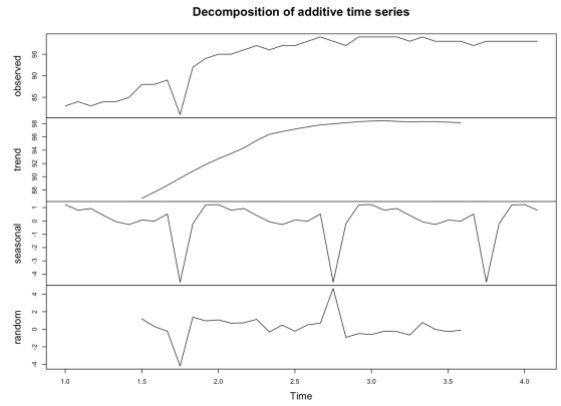
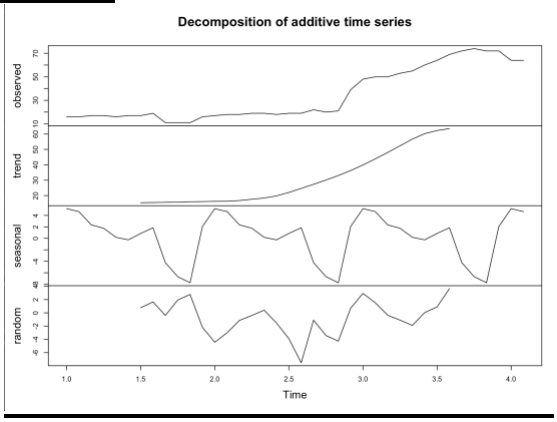
Introductory/Descriptive Reports

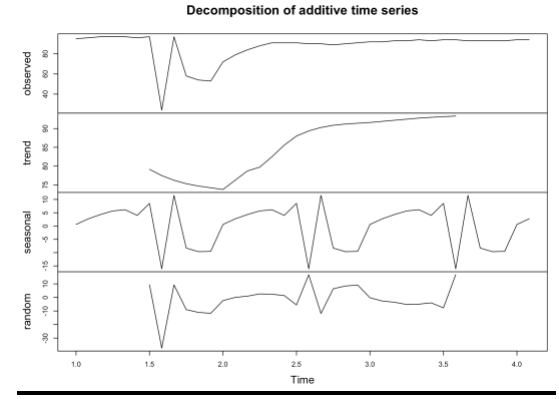
DTP1



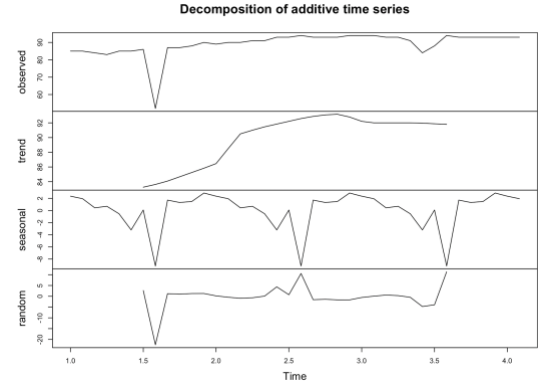
HepB_BD



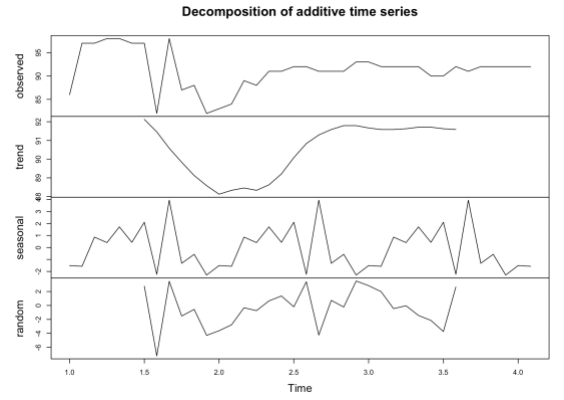
Pol3



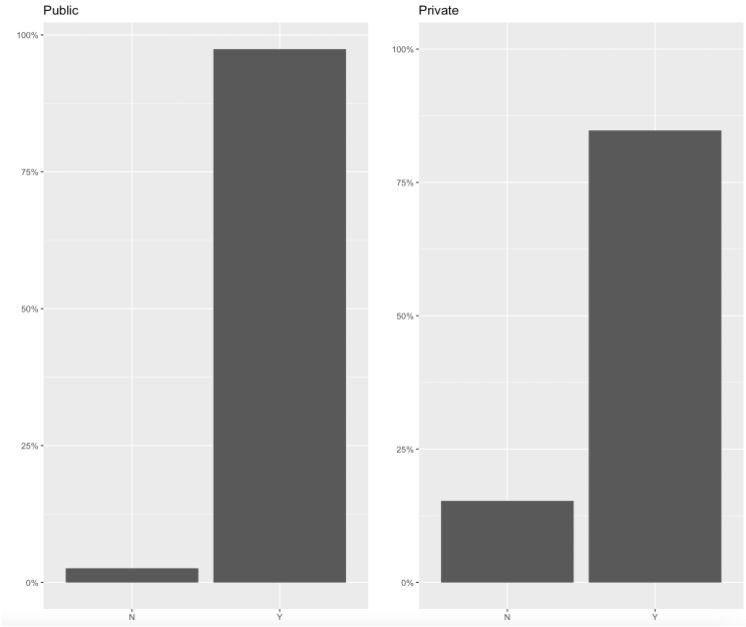
Hib3



MCV1



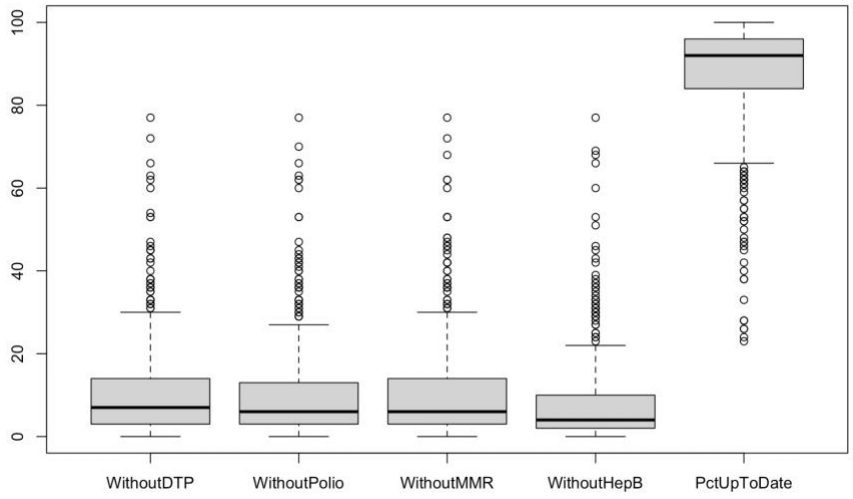
Vaccination rates have varied greatly between 1980 and 1991. After 1991 however, there was a steady positive trend up to around 90% for DTP1, Pol3, Hib3, and MCV1, and hovering around 65% for HepB_BD. DTP1 has the highest vaccination rate at the conclusion of this time series at ~98% while HepB_BD has the lowest at ~64%. MCV1 was noted to have the greatest variability as noted in the graph above.



pubpriv	reported	n	freq
<chr>	<chr>	<int>	<dbl>
1 PRIVATE	N	252	0.153
2 PRIVATE	Y	1397	0.847
3 PUBLIC	N	148	0.0258
4 PUBLIC	Y	5584	0.974

```
        2-sample test for equality of proportions with continuity correction

data:  c(5584, 1397) out of c(5584 + 148, 1397 + 252)
X-squared = 400.49, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1087641 0.1452357
sample estimates:
   prop 1    prop 2
0.9741800 0.8471801
```

Public Schools had a 97.4% report rate, while Private Schools had an 84.7% report rate. We ran a t-test that constructed a 95% confidence interval around the mean proportion difference of 12.7% between Public and Private School reporting rates, which ranged from 10.9% to 14.5%. It should be noted that this confidence interval may or may not contain the true population proportion difference. In this case, the p-value is less than alpha level .05, so results are statistically significant and favor the alternative hypothesis, which suggests there's a credible difference between rates of public and private school reporting.



The mean vaccination rates across 700 California schools in 2013 were:

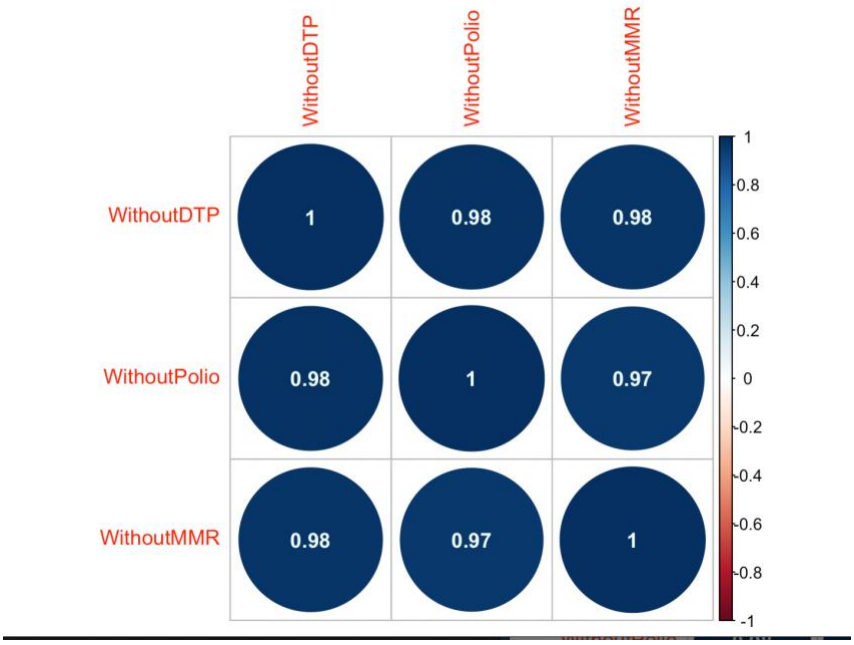
- *DOT: 90.08%*
- *Polio: 90.5%,*
- *MMR: 90.05%*
- *HepB: 92.49%*

Based on the boxplot analysis above, it seems that there is a lot of variability among the vaccination rates. However, outliers can skew the mean rates among the different vaccinations. The long upper bound whiskers indicate that the data is skewed towards higher values and has more uncertainty around the mean estimates.

Overall US rates.

- *DOT: 98.0%*
- *Polio: 93.0%,*
- *MMR: 92.0%*
- *HepB: 74%*

California schools were 8%, 3% and 2% below the national average of vaccination rates for DOT, Polio and MMR respectively. On the other hand, California schools were 18% above the national average for rates of vaccination for HepB.



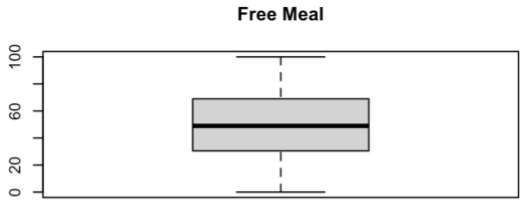
There are very strong positive correlations between individual unvaccinated rates. This indicates that if a student is missing one vaccine, there is a strong probability they are missing others.

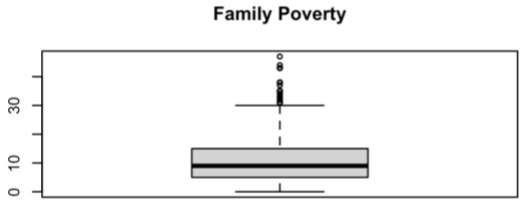
Predictive Analyses

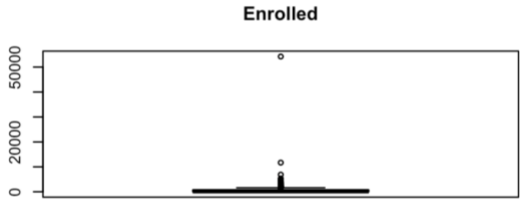
Predictor Variables Analysis:

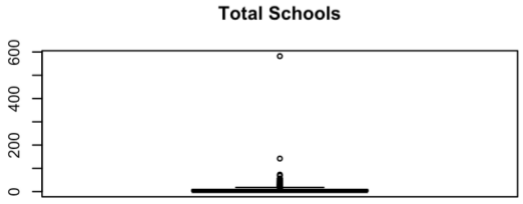
Outliers – outliers were detected in all but the percentage of students eligible for free meals, with extreme outliers being noted for number of students enrolled and total schools in district. The presence of outliers can dramatically change the magnitude and direction of coefficient signs, which impacts the accuracy of the model. Logistic regression models are more robust to outliers since the sigmoid function reduces any skewness by pulling in larger values to the center, making it appear more normal. However, extreme values can still affect the accuracy of the model. The number of outliers is determined by a data point below the 25th percentile or above the 75th percentile by a factor of 1.5*IQR. Reviewing the outliers, it was noted that Los Angeles Unified and San Diego Unified are the heavy

influencers for Enrolled and Total Schools. Family Poverty outliers consisted of a range between 31%-47% for 23 districts. Child Poverty outliers consisted of a range between 54%-72% for 13 districts. For both Child and Family poverty, Earlimart Elementary had the highest rates. There is a lot of gray area when dealing with outliers. I have chosen to include them as they truly represent the population being analyzed.









Multi-collinearity - An analysis of PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools variables shows high positive correlations, which violates the multi-collinearity assumption as indicated by the below correlation matrix and VIF analysis. For VIF analysis, anything above 2.5 indicates a high correlation. Multi-collinearity reduces the accuracy of the estimated coefficients and thus impacts the t-ratio leading to higher p-values.

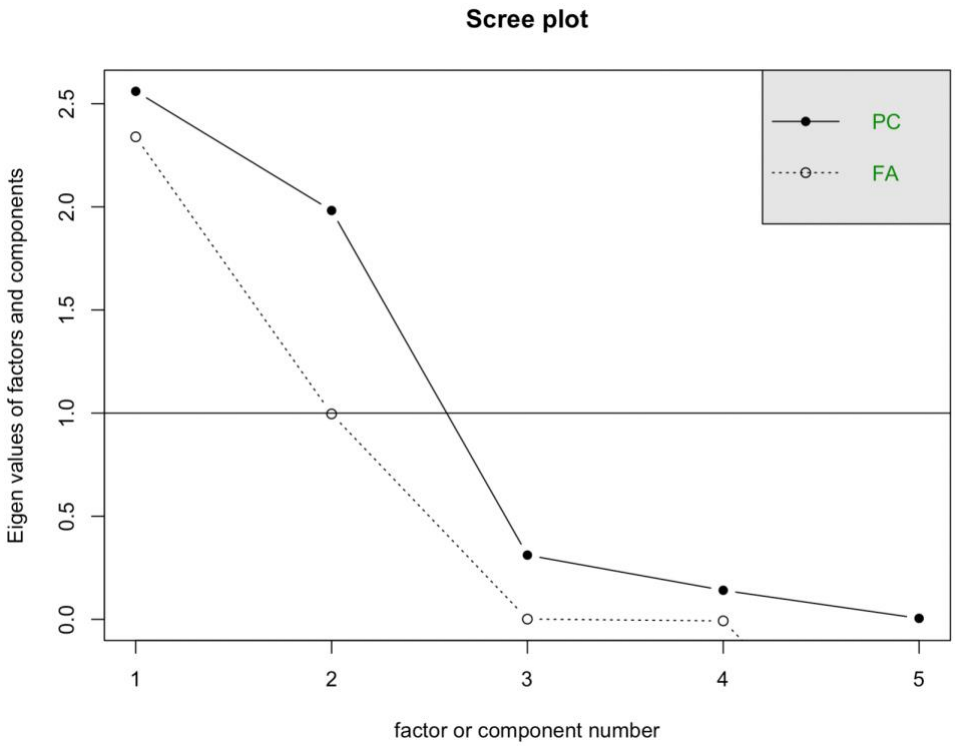
VIF Multi-Collinearity Test Results:

```
PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled     TotalSchools
       4.733440         2.174555         4.258903       105.155548       105.159284
```

A factor analysis was conducted to identify the relationship among the predictor variables to narrow them down to a smaller number influencing factors. We first determine the factorability using the Kaiser-Meyer-Olkin (KMO) measure to see how suited the variables are for factoring. Values less than 0.6 indicate that the sampling is not adequate and requires justification. Both Enrolled and TotalSchools have values less than 0.6. However, these two variables are essentially measuring the same thing. More students that enroll would likely lead to more total schools. Considering this, I believe it is reasonable to factor these two variables together.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = myDdistricts[, 9:13])
Overall MSA =  0.63
MSA for each item =
 PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled     TotalSchools
            0.68             0.85             0.70             0.50             0.50
>
```

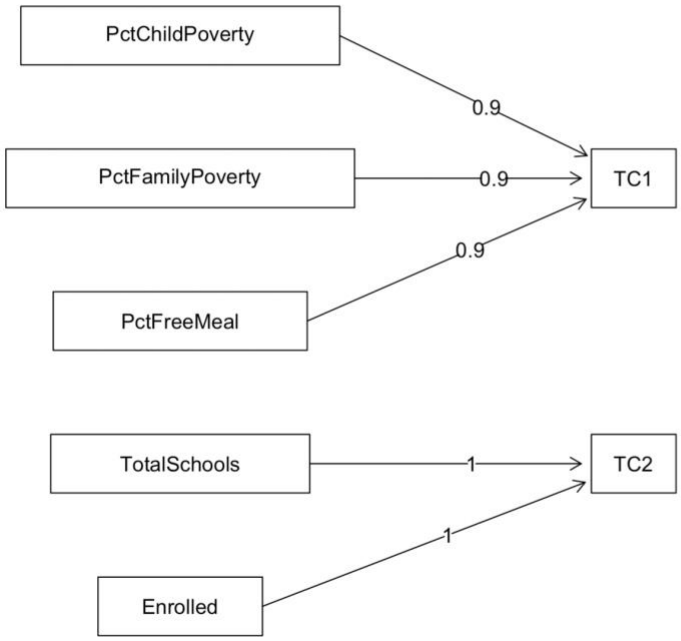
Next, using a scree plot with eigenvalues on the y-axis and number of factors on x-axis, we need to retain two factors. *The Kaiser* criterion recommends that the number of principal components or factors is the number of dots above the 1.0 line.



Two methods for combining variables into factors are **Factor Analysis** and **Principal Component Analysis (PCA).** PCA is very similar to Factor Analysis, and the two procedures are sometimes confused. Both procedures are built on the same mathematical techniques. They both identify a set of underlying factors that explain the relationships between correlated variables.

PCA

Components Analysis



Factor loadings (values on lines above) represent the strength and directionality of the relationship between each item and the underlying factor, and they can range from -1 to 1.

```
Principal Components Analysis
Call: principal(r = myDistricts[, 9:13], nfactors = 2, rotate = "oblimin")
Standardized loadings (pattern matrix) based upon correlation matrix
                  TC1   TC2   h2     u2 com
PctChildPoverty  0.94 -0.02 0.89 0.1087   1
PctFreeMeal      0.89  0.03 0.79 0.2143   1
PctFamilyPoverty 0.93  0.00 0.87 0.1294   1
Enrolled         0.00  1.00 1.00 0.0025   1
TotalSchools     0.00  1.00 1.00 0.0026   1

                       TC1  TC2
SS loadings           2.55 2.00
Proportion Var        0.51 0.40
Cumulative Var        0.51 0.91
Proportion Explained  0.56 0.44
Cumulative Proportion 0.56 1.00
```

Factor Analysis

```
Call:
factanal(x = myDistricts[, 9:13], factors = 2)

Uniquenesses:
 PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled     TotalSchools
           0.108            0.377            0.177            0.005            0.005

Loadings:
                Factor1 Factor2
PctChildPoverty   0.944
PctFreeMeal       0.787
PctFamilyPoverty  0.907
Enrolled                  0.998
TotalSchools              0.998

               Factor1 Factor2
SS loadings      2.334   1.995
Proportion Var   0.467   0.399
Cumulative Var   0.467   0.866

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.6 on 1 degree of freedom.
The p-value is 0.206
```

Using Factor Analysis, we get a summary output called “Uniqueness”, which is the variance that is unique to the variable and not shared with other variables. The null hypothesis for test states that two factors are sufficient. With a p-value >0.05, we accept the null that two factors are sufficient.

Factors	***Variables***	***Short Interpretation***
Poverty	PctChildPoverty, PctFamilyPoverty, PctFreeMeal	All variables related to percentage of children and families living below the poverty line and percentage eligible for free meals
Students	Enrolled, TotalSchools	All variables related to enrolled students and number of schools in district

1. What variables predict whether a district's reporting was complete?

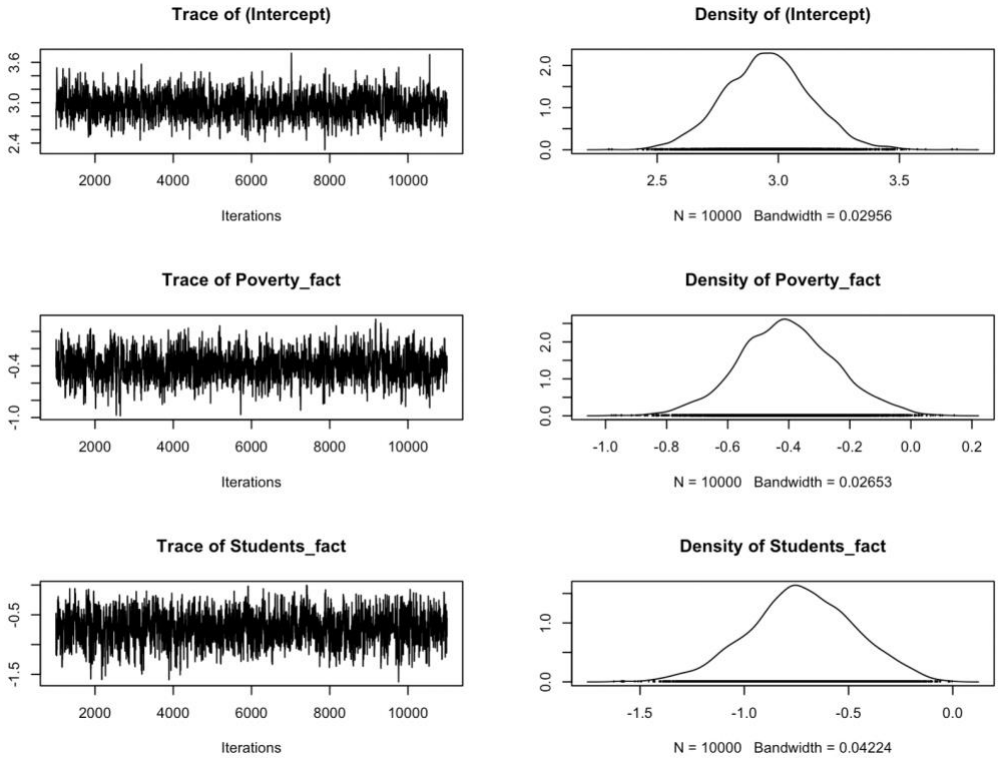
```
Call:
glm(formula = DistrictComplete ~ Poverty_fact + Students_fact,
    family = binomial(), data = myDistricts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6465   0.2568   0.2997   0.3582   1.5100

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.9313     0.1798  16.299  < 2e-16 ***
Poverty_fact   -0.4135     0.1575  -2.625  0.00867 **
Students_fact  -0.6900     0.2555  -2.701  0.00691 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 306.65  on 699  degrees of freedom
Residual deviance: 287.49  on 697  degrees of freedom
AIC: 293.49
```

Using logistic regression, we tested whether the percentage of children living below poverty line, percentage of children eligible for free student meals, percentage of families living below poverty line, number of enrolled students, and number of different schools in the district could predict whether a California School District's reporting was complete.

A Wald's z-test indicated that Poverty (z=-2.625, p<.01) and Students (z=-2.701, p<.01) were statistically significant predictors. The plain odds for Poverty were 0.66:1, with a CI ranging from 0.49 to 0.90. The plain odds for Students were 0.50:1, with a CI ranging from 0.30 to 0.82. The bands represent our uncertainty around these coefficients.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: DistrictComplete

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       699     306.65
Poverty   1   7.3049       698     299.34 0.0068768 **
Students  1  11.8506       697     287.49 0.0005764 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A chi-square omnibus test on the logistic results were significant for both Poverty, [X2(1)=7.3, p<.01], and Students [X2(1)=11.85, p<.001]. The results show that both predictors improve the model over an intercept-only model.



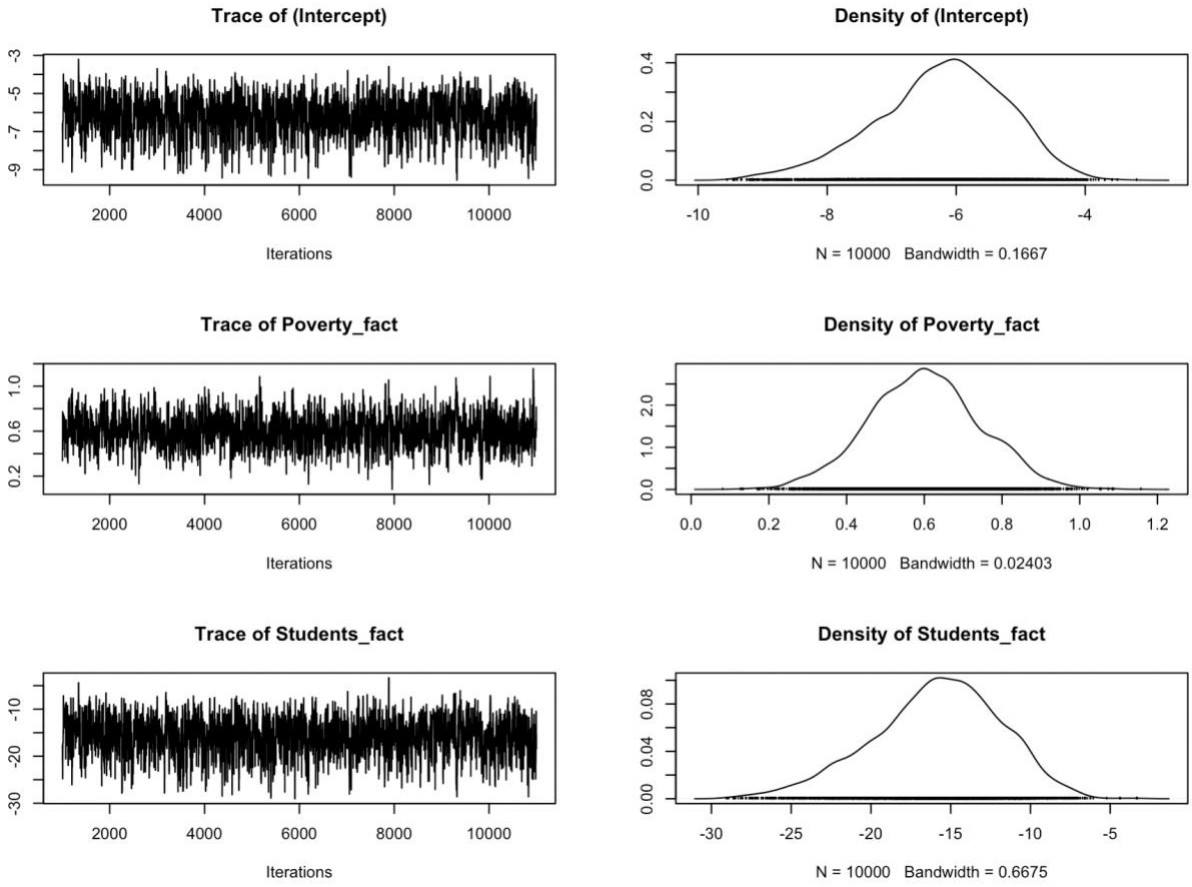

```
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean     SD Naive SE Time-series SE
(Intercept)    2.9546 0.1775 0.001775       0.005829
Poverty_fact  -0.4039 0.1586 0.001586       0.005335
Students_fact -0.7237 0.2531 0.002531       0.008394

2. Quantiles for each variable:

                2.5%     25%     50%     75%    97.5%
(Intercept)    2.6057  2.8325  2.9543  3.0683  3.30882
Poverty_fact  -0.7189 -0.5121 -0.4066 -0.3005 -0.07935
Students_fact -1.2421 -0.8872 -0.7260 -0.5502 -0.23369
```

Bayesian logistic analysis provided us with more clarity and showed similarities to conventional analysis, with only minor differences in coefficients. The 95% HDI's for both Poverty and Students did not overlap with zero, providing further support that these are meaningful predictors. When converted to regular odds, the mean value of the posterior distribution for Poverty was 0.68:1 and 0.50:1 for Students. The odds of a school district having complete reports decrease as the percentage of students/families living in poverty and the number of students/schools increases.

```
Call:
glm(formula = PctUpToDate_bin ~ Poverty_fact + Students_fact,
    family = binomial(), data = myDistricts)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.07155  -0.40778  -0.21395  -0.02646   2.89752

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.9515     0.9903  -6.010 1.85e-09 ***
Poverty_fact    0.5887     0.1429   4.118 3.81e-05 ***
Students_fact -14.8750     3.9894  -3.729 0.000193 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.66  on 699  degrees of freedom
Residual deviance: 261.26  on 697  degrees of freedom
AIC: 267.26
```

A logistic regression analysis was performed to test if Poverty and Students variables significantly predict the percentage of students fully vaccinated. A Wald's z-test indicated that Poverty (z=4.118, p<.001) and Students (z=-3.729, p<.001) were statistically significant predictors. The plain odds for Poverty were 1.80:1, with a CI ranging from 1.36 to 2.40. The plain odds for Students were essentially 0:1. The fractional CI was also very small, indicating as students and schools increase, the odds of enrolled students being fully vaccinated decreases exponentially. The bands represent our uncertainty around these coefficients.

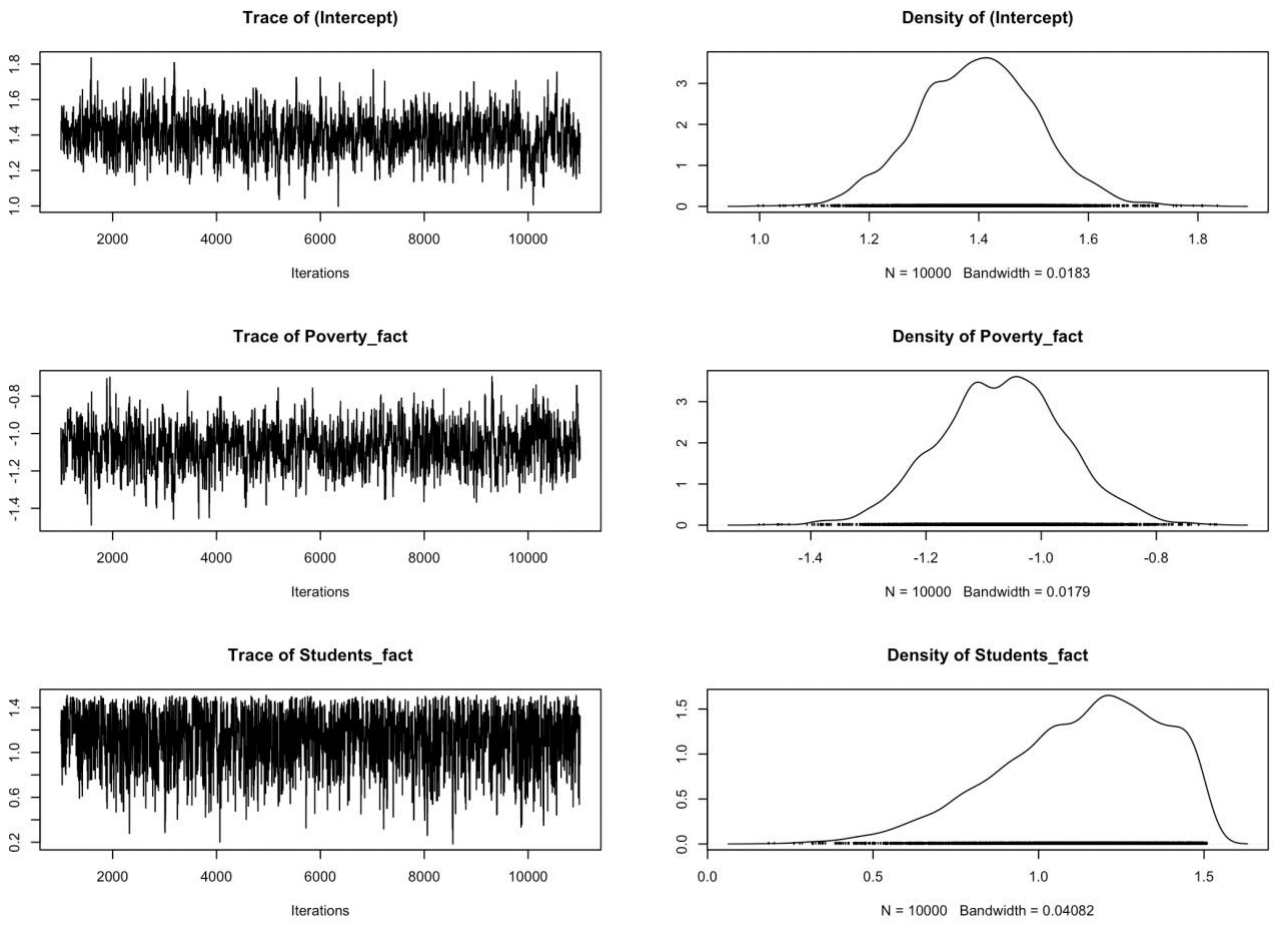
```
                    2.5 %       97.5 %
(Intercept)   2.729413e-04 0.0133650573
Poverty_fact  1.363976e+00 2.3950472578
Students_fact 4.541755e-11 0.0002870269
```



Using the Bayesian approach to logistic regression we see that the 95% HDI for both betas do not overlap with zero, providing further evidence that the population values for these coefficients are credibly different from zero. When converted to regular odds, the mean value of the posterior distribution for Poverty was 1.84:1 and 0:1 for Students, which are identical to the conventional logistic test. The odds of enrolled students being fully vaccinated increase as the percentage of students/families living in poverty increases and decreases as the number of students/schools increases.

```
Call:
glm(formula = PctBeliefExempt_bin ~ Poverty_fact + Students_fact,
    family = binomial, data = myDistricts)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.36039   0.01866   0.47114   0.68766   1.89857

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.4104     0.1108  12.724  < 2e-16 ***
Poverty_fact   -1.0666     0.1079  -9.887  < 2e-16 ***
Students_fact   1.2710     0.3607   3.524 0.000425 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 785.06  on 699  degrees of freedom
Residual deviance: 642.32  on 697  degrees of freedom
AIC: 648.32
```

A logistic regression analysis was performed to test if Poverty and Students variables significantly predict the percentage of all enrolled students with belief exceptions. A Wald's z-test indicated that Poverty (z=-9.887, p<.001) and Students (z=3.524, p<.001) were statistically significant predictors. The plain odds for Poverty were 0.34:1, with a CI ranging from 0.28 to 0.42. The plain odds for Students were 3.56:1, with a CI ranging from 1.86 to 7.66. The bands represent our uncertainty around these coefficients.



Using the Bayesian approach to logistic regression we see that the 95% HDI for both betas do not overlap with zero, providing further evidence that the population values for these coefficients are credibly different from zero. When converted to regular odds, the mean value of the posterior distribution for Poverty was 0.35:1 and 3.18:1 for Students, which resembles the conventional logistic test. As noted in the graph above, the HDI for Students is left skewed indicating more uncertainty around this estimate. The conventional test also had a wide CI for this estimate. This is likely due to the two largest districts, Los Angeles and San Diego Unified, heavily skewing the dataset. As mentioned in the opening of the predictive analysis section, the presence of extreme outliers can dramatically impact the accuracy of a model.

Reporting compliance was affected by the number of students/schools in the district and the percentage of students living below the poverty line. The probability a district had complete reports decreased as the number of students/schools and ratio of students living in poverty increased.

Vaccination rates were also affected by the number of students/schools and percentage of students living below the poverty line. As the number of students and schools increased, the probability that all students were fully vaccinated decreased. In contrast, as the ratio of students living in poverty increased, the probability that all students were fully vaccinated increased.

For belief exceptions, as the number of students and schools increased, the probability that a student had a belief exception increased. In contrast, as the ratio of students living in poverty increased, the probability that a student had a belief exception decreased.

Furthermore, districts with more students living in poverty also had schools with higher enrollment as compared those districts with fewer students living in poverty. To improve reporting compliance, more resources should be allocated to districts with higher enrollment and higher rates of poverty. To improve vaccination rates, resources should also be allocated to districts with high enrollment as the odds of student not being fully vaccinated increase with headcounts.