Impact of State Policies on New Covid19 Cases

By Sam Sarmad, Andrew Zhou, and Nalani Chun

Summary

We set out to find out the impact of governmental policies on rates of COVID-19 infections. After reviewing available data and their quality we decided to narrow our focus on non pharmaceutical interventions (NPI) and the US population.

For our study we choose 2 data sets:

- NY Times historical cases and deaths
- Imperial college Non-Pharmaceutical Interventions by state effective date

Upon initial statistical analysis and variability between states we realize that policy measures should be analyzed on a per state basis.

Due to the amount of time available, we decided to analyze the top 3 states with most cases in which also provides more samples: California, New York, and Pennsylvania. Our final process, as described below, was applied to all states. The data files and outputs for all states are available on our Github repository for review. We only include the process and outputs for California in this summary. California summary can be found at https://github.com/NalaniKai/Covid19-Hackathon/blob/master/CA/CA_visuals_summary.pdf. New York summary can be found at https://github.com/NalaniKai/Covid19-Hackathon/blob/master/NY/NY_visuals_summary.pdf. Pennsylvania summary can be found at https://github.com/NalaniKai/Covid19-Hackathon/blob/master/PA/PA_visuals_summary.pdf.

We started by performing regression analysis on case increases per day using all features. After eliminating the features which did not statistically contribute to the model, the ones with high covariance and multicollinearity and high VIFs we arrived at a final initial regression model.

In our analysis general analysis, some conclusions did not make sense and factors that seem to be late interventions such as wearing masks, were showing heavier influence than social distancing. This is due to the fact that when the data is looked at in totality, features are matched with all effects regardless of time series impact.

We proceeded to run a time series analysis with distancing and ARIMA to remove lag effects and get a more accurate match for feature influence. As a result, we were able to generate a

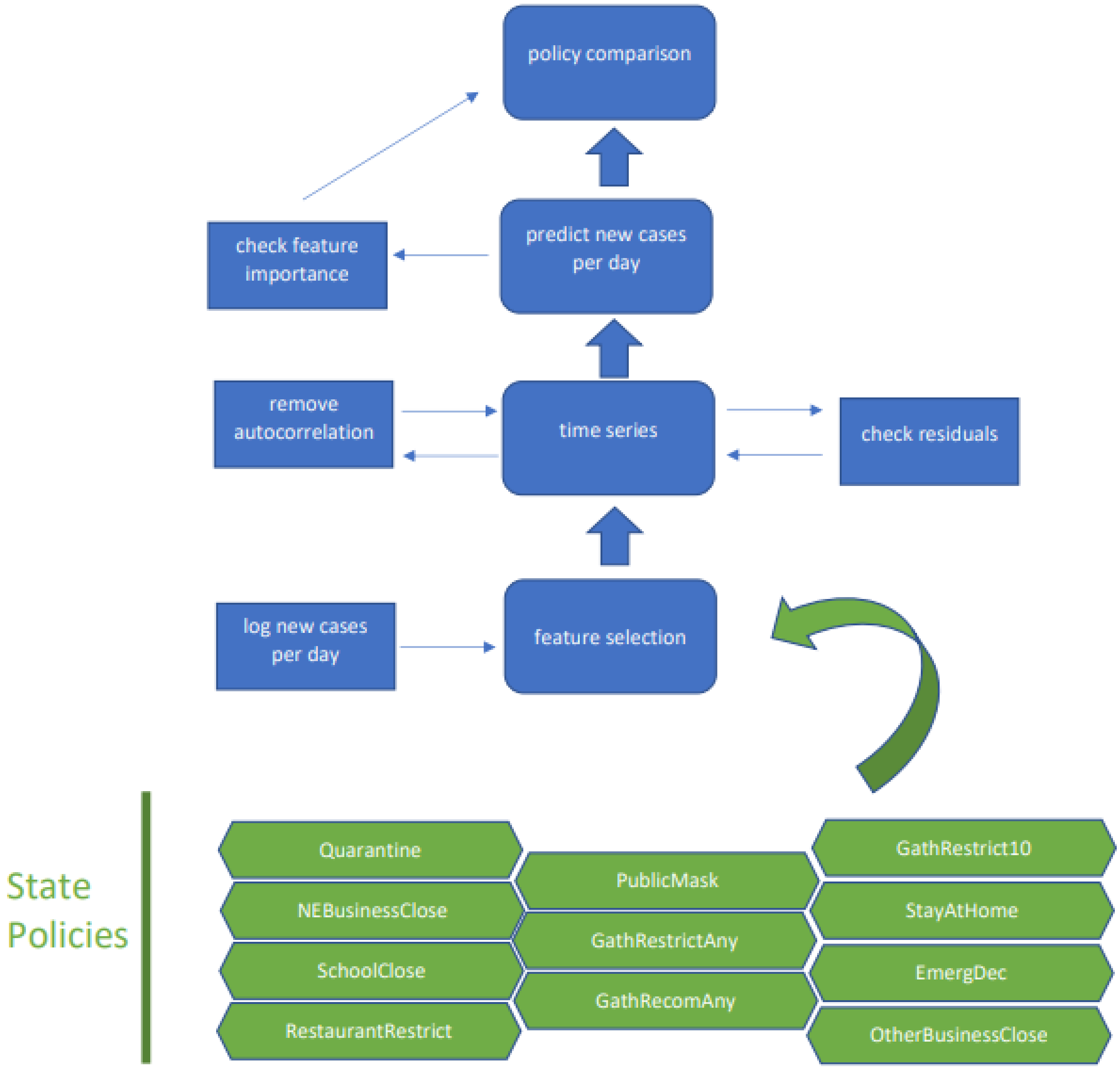
new predicted outcome variable model that we rerun in a multi regression model. After eliminating features achieved a result that was more accurate and made more sense.

For example, in California the first multi regression run eliminated the Declaration of Emergency as a significant factor. After our adjustments for the time series effects emergency declaration shows as the most prominent feature. Looking at the day the emergency declaration was put in place, March 4, and day graphs of data, it makes much more sense that it should have had a higher impact through social distancing and reducing contacts than a later effect which was initially classified as significant.

The results could be used to compare policies across states.

https://github.com/NalaniKai/Covid19-Hackathon/

Process



State policies

We extracted the state policies from the Imperial College dataset. Each row corresponded to one day and one state with each state policy as a feature column filled with a binary value for whether the policy was active.

Log new cases per day

We extracted the number of new cases per day from the NY Times dataset and added a column for the log of the new cases. Then combined this dataset with the modified Imperial College dataset.

Feature selection

To determine which policies were the most important for predicting the log of new cases per day, we used standard least squares regression and analyzed the log probabilities, p-values, and VIF for each policy. Then we iteratively removed biased and non-contributing policies until we reached a group of policies that were statistically significant.

Time series

We modeled the log of new cases using autoregression and compared models using the differences in AIC.

Remove autocorrelation

To choose the lag we used the autocorrelation plot, p-values, and AIC.

Check residuals

We check residual plots and confidence intervals to confirm model accuracy.

Predict new cases per day

We used the ARIMA adjusted predicted cases per day to rerun our regression analysis which can both describe and predict new cases per day using time series models with trained features.

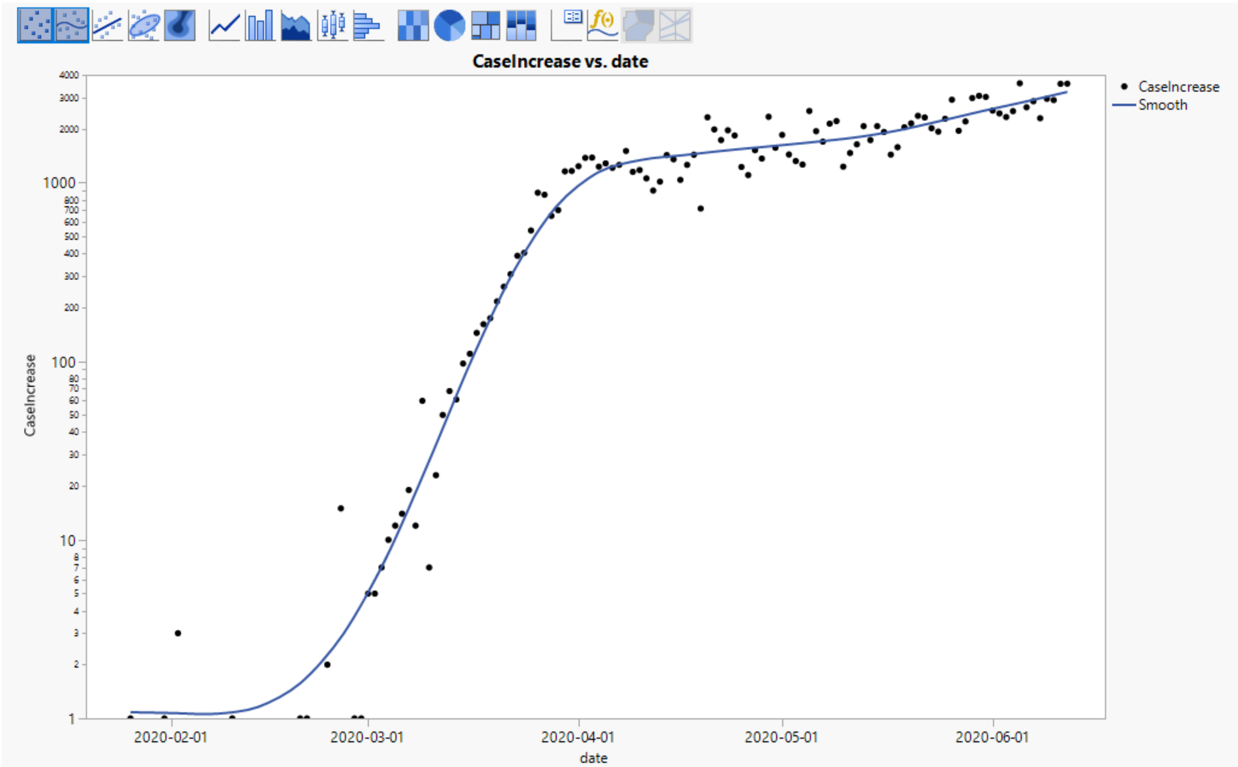
Check feature importance

After predicting the log of new cases per day, we used standard least squares regression similar to the feature selection step above.

Policy comparison

This model can be used to compare the policy significance across states in a time series.

Data Fit



Scatter Plot of Log(Case Increase) vs Date for California

Feature Selection

Response Log[CaseIncrease]

Singularity Details

Term	Details
GathRecomAny	= 0
GathRestrict10	= 0
Quarantine	= 0
OtherBusinessClose	=RestaurantRestrict
NEBusinessClose	=StayAtHome

Effect Summary

Source	LogWorth	PValue
PublicMask	5.439	0.00000
EmergDec	3.285	0.00052
GathRestrictAny	0.630	0.23441
SchoolClose	0.490	0.32382
StayAtHome	.	.
RestaurantRestrict	.	.
Quarantine	.	.
OtherBusinessClose	.	.
NEBusinessClose	.	.
GathRestrict10	.	.
GathRecomAny	.	.

Summary of Fit

RSquare	0.87826
RSquare Adj	0.871433
Root Mean Square Error	0.859742
Mean of Response	6.057394
Observations (or Sum Wgts)	114

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	570.56964	95.0949	128.6534
Error	107	79.08968	0.7392	Prob > F
C. Total	113	649.65931		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept		1.2747568	0.229776	5.55	<.0001*	.
EmergDec		1.4246931	0.397983	3.58	0.0005*	2.6315789
GathRecomAny	Zeroed	0	0	.	.	0
GathRestrict10	Zeroed	0	0	.	.	0
GathRestrictAny		0.8243087	0.689327	1.20	0.2344	11.013158
NEBusinessClose	Biased	1.2940566	0.550675	2.35	0.0206*	8.8708952
OtherBusinessClose	Biased	0.6155352	0.701976	0.88	0.3825	13.380117
PublicMask		1.2657812	0.259083	4.89	<.0001*	2.4089069
Quarantine	Zeroed	0	0	.	.	0
RestaurantRestrict	Zeroed	0	0	.	.	0

SchoolClose		0.7779389	0.784833	0.99	0.3238	15.299708
StayAtHome	Zeroed	0	0	.	.	0

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
EmergDec	1	1	9.472168	12.8148	0.0005*	
GathRecomAny	1	0	0.000000	.	.	LostDFs
GathRestrict10	1	0	0.000000	.	.	LostDFs
GathRestrictAny	1	1	1.056976	1.4300	0.2344	
NEBusinessClose	1	0	0.000000	.	.	LostDFs
OtherBusinessClose	1	0	0.000000	.	.	LostDFs
PublicMask	1	1	17.643072	23.8692	<.0001*	
Quarantine	1	0	0.000000	.	.	LostDFs
RestaurantRestrict	1	0	0.000000	.	.	LostDFs
SchoolClose	1	1	0.726227	0.9825	0.3238	
StayAtHome	1	0	0.000000	.	.	LostDFs

Initial Regression Model with all Features

Response Log[CaseIncrease]

Effect Summary

Source	LogWorth	PValue
PublicMask	5.402	0.00000
StayAtHome	5.336	0.00000
GathRestrictAny	3.413	0.00039
EmergDec	3.260	0.00055

Summary of Fit

RSquare	0.874663
RSquare Adj	0.870064
Root Mean Square Error	0.864308
Mean of Response	6.057394
Observations (or Sum Wgts)	114

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	568.23317	142.058	190.1644
Error	109	81.42614	0.747	Prob > F
C. Total	113	649.65931		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.2747568	0.230996	5.52	<.0001*
EmergDec	1.4246931	0.400097	3.56	0.0005*
GathRestrictAny	1.6385886	0.447322	3.66	0.0004*
PublicMask	1.2657812	0.26046	4.86	<.0001*
StayAtHome	1.8732508	0.388384	4.82	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
EmergDec	1	1	9.472168	12.6798	0.0005*
GathRestrictAny	1	1	10.023897	13.4184	0.0004*
PublicMask	1	1	17.643072	23.6177	<.0001*
StayAtHome	1	1	17.378243	23.2632	<.0001*

Correlation of Estimates

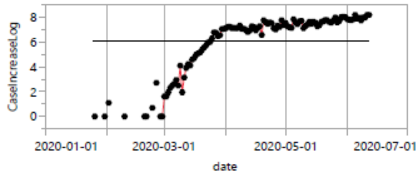
Corr

	Intercept	EmergDec	GathRestrictAny	PublicMask	StayAtHome
Intercept	1.0000	-0.5774	0.0000	0.0000	-0.0000
EmergDec	-0.5774	1.0000	-0.5963	-0.0000	0.0000
GathRestrictAny	0.0000	-0.5963	1.0000	0.0000	-0.5375
PublicMask	0.0000	-0.0000	0.0000	1.0000	-0.5681
StayAtHome	-0.0000	0.0000	-0.5375	-0.5681	1.0000

Final Regression model with significant features

Time Series

Time Series CaseIncreaseLog



Mean	6.0573937
Std	2.3872088
N	114
Zero Mean ADF	0.8209779
Single Mean ADF	-3.471114
Trend ADF	-3.225744

Time Series Basic Diagnostics

Lag	AutoCorr	Ljung-Box Q	p-Value
0	1.0000	.	.
1	0.6284	46.2067	<.0001*
2	0.6418	94.8349	<.0001*
3	0.5946	136.961	<.0001*
4	0.6113	181.890	<.0001*
5	0.6245	229.208	<.0001*
6	0.5334	264.047	<.0001*
7	0.5806	305.706	<.0001*
8	0.6111	352.301	<.0001*
9	0.5223	386.660	<.0001*
10	0.5929	431.353	<.0001*
11	0.5213	466.248	<.0001*
12	0.4121	488.265	<.0001*
13	0.3956	508.758	<.0001*
14	0.4041	530.356	<.0001*
15	0.4003	551.758	<.0001*
16	0.3426	567.594	<.0001*
17	0.2785	578.167	<.0001*
18	0.3636	596.377	<.0001*
19	0.3562	614.034	<.0001*
20	0.3124	627.762	<.0001*
21	0.2880	639.554	<.0001*
22	0.2486	648.440	<.0001*
23	0.1839	653.354	<.0001*
24	0.2376	661.653	<.0001*
25	0.2001	667.599	<.0001*

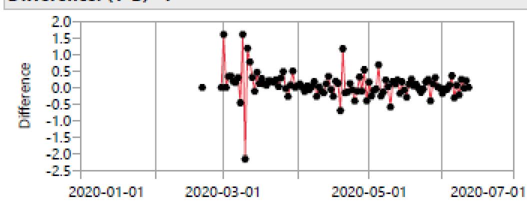
Lag	Partial
0	1.0000
1	0.6284
2	0.4080
3	0.1973
4	0.2048
5	0.2063
6	-0.0435
7	0.1149
8	0.2034
9	-0.0916
10	0.1464
11	0.0063
12	-0.3459
13	-0.1548
14	0.0595
15	-0.1197
16	-0.0727
17	-0.0559
18	0.0361
19	0.1675
20	0.1187
21	0.0235
22	0.0575
23	-0.1343
24	0.0779
25	0.0402

Time Series CaseIncreaseLog

Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights
☑	☐	ARI(1, 1)	104	0.1441364	97.671952	102.99883	0.976	93.671952	0.934554
☑	☐	ARI(20, 1)	85	0.1270273	104.14000	160.07222	0.981	62.139996	0.036820
☑	☐	ARI(14, 1)	91	0.1368762	104.64346	144.59505	0.979	74.643464	0.028626

Difference: (1-B)^1



Mean	0.0771887
Std	0.4159142
N	106
Zero Mean ADF	-15.12105
Single Mean ADF	-15.94364
Trend ADF	-17.01925

Time Series CaseIncreaseLog

Model: ARI(1, 1)

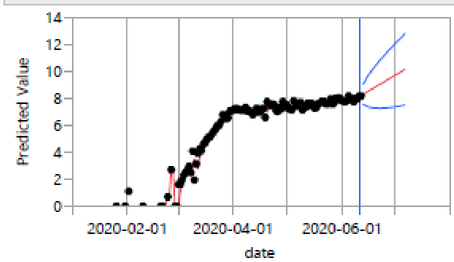
Model Summary

DF	104	Stable	Yes
Sum of Squared Errors	14.990183	Invertible	Yes
Variance Estimate	0.14413638		
Standard Deviation	0.37965297		
Akaike's 'A' Information Criterion	97.6719516		
Schwarz's Bayesian Criterion	102.99883		
RSquare	0.97553015		
RSquare Adj	0.9753097		
MAPE	.		
MAE	0.23716256		
-2LogLikelihood	93.6719516		

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
AR1	1	-0.4234418	0.0872822	-4.85	<.0001*	0.11102155	0.07799514
Intercept	0	0.0779951	0.0257955	3.02	0.0031*		

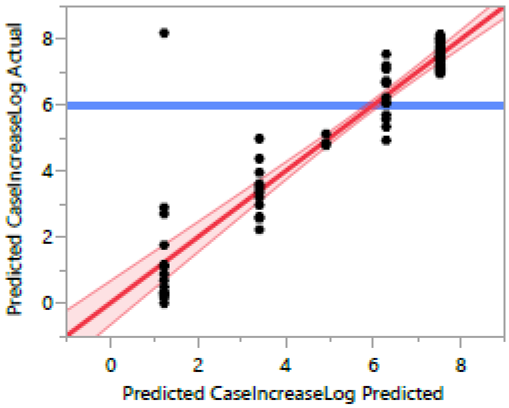
Forecast



Final Model

Response Predicted CaseIncreaseLog

Actual by Predicted Plot

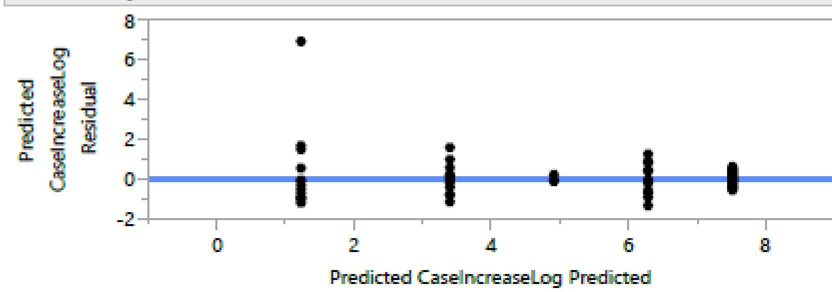


RMSE=0.8712 RSq=0.88 PValue=<.0001

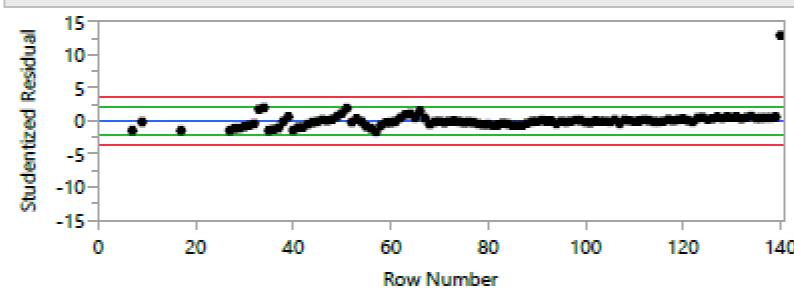
Effect Summary

Source	LogWorth	PValue
EmergDec	8.843	0.00000
PublicMask	5.097	0.00001
RestaurantRestrict	2.088	0.00816
StayAtHome	1.809	0.01553

Residual by Predicted Plot



Studentized Residuals



Response Predicted CaseIncreaseLog

Studentized Residuals

Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

AICc	BIC
307.4114	323.2209

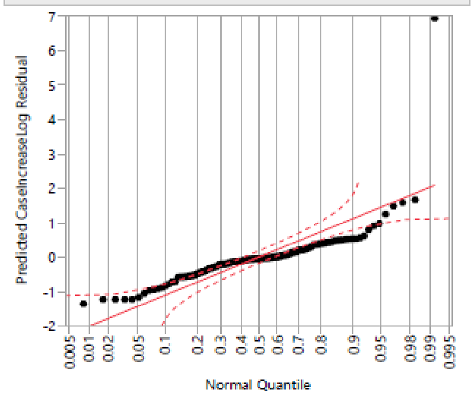
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.2323953	0.211289	5.83	<.0001*
EmergDec	2.1673596	0.328462	6.60	<.0001*
PublicMask	1.2294031	0.262526	4.68	<.0001*
RestaurantRestrict	1.5146409	0.562335	2.69	0.0082*
StayAtHome	1.3712179	0.557993	2.46	0.0155*

Effect Tests

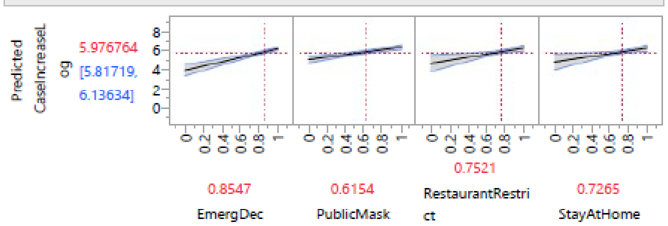
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
EmergDec	1	1	33.044115	43.5404	<.0001*
PublicMask	1	1	16.643535	21.9302	<.0001*
RestaurantRestrict	1	1	5.505929	7.2548	0.0082*
StayAtHome	1	1	4.583082	6.0389	0.0155*

Residual Normal Quantile Plot



Response Predicted CaseIncreaseLog

Prediction Profiler



Final Arima Model Results

.