**Introduction:**

The objective of this project was to predict the birth weight of infants based on the NC\_BIRTH dataset provided us. A particular area of interest was in the difference between and influence of months on the resulting analysis of the data. I split the data into twelve sets and fit separate models on each of them and then compared the resulting models.

**Data Analysis:**

Data analysis was performed in R 3.4.4 using RStudio. In addition to partitioning the data into 12 sets to model the effect of months, it was also split in a 90:10 ratio for cross validation. Models were fit on 90 percent of the data and then the adj-R2 and RMSE were calculated based of the models’ fit with the other 10 percent. This allows a much better check of model viability in predicting real world scenarios. Since we are using real data, it is important to always cross validate before coming to conclusions.

**JANUARY:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2139.971 435.966 -4.909 1.21e-06 \*\*\*

Plurality -334.977 85.302 -3.927 9.68e-05 \*\*\*

Gender -98.196 29.305 -3.351 0.000860 \*\*\*

RaceOther 83.637 73.580 1.137 0.256159

RaceWhite 168.843 40.612 4.157 3.73e-05 \*\*\*

TotalPreg 33.099 9.928 3.334 0.000914 \*\*\*

LowBirthNorm 814.600 60.986 13.357 < 2e-16 \*\*\*

SmokerNo 109.084 46.229 2.360 0.018637 \*

WtGain 5.647 1.072 5.267 1.98e-07 \*\*\*

HYDRAM -233.363 98.272 -2.375 0.017904 \*

PINFANT 859.073 244.557 3.513 0.000480 \*\*\*

PROLAPSE 768.612 349.508 2.199 0.028280 \*

HEARTMAL -1017.673 368.690 -2.760 0.005966 \*\*

GestAge 111.101 8.481 13.100 < 2e-16 \*\*\*

Apgarscoreoneminute -33.014 12.821 -2.575 0.010278 \*

InfantTran 335.546 156.858 2.139 0.032857 \*

HYALINE 1068.391 406.845 2.626 0.008876 \*\*

VENTMORE -577.022 155.369 -3.714 0.000225 \*\*\*

Cross Validated adj-r2 = 0.606

RMSE = 396.76

**February:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3662.734 541.885 -6.759 3.83e-11 \*\*\*

Gender -181.134 33.041 -5.482 6.64e-08 \*\*\*

RaceOther 81.395 87.348 0.932 0.351859

RaceWhite 134.099 43.824 3.060 0.002331 \*\*

Numchild 30.976 15.027 2.061 0.039781 \*

LowBirthNorm 649.110 83.227 7.799 3.58e-14 \*\*\*

SmokerNo 236.438 55.867 4.232 2.75e-05 \*\*\*

WtGain 4.452 1.202 3.705 0.000235 \*\*\*

HYPERCH 576.069 119.931 4.803 2.06e-06 \*\*\*

MONITOR -190.417 62.692 -3.037 0.002510 \*\*

FEBRILE 239.338 105.486 2.269 0.023694 \*

PRECLAB -318.255 119.806 -2.656 0.008147 \*\*

DISTRESS -138.762 70.421 -1.970 0.049328 \*

GestAge 138.641 10.106 13.719 < 2e-16 \*\*\*

MomTran 497.126 220.278 2.257 0.024446 \*

Cross Validated adj-r2 = 0.583

RMSE = 379.07

**MARCH:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1544.753 391.670 -3.944 9.03e-05 \*\*\*

Plurality -454.241 87.277 -5.205 2.73e-07 \*\*\*

Gender -103.942 34.000 -3.057 0.002341 \*\*

Raceofchild -64.523 30.790 -2.096 0.036567 \*

RaceOther 441.036 158.253 2.787 0.005502 \*\*

RaceWhite 49.584 53.598 0.925 0.355305

Educationmother 17.650 6.224 2.836 0.004735 \*\*

TotalPreg 94.031 16.994 5.533 4.83e-08 \*\*\*

Terms -118.899 28.408 -4.185 3.30e-05 \*\*\*

Gestage 26.808 11.294 2.374 0.017952 \*

LowBirthNorm 672.992 82.183 8.189 1.79e-15 \*\*\*

SmokerNo 232.651 61.715 3.770 0.000181 \*\*\*

WtGain 7.506 1.338 5.610 3.18e-08 \*\*\*

ECLAMPSIA -368.598 171.621 -2.148 0.032163 \*

INDUCT 99.145 44.959 2.205 0.027846 \*

DISTRESS -295.891 91.538 -3.232 0.001300 \*\*

OTHLABOR 133.729 52.536 2.545 0.011180 \*

VAGINAL -96.365 36.797 -2.619 0.009062 \*\*

GestAge 79.301 14.205 5.583 3.70e-08 \*\*\*

VENTLESS 416.020 108.979 3.817 0.000150 \*\*\*

Cross Validated adj-r2 = 0.577

RMSE = 396.20

**APRIL:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2944.039 375.412 -7.842 2.27e-14 \*\*\*

Plurality -210.542 92.038 -2.288 0.022537 \*

Gender -130.400 30.974 -4.210 2.98e-05 \*\*\*

RaceOther 193.949 77.964 2.488 0.013150 \*

RaceWhite 138.227 41.705 3.314 0.000978 \*\*\*

Educationmother 16.259 5.501 2.956 0.003251 \*\*

Prenatal -41.289 17.287 -2.388 0.017250 \*

Visits 16.033 5.054 3.172 0.001595 \*\*

Numchild 39.294 16.379 2.399 0.016768 \*

LowBirthNorm 744.975 78.601 9.478 < 2e-16 \*\*\*

WtGain 5.023 1.106 4.543 6.80e-06 \*\*\*

BREECH -215.651 91.848 -2.348 0.019227 \*

PRCSECT 97.188 46.069 2.110 0.035337 \*

GestAge 133.516 9.525 14.017 < 2e-16 \*\*\*

KessnerIndex 156.652 49.895 3.140 0.001781 \*\*

Cross Validated adj-r2 = 0.447

RMSE = 367.62

**MAY:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2483.443 348.801 -7.120 3.62e-12 \*\*\*

Gender -75.259 33.329 -2.258 0.024357 \*

Educationoffather 16.473 5.479 3.007 0.002770 \*\*

TotalPreg -260.175 111.024 -2.343 0.019485 \*

Terms 245.292 113.572 2.160 0.031247 \*

Numchild 293.480 112.627 2.606 0.009430 \*\*

LowBirthNorm 783.405 79.354 9.872 < 2e-16 \*\*\*

VAGINAL -134.818 35.973 -3.748 0.000198 \*\*\*

GestAge 137.555 9.131 15.065 < 2e-16 \*\*\*

Cross Validated adj-r2 = 0.722

RMSE = 336.77

**JUNE:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2669.671 373.377 -7.150 2.67e-12 \*\*\*

Plurality -241.049 91.802 -2.626 0.008878 \*\*

Gender -90.623 30.681 -2.954 0.003269 \*\*

RaceOther -58.478 77.681 -0.753 0.451881

RaceWhite 133.634 41.677 3.206 0.001419 \*\*

Ageofmother 6.556 2.929 2.238 0.025587 \*

Prenatal -43.946 12.200 -3.602 0.000343 \*\*\*

BirthAttendant 90.235 25.228 3.577 0.000377 \*\*\*

Numchild 58.534 15.927 3.675 0.000260 \*\*\*

LowBirthNorm 710.111 72.516 9.792 < 2e-16 \*\*\*

SmokerNo 223.135 57.953 3.850 0.000131 \*\*\*

WtGain 3.583 1.200 2.986 0.002952 \*\*

RHSEN -354.960 166.899 -2.127 0.033866 \*

VAGINAL -69.757 35.275 -1.978 0.048465 \*

GestAge 132.186 9.370 14.107 < 2e-16 \*\*\*

IANEMIA 1421.761 373.807 3.803 0.000158 \*\*\*

BINJURY 1530.044 370.853 4.126 4.25e-05 \*\*\*

Cross Validated adj-r2 = 0.570

RMSE = 349.25

**July:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2059.744 308.114 -6.685 5.44e-11 \*\*\*

Plurality -275.144 93.353 -2.947 0.003334 \*\*

Gender -70.663 31.157 -2.268 0.023697 \*

RaceOther 42.517 80.931 0.525 0.599540

RaceWhite 167.712 42.042 3.989 7.48e-05 \*\*\*

Ageoffather 8.234 2.482 3.318 0.000965 \*\*\*

Numchild 52.241 14.681 3.558 0.000404 \*\*\*

LowBirthNorm 764.484 79.244 9.647 < 2e-16 \*\*\*

SmokerNo 221.069 51.738 4.273 2.26e-05 \*\*\*

WtGain 4.954 1.162 4.264 2.34e-05 \*\*\*

HYPERCH -281.280 136.031 -2.068 0.039106 \*

HYPERPR -147.371 69.458 -2.122 0.034284 \*

RUPTURE -255.010 110.030 -2.318 0.020817 \*

CEPHALO 336.584 136.481 2.466 0.013946 \*

OMPHALAC 804.240 376.525 2.136 0.033102 \*

GestAge 110.370 8.194 13.469 < 2e-16 \*\*\*

ASPIRATE 859.871 265.140 3.243 0.001250 \*\*

Cross Validated adj-r2 = 0.623

RMSE = 400.62

**AUGUST:**

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1656.510 347.368 -4.769 2.33e-06 \*\*\*

Plurality -202.013 83.217 -2.428 0.015495 \*

Gender -118.486 29.882 -3.965 8.22e-05 \*\*\*

Raceofchild -103.353 34.415 -3.003 0.002784 \*\*

RaceOther 499.658 152.151 3.284 0.001083 \*\*

RaceWhite 62.328 52.051 1.197 0.231609

TotalPreg 27.039 10.458 2.586 0.009957 \*\*

LowBirthNorm 725.242 68.594 10.573 < 2e-16 \*\*\*

SmokerNo 166.285 50.128 3.317 0.000964 \*\*\*

WtGain 5.137 1.133 4.533 7.03e-06 \*\*\*

PRETERM -355.109 132.803 -2.674 0.007701 \*\*

MONITOR -111.739 52.984 -2.109 0.035367 \*

BREECH -150.787 73.957 -2.039 0.041905 \*

CEPHALO 301.292 122.034 2.469 0.013830 \*

PRCSECT 102.711 43.331 2.370 0.018084 \*

GestAge 115.119 8.413 13.683 < 2e-16 \*\*\*

ISEIZURE 1182.769 365.887 3.233 0.001294 \*\*

Cross Validated adj-r2 = 0.513

RMSE = 406.13

**SEPTEMBER:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1949.666 344.433 -5.661 2.26e-08 \*\*\*

Plurality -283.665 92.697 -3.060 0.002303 \*\*

Gender -155.202 28.582 -5.430 7.95e-08 \*\*\*

RaceOther -20.123 76.036 -0.265 0.791359

RaceWhite 175.169 36.594 4.787 2.10e-06 \*\*\*

Terms -65.294 20.596 -3.170 0.001594 \*\*

Numchild 77.982 13.543 5.758 1.31e-08 \*\*\*

LowBirthNorm 710.979 74.953 9.486 < 2e-16 \*\*\*

SmokerNo 115.471 51.439 2.245 0.025114 \*

WtGain 3.203 1.095 2.926 0.003557 \*\*

DIABETES 250.035 103.980 2.405 0.016465 \*

RENAL 551.326 260.486 2.117 0.034676 \*

CEPHALO 490.275 133.039 3.685 0.000247 \*\*\*

VAGINAL -71.575 31.423 -2.278 0.023062 \*

OTHURO 801.989 374.654 2.141 0.032674 \*

POLYDACT 771.662 368.062 2.097 0.036416 \*

GestAge 123.172 8.539 14.424 < 2e-16 \*\*\*

VENTLESS 192.202 94.402 2.036 0.042154 \*

Cross Validated adj-r2 = 0.686

RMSE = 386.1

**OCTOBER:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2169.338 329.747 -6.579 1.04e-10 \*\*\*

Gender -149.557 31.745 -4.711 3.06e-06 \*\*\*

Raceofchild -42.194 11.877 -3.552 0.000412 \*\*\*

Educationoffather 11.418 5.411 2.110 0.035250 \*

Terms 54.330 19.031 2.855 0.004456 \*\*

Gestage 20.651 9.880 2.090 0.037028 \*

LowBirthNorm 878.430 72.721 12.080 < 2e-16 \*\*\*

SmokerNo 213.238 55.773 3.823 0.000145 \*\*\*

WtGain 5.845 1.163 5.025 6.67e-07 \*\*\*

DIABETES 388.173 90.826 4.274 2.24e-05 \*\*\*

HERPES -329.621 147.356 -2.237 0.025660 \*

PINFANT 639.815 193.816 3.301 0.001020 \*\*

OTHLABOR 140.567 44.111 3.187 0.001514 \*\*

GestAge 92.502 12.464 7.421 3.99e-13 \*\*\*

Cross Validated adj-r2 = 0.581

RMSE = 407.48

**NOVEMBER:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1078.191 505.587 -2.133 0.033396 \*

Plurality -279.471 93.149 -3.000 0.002817 \*\*

Gender -130.042 30.308 -4.291 2.10e-05 \*\*\*

Educationmother 13.869 5.585 2.483 0.013310 \*

TotalPreg 39.109 10.581 3.696 0.000240 \*\*\*

Marital -120.745 35.940 -3.360 0.000834 \*\*\*

LowBirthNorm 759.862 75.661 10.043 < 2e-16 \*\*\*

SmokerNo 147.539 55.764 2.646 0.008379 \*\*

WtGain 5.533 1.154 4.795 2.09e-06 \*\*\*

PINFANT 524.213 211.739 2.476 0.013591 \*

TOCOLY -262.480 89.333 -2.938 0.003437 \*\*

MECONIUM 183.276 68.811 2.663 0.007957 \*\*

DISTRESS -193.979 74.041 -2.620 0.009034 \*\*

GestAge 110.288 9.731 11.333 < 2e-16 \*\*\*

Apgarscorefiveminutes 56.278 26.973 2.087 0.037385 \*

InfantTran -505.015 195.328 -2.585 0.009976 \*\*

BINJURY 775.822 266.555 2.911 0.003751 \*\*

VENTMORE 353.278 152.957 2.310 0.021270 \*

Cross Validated adj-r2 = 0.562

RMSE = 415.88

**DECEMBER:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2512.047 310.350 -8.094 3.23e-15 \*\*\*

Gender -133.291 30.428 -4.380 1.40e-05 \*\*\*

Raceofchild -43.067 10.891 -3.954 8.60e-05 \*\*\*

Ageofmother 6.674 2.629 2.538 0.011386 \*

TotalPreg 49.672 13.299 3.735 0.000206 \*\*\*

Terms -52.115 21.723 -2.399 0.016741 \*

Gestage 42.990 9.138 4.704 3.16e-06 \*\*\*

LowBirthNorm 862.516 68.935 12.512 < 2e-16 \*\*\*

SmokerNo 189.483 58.363 3.247 0.001233 \*\*

WtGain 3.004 1.074 2.796 0.005338 \*\*

DIABETES 238.706 92.662 2.576 0.010230 \*

HYPERCH -518.166 167.005 -3.103 0.002008 \*\*

OTHMS -544.834 265.300 -2.054 0.040443 \*

GestAge 84.636 11.888 7.119 3.11e-12 \*\*\*

Apgarscoreoneminute -22.474 11.298 -1.989 0.047133 \*

BINJURY 1694.742 371.286 4.565 6.08e-06 \*\*\*

Cross Validated adj-r2 = 0.408

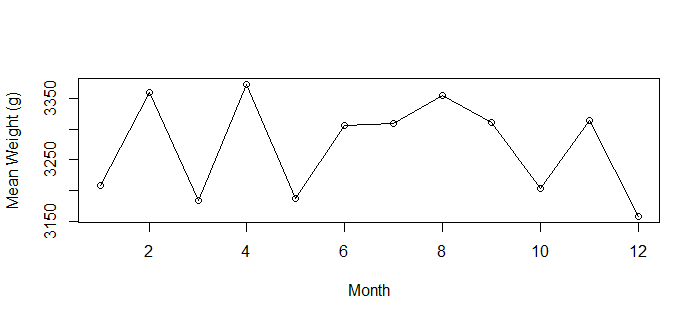
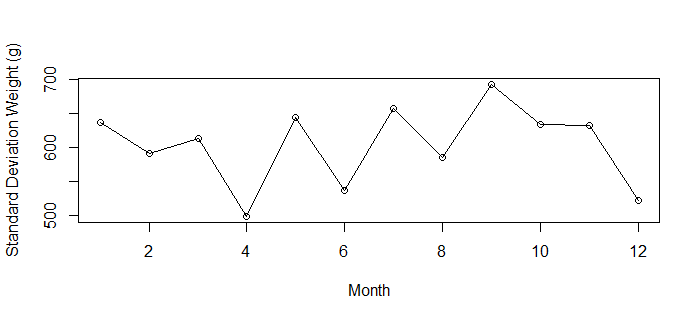
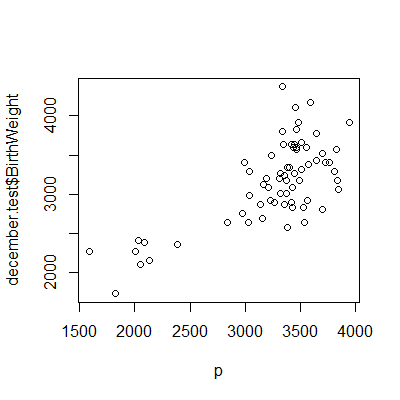
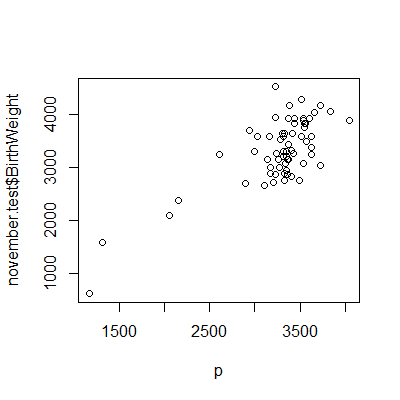
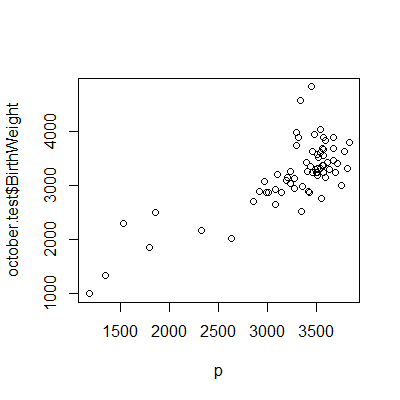
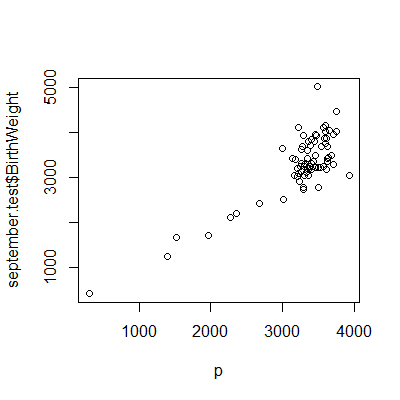
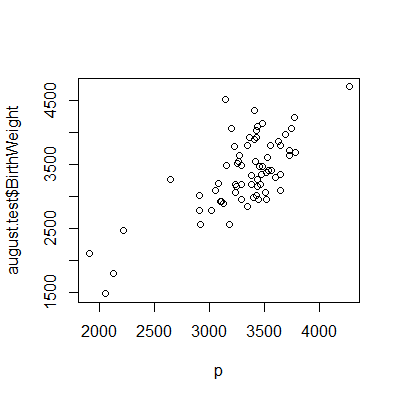
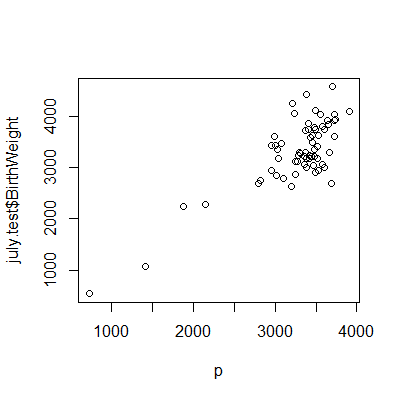
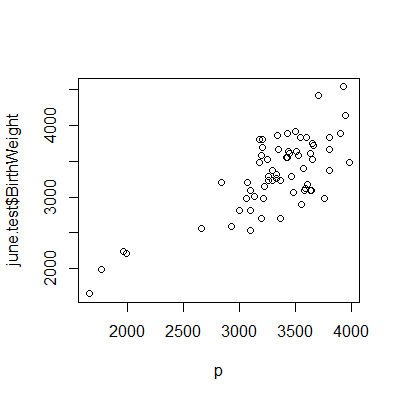
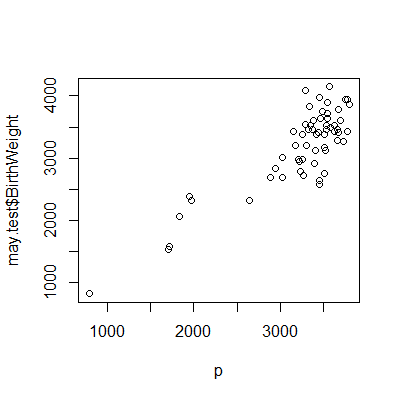
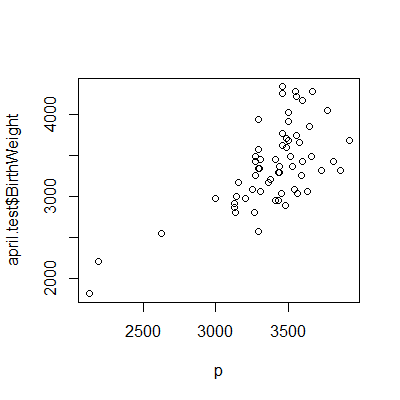
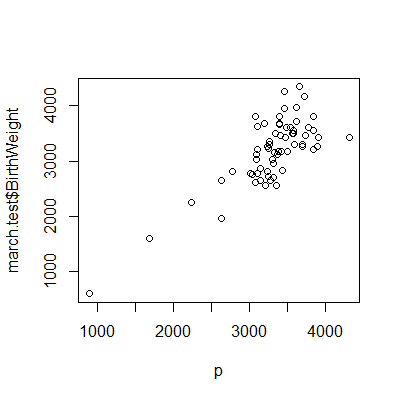
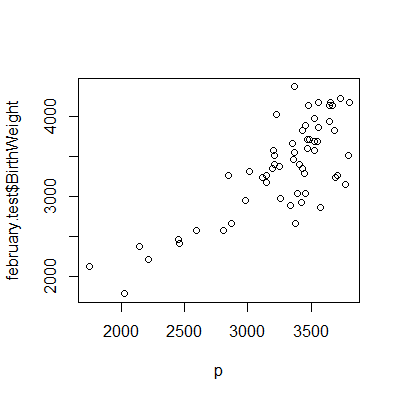
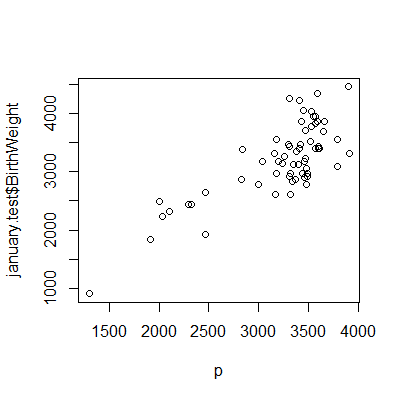
RMSE = 398.98

In preliminary analysis I’ve taken note of the coefficients that are common to all models: gender, lowbirthnorm, and gestage. These variables all seem fairly obvious. Gender is always a very negative coefficient indicating that male babies weigh more than female ones. GestAge always has a positive effect as this indicates the amount of time the baby has had to grow before being born. Low birth weight is also a fairly obvious indicator of birth weight.

Close seconds in coefficient popularity were weightGain, smoker, plurality, race, totalPreg, and numchild. There were 62 total unique coefficients over all of the models. Many of the models had a variety of numbers of significant coefficients with the smallest 8 and the largest at 18. Almost every model had an r2 over 0.5 with December being the only month lower at 0.408. Most of the RMSEs were under 400 with October and November being over.

Upon plotting and checking residual assumptions of each model everything looked in order. Each month had a few high leverage points with the exception of August which was tightly grouped in the cooks distance plot. I also plotted the y vs. yhat for each model and will include the plots below.

Furthermore, I have plotted the mean and standard deviation of the birthweights for the months for greater analysis of months’ effects.



**CONCLUSION:**

This was a very interesting project and there was a greater amount of disparity between the months than I had initially assumed there would be. The popular predictors all seemed fairly obvious in their placement as they all measured weight, time, or gender. An interesting note was the significance and high coefficient on not being a smoker. It seems that smoking tends to decrease the weight of the child through the pregnancy. An understandable one was the negative coefficient on plurality meaning that having more than one baby at a time decreases the weight of the lot as, I assume, there is less room and nutrients each. “This town ain’t big enough for the two of us” kinda stuff. It is also interesting that geing white correlated much stronger with a higher birth weight than the other races. Total pregnancies was also correlated positively with the birth weight, which I find interesting.

There was a very wide disparity with the ability and complexity of prediction for the partitioned months. May was very efficiently and parsimoniously predicted at r2 of 0.722 with only 8 predictors while December and November had much lower r2 and December and January had very high numbers of significant predictors. I can posit that this may be due to increased illness in the winter months possibly caused by a vitamin D deficiency but anything more than a pure observation borders on dishonesty. An interesting further study would be to segment the data further by year and add in data for the known weather patterns in the area and time of interest to investigate the weathers effect on birth.

Observing the y vs yhat graphs I notice a few interesting trends. Most births seem to form a single cluster roughly centered around the line, y = yhat. However, in every graph there is a small trail of lower values. These seem to be a good explanation for our few high leverage points in every set. It is interesting to also note that there is a lower tail but no upper tail. This could be explained anatomically in the fact that a babies size seems to be mostly a function of the time that it is alive, hence the prolific nature of the Gestation Age variable in explaining birth weight. I know of medical emergencies that would cause a baby to need to be born early but none that would require a baby to stay in the womb longer. Modern medicine now has drugs to induce labor and the facilities and understanding to sufficiently incubate premature babies and thus we might explain the prevalence of small babies but the mysterious lack of giant babies with an understanding of modern medicine and human anatomy.

Observing our line graphs for the mean and standard deviation of the birth weight of babies over the months, I see no obvious trend. It seems to be an incredibly variable measure and though I might be able to make a statement about high variability here and stability there, I would rather not stray from the realm of credibility by making assumptions and asserting half-truths.

Overall, the factors affecting birth weight were synonymous with my assumptions and the difference between the months turned out to be harder to pin and very non-obvious. The r2 values of greater than 0.5 lead me to believe that I have developed a pretty accurate measure with the models created. I hesitate to draw conclusions from the summary statistics of the months as I see too much variability to accurately state anything. I’ve explored a few interesting conclusions like the range of baby sizes and the apparent lack of giant babies. I can state that I’ve gained a greater appreciation of the complexities of birth and of the intricacies of modeling real world phenomena.