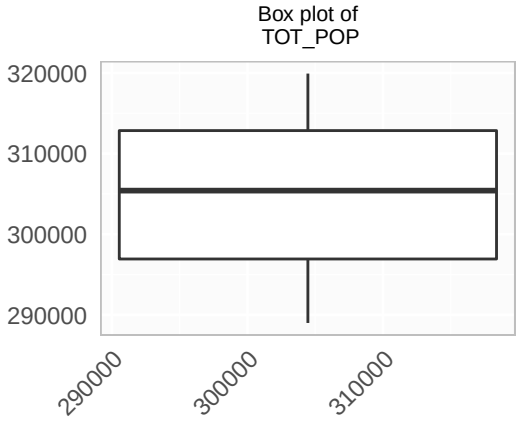
The *TOT_POP* variable is the total population per month as esimated by the Census Bureau.

Table 3: TOT_POP Variable Statistics

min	mean	stdev	median	max
288998.8	304885.4	9171.506	305409.3	319925.2





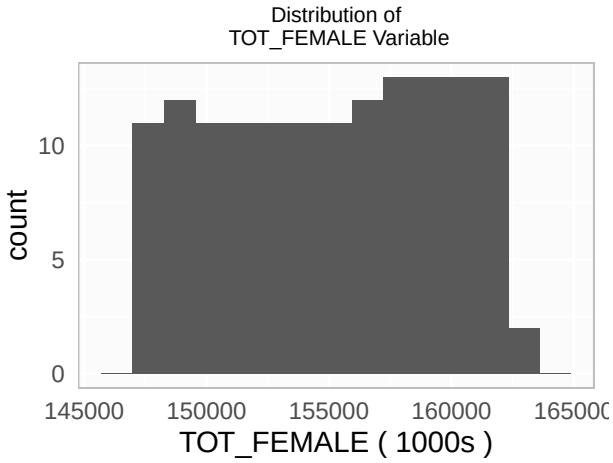


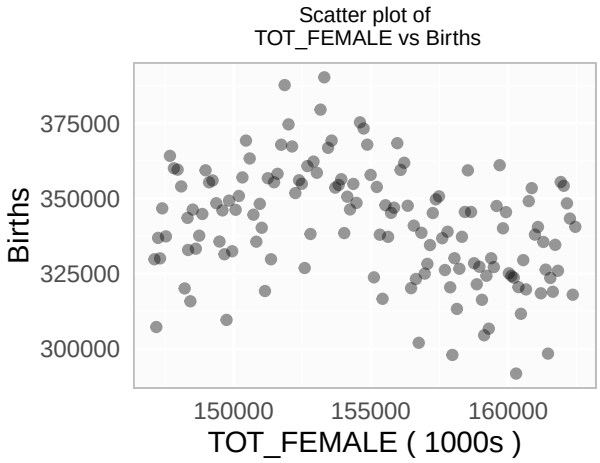
1.5 Variable TOT_FEMALE

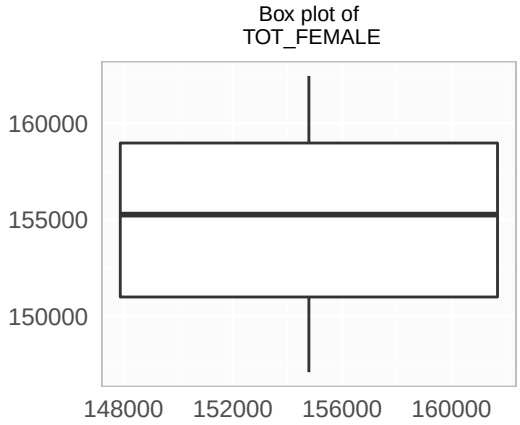
The *TOT_FEMALE* variable is the total population of females per month as estimated by the Census Bureau.

Table 4: TOT_FEMALE Variable Statistics

min	mean	stdev	median	max
147114.4	154997.1	4561.405	155272.1	162452.2





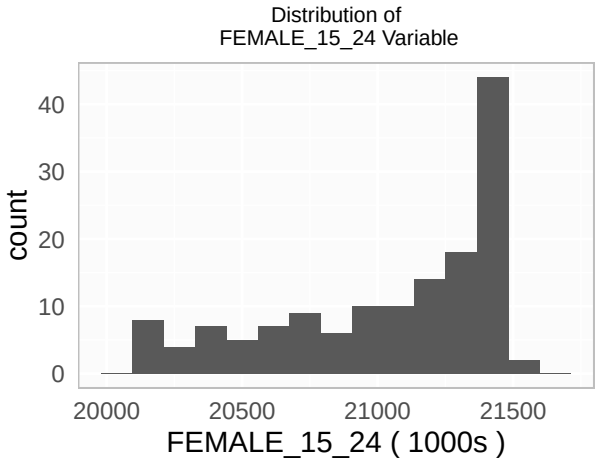


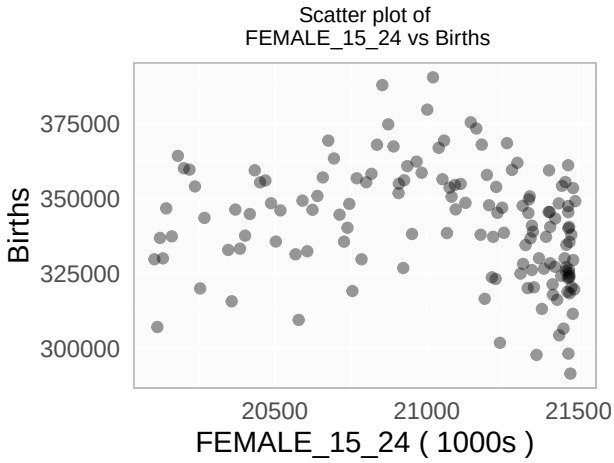
1.6 Variable FEMALE_15_24

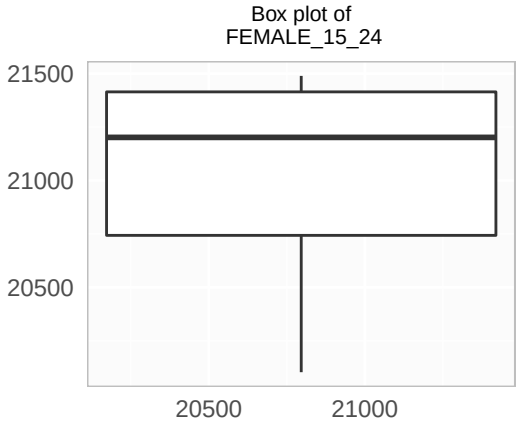
The *FEMALE_15_24* variable is the total population of females ages 15-24 per month as estimated by the Census Bureau.

Table 5: FEMALE_15_24 Variable Statistics

min	mean	stdev	median	max
20103.14	21046.7	422.1778	21201.43	21489.1





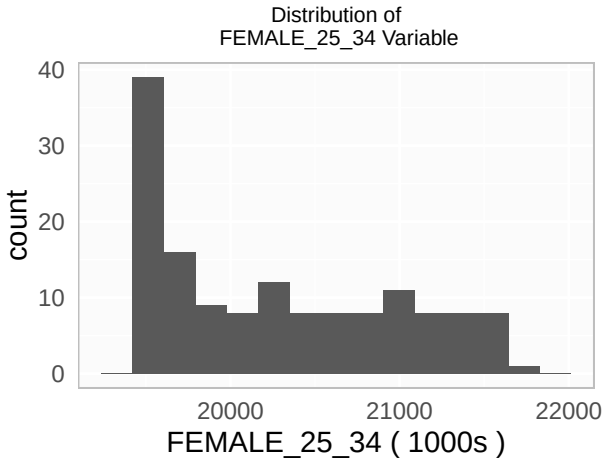


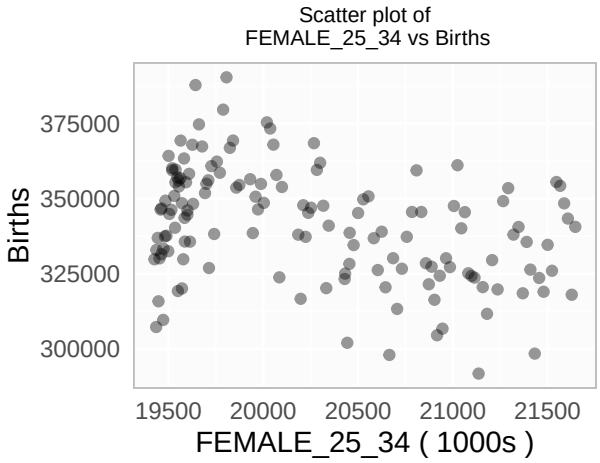
1.7 Variable FEMALE_25_34

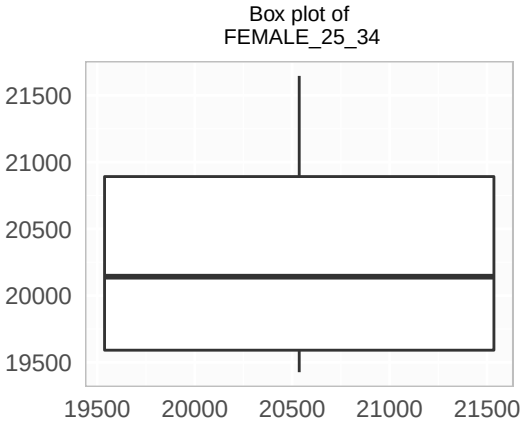
The *FEMALE_25_34* variable is the total population of females ages 25-34 per month as estimated by the Census Bureau.

Table 6: FEMALE_25_34 Variable Statistics

min	mean	stdev	median	max
19426.37	20274.31	701.1676	20141.73	21646.13





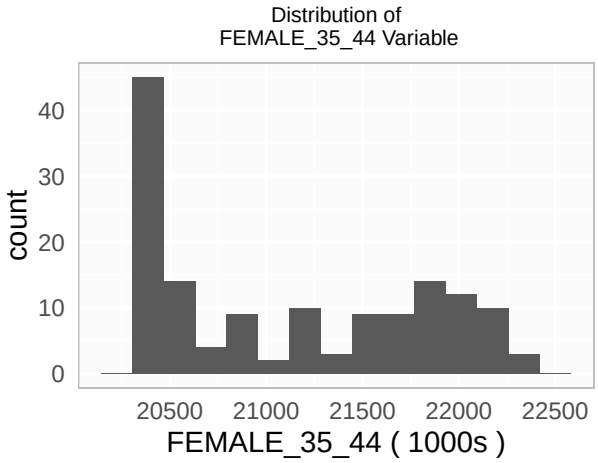


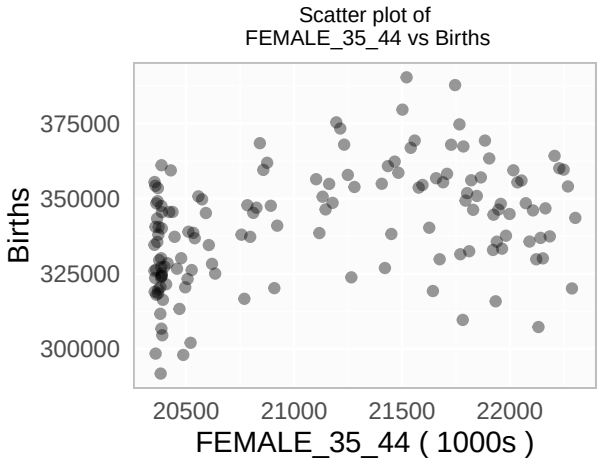
1.8 Variable FEMALE_35_44

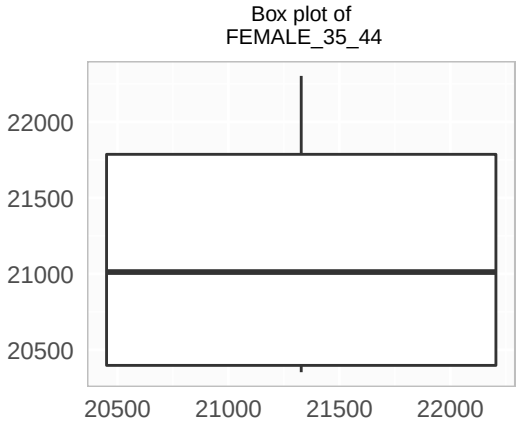
The *FEMALE_35_44* variable is the total population of females ages 35-44 per month as estimated by the Census Bureau.

Table 7: FEMALE_35_44 Variable Statistics

min	mean	stdev	median	max
20353.37	21120.04	683.5963	21012.17	22302.87





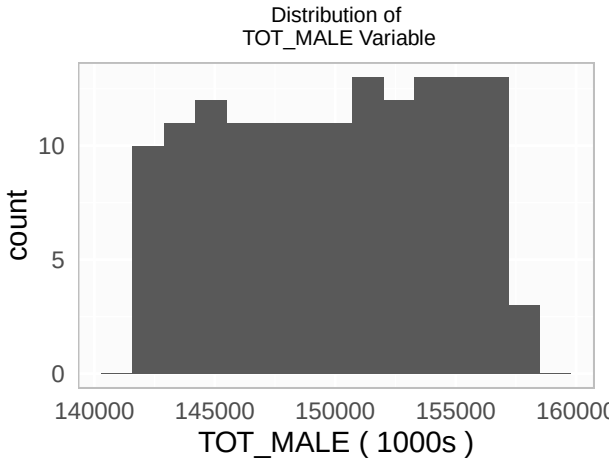


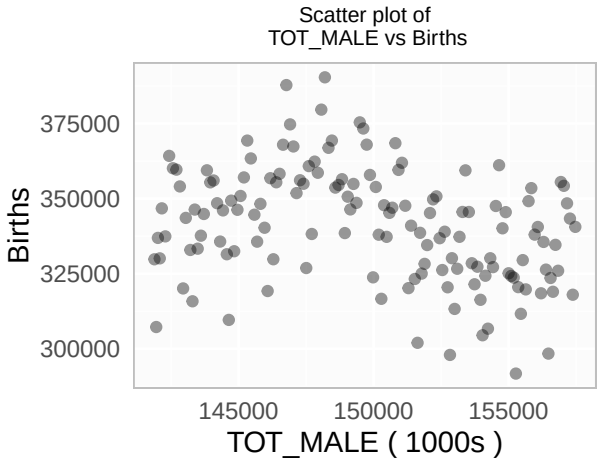
1.9 Variable TOT_MALE

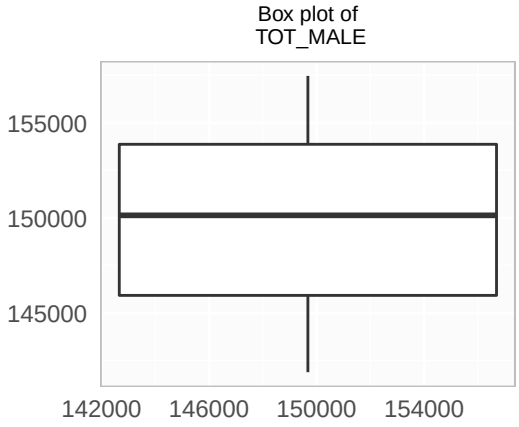
The *TOT_MALE* variable is the total population of females per month as esimated by the Census Bureau.

Table 8: TOT_MALE Variable Statistics

min	mean	stdev	median	max
141884.4	149888.3	4610.232	150137.2	157472.9





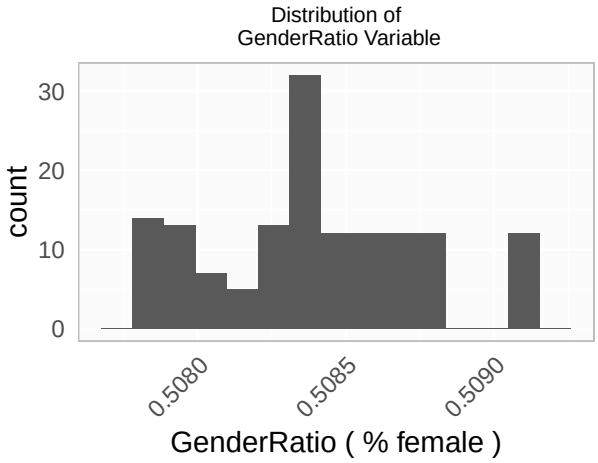


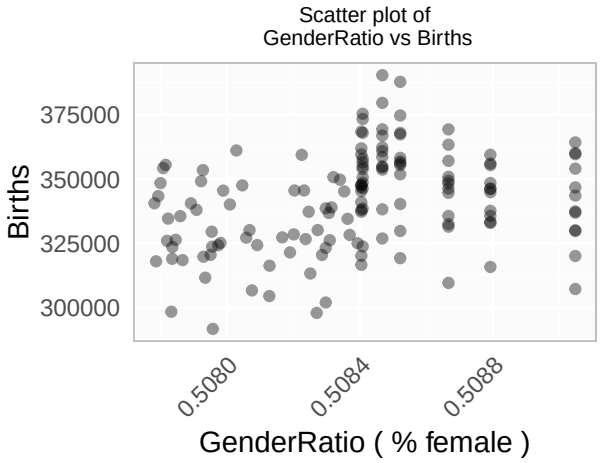
1.10 Variable GenderRatio

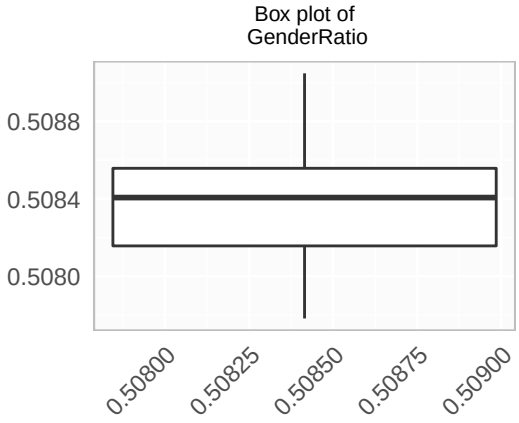
The *GenderRatio* variable is the percentage of the total population which are females per month derived from data from the Census Bureau. In cases where month data was not available, the annual gender ratio was computed and applied to the monthly total population.

Table 9: GenderRatio Variable Statistics

min	mean	stdev	median	max
0.507782	0.5083882	0.0003426	0.5084067	0.5090486





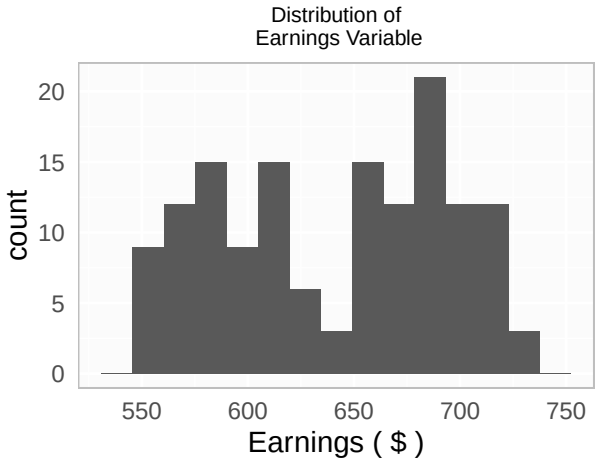


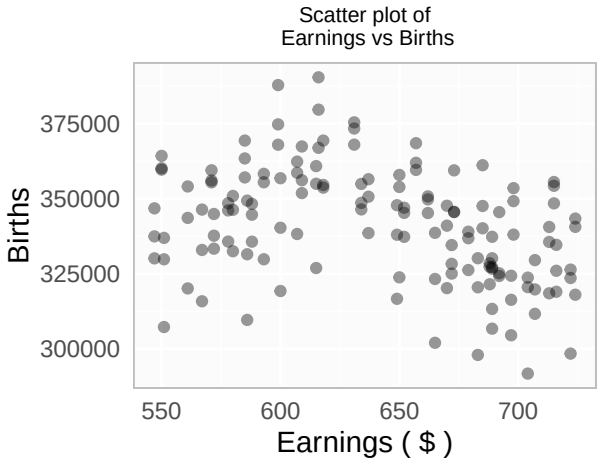
1.11 Variable Earnings

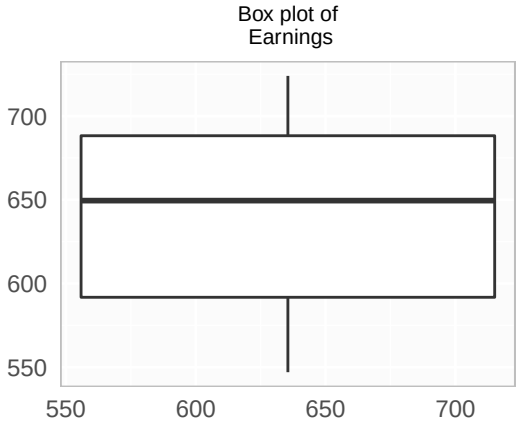
The *Earnings* variable is womoen's weekly earnings in current dollars based on data from the Bureau of Labor Statistics. The original values were provided quarterly and were expanded to a monthly format for data analysis purposes.

Table 10: Earnings Variable Statistics

min	mean	stdev	median	max
547	640.5417	53.55213	649.5	724





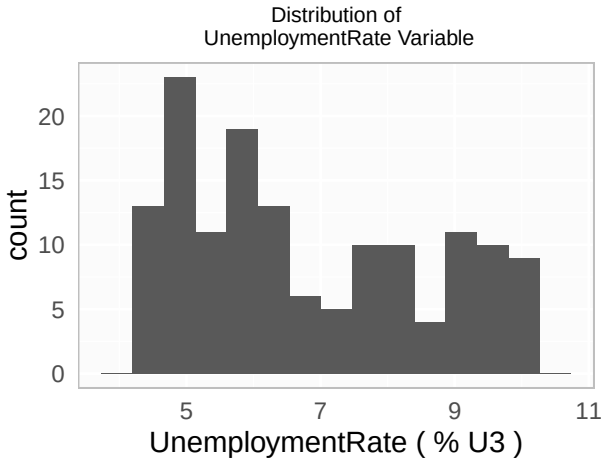


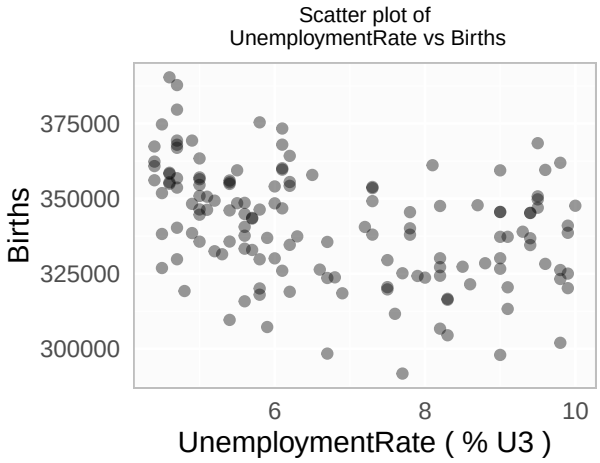
1.12 Variable UnemploymentRate

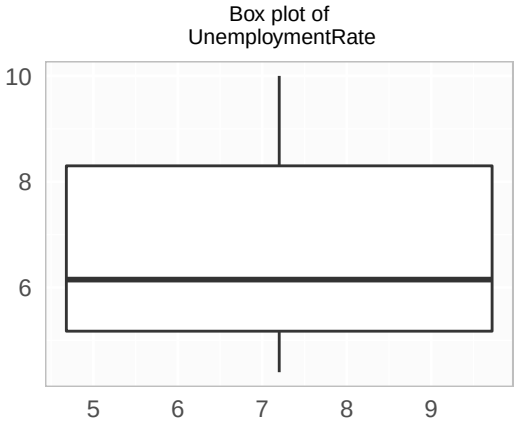
The *UnemploymentRate* variable is the unemployment rate per month (U3) based on data from the Bureau of Labor Statistics.

Table 11: UnemploymentRate Variable Statistics

min	mean	stdev	median	max
4.4	6.756944	1.789466	6.15	10







2 Build Models

2.1 All Variables Linear Model

The first multiple linear regression model uses all 10 predictor variables. The adjusted R^2 value for this model is 0.61129.

Table 12: All Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept	-107271160.660574	0.7809893
Month *	2217.241706	0.0006441
TOT_POP	311.447873	0.8070423
GenderRatio	210610661.121019	0.7815692
TOT_FEMALE	-595.803571	0.8124195
FEMALE_15_24	-30.890272	0.5719171
FEMALE_25_34	-74.057899	0.0513369
FEMALE_35_44	5.958233	0.7613287
Earnings	-84.575677	0.7558396
UnemploymentRate	-2839.165366	0.1719684
Month9Ago *	2600.816380	0.0000000

Table 13: All Variables Linear Model VIFs

Month	4.4144849
TOT_POP	120650287.7957136
GenderRatio	57975.4107139
TOT_FEMALE	115783897.4451604
TOT_MALE	55884.8603223
FEMALE_15_24	217.5425337
FEMALE_25_34	170.6400888
FEMALE_35_44	30258.7816115
Earnings	10734.8169565
UnemploymentRate	0.5494916
Month9Ago	4.3240384

2.2 Signficant Variables Linear Model

The second multiple linear regression model uses predictor variables indicated as significant from the All Variables model. The variables selected here were based on the All Variables model prior to the inclusion of the Month9Ago generated variable which appears to have affected the significant variables. The adjusted R^2 value for this model is 0.47714.

Table 14: Signficant Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept	514568208.68924	0.2046065
TOT_POP	-1765.33807	0.1855940
GenderRatio	-1018540784.31186	0.2017759
TOT_FEMALE	3501.25182	0.1822808
FEMALE_15_24	-34.77073	0.5600065
FEMALE_25_34	-15.46851	0.6855427
FEMALE_35_44 *	47.00020	0.0000874
Earnings *	-1277.29819	0.0000000

Table 15: 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

2.3 High Correlation Variables Linear Model

The third multiple linear regression model uses the six predictor variables with the highest correlation. The adjusted R^2 value for this model is 0.47889.

	Estimate	Pr(>|t|)

Table 16: High Correlation Variables Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	-1962175.32283	0.0203610
FEMALE_25_34 *	-47.74490	0.0000000
UnemploymentRate	713.91408	0.6637059
FEMALE_35_44	22.29315	0.2121244
Earnings *	-1182.65357	0.0000014
Month	988.80549	0.0908859
TOT_FEMALE *	22.88310	0.0000006

Table 17: High Correlation Variables Linear Model VIFs

FEMALE_25_34	20.572017
UnemploymentRate	6.361599
FEMALE_35_44	96.756161
Earnings	100.430108
Month	2.785238
TOT_FEMALE	257.428151

2.4 Step Linear Model

The *step* function was used to produce the next multiple linear regression model. The adjusted R^2 value for this model is 0.62001.

Table 18: Step Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	320506.481281	0.0000000
Month *	2531.412763	0.0000000
TOT_POP *	2.630363	0.0000005
FEMALE_25_34 *	-39.214446	0.0000000
UnemploymentRate *	-2955.157296	0.0000197
Month9Ago *	2645.246189	0.0000000

Table 19: Step Linear Model VIFs

Month	1.063277
TOT_POP	18.523696
FEMALE_25_34	17.820910
UnemploymentRate	1.427085
Month9Ago	1.028006

2.5 Significant Variables Minus Linear Model

The next model was aimed at removing variables with multicolinearity evidenced by the high VIFs we'd seen on earlier models. The adjusted R^2 value for this model is 0.36007.

Table 20: Significant Variables Minus Linear Model Coefficient Estimates

	Estimate	Pr(>|t|)
Intercept *	36913929.49606	0.0000048
Month *	2863.98730	0.0000000
GenderRatio *	-69493643.00948	0.0000069
FEMALE_25_34 *	-25.38478	0.0006337
FEMALE_35_44	-20.81506	0.0506143
Earnings *	-479.99715	0.0138361

Table 21: Significant Variables Minus Linear Model VIFs

Month	1.415159
GenderRatio	13.281871
FEMALE_25_34	14.099426
FEMALE_35_44	27.701955
Earnings	56.329990

2.6 Significant Variables Limited Linear Model

A manual review of features and the introduction of a 9 month lag variable brought us to the next model. The adjusted R^2 value for this model is 0.52589.

Table 22: 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 23: Significant Variables Limited Linear Model VIFs

Month	1.062619
Month9Ago	1.028000
FEMALE_25_34	1.400205
UnemploymentRate	1.360376

3 Select Models

A validation data set (VS) was created from a subset of the full dataset for use in the mulitple linear regression. This VS data set was used to perform a level of independent validation of the previously described models. The validation metric for the multiple linear regression models is the mean squared error from the validation set.

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

Table 24: Linear Model Validation Error Results

Model	VS Error	Adj R^2	AIC	Variables	VIF
All Variables	161072343	0.6112935	2504.690	11	BAD
Neg Bionomial	161290241	NA	2506.956	10	NA
Poisson Step	161316049	NA	40569.475	10	NA
Step	172024296	0.6200088	2497.456	5	BAD
Poisson Signif Ltd	176055186	NA	51551.445	4	NA
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

Based on the criteria of least complex model with lowest validation error, highest R^2 and no multicollinearity issues, the Significant Limited model is favored for further investigation.

3.1 Evaluation: Significant Limited Linear Model

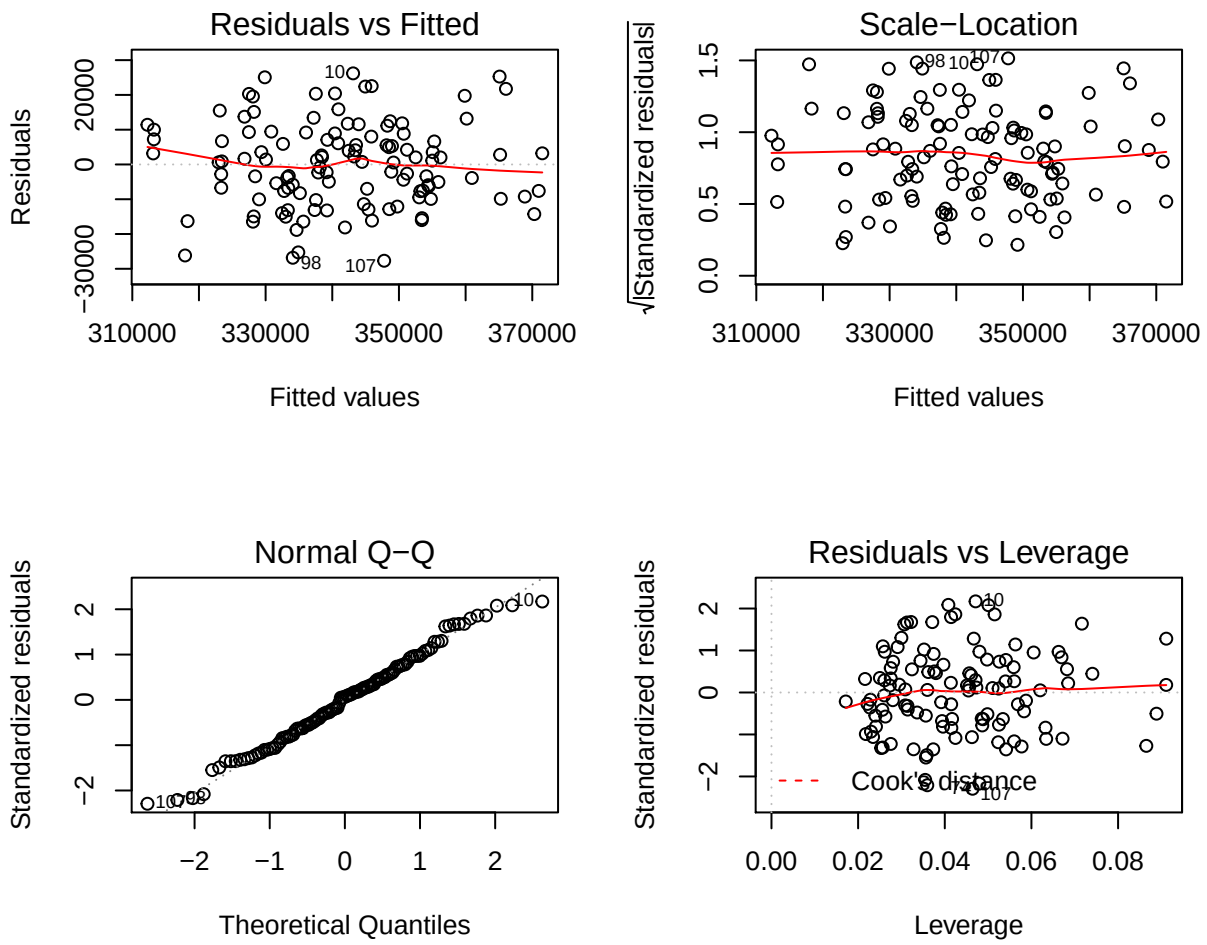
The Significant Limited model has an F-statistic of 32.89 and a mean squared error (MSE) of 146358133.79.

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

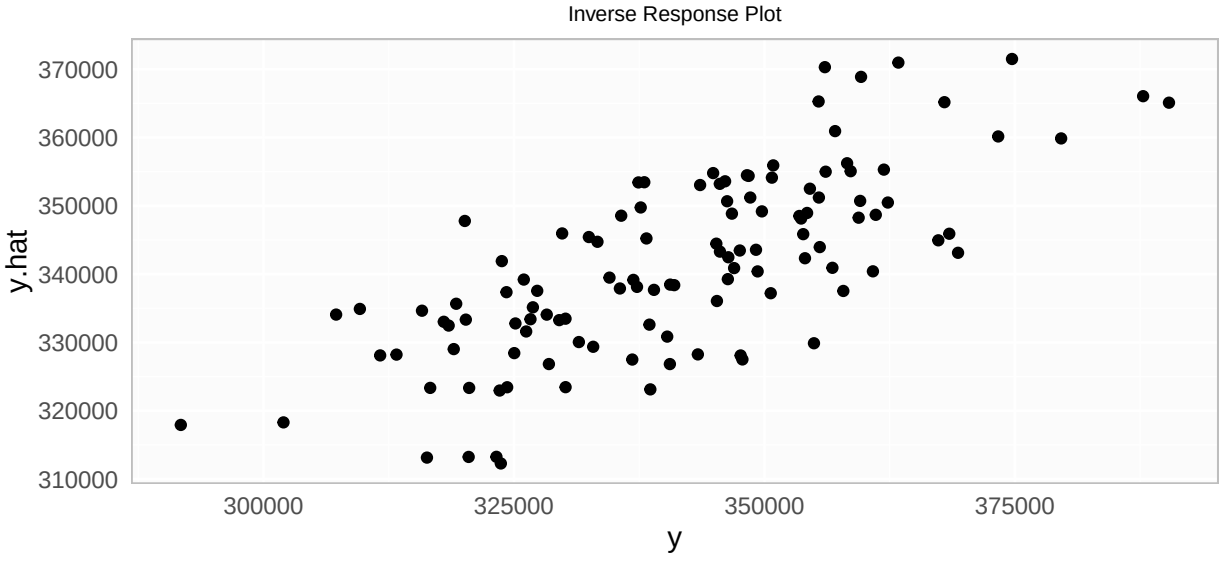
We can interpret the coefficients in the following manner. Holding all other predictors constant, for variable:

- *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.

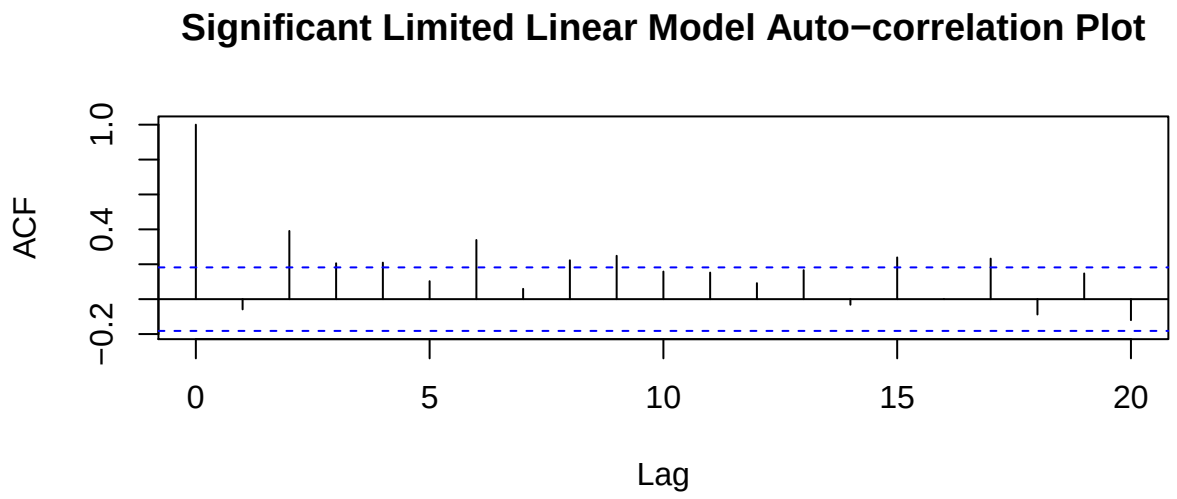
Linear regression diagnostic plots are shown below. Residuals appear to be normally distributed and variance seems to be fairly constant. There are definitely some outliers which the model is not capturing.



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



Running a more targeted auto-correlation analysis with R's *acf* function shows a possible pattern of negative and positive residual oscilation and imparticular at the lag of 12 .


Significant Limited Linear Model Auto−correlation Plot
ACF
1.0
0.4
−0.2
0
5
10
15
20
Lag


Signif Limited Model vs Validation Set
model
360000
340000
320000
300000
2004
2006
2008
2010
2012
2014
Date