Lab 4: Does Prenatal Care Improve Infant Health?

W203: Statistics for Data Science

Dan Guevarra, Bas Hendri, Saurabh Jaju (beatspace9,dip,saurabh.jaju@berkeley.edu)

December 8, 2016

```r
load("bwght_w203.RData")
desc
```

```
##    variable                           label
## 1      mage             mother's age, years
## 2     meduc            mother's educ, years
## 3    monpre       month prenatal care began
## 4     npvis total number of prenatal visits
## 5      fage             father's age, years
## 6     feduc            father's educ, years
## 7     bwght             birth weight, grams
## 8     omaps          one minute apgar score
## 9     fmaps         five minute apgar score
## 10     cigs          avg cigarettes per day
## 11    drink             avg drinks per week
## 12      lbw             =1 if bwght <= 2000
## 13     vlbw             =1 if bwght <= 1500
## 14     male                 =1 if baby male
## 15    mwhte              =1 if mother white
## 16    mblck              =1 if mother black
## 17     moth           =1 if mother is other
## 18    fwhte              =1 if father white
## 19    fblck              =1 if father black
## 20     foth           =1 if father is other
## 21   lbwght                      log(bwght)
## 22   magesq                          mage^2
## 23  npvissq                         npvis^2
```

```r
library(car)
library(sandwich)
library(stargazer)
library(reshape2)
library(ggplot2)
library(data.table)
library(lmtest)
```

1 Introduction

According to the National Institutes of Health, "prenatal care should begin as soon as a woman knows or suspects she is pregnant… prenatal care is important to help promote a healthy pregnancy. Women who do not seek prenatal care are three times as likely to deliver a low birth weight infant, and lack of prenatal care can also increase the risk of infant death." The following statistical analysis aims to substantiate these claims using a classical linear regression model.

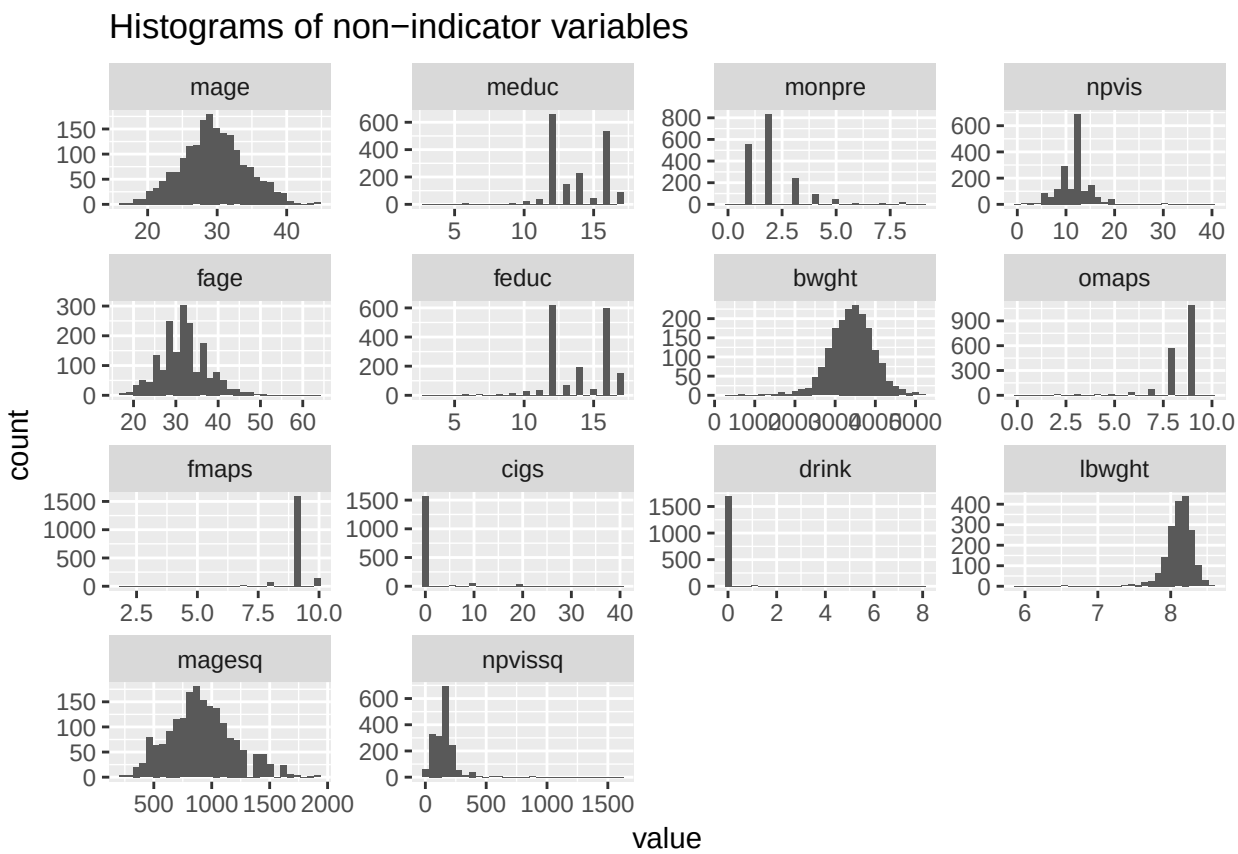
https://www.nichd.nih.gov/health/topics/preconceptioncare/conditioninfo/Pages/default.aspx

2 Preliminary Explorations

```
# race
race <- c(1,2,NA,3)
ilbw <- c(1,2,3)
data$mrace <- (race[apply(t(t(data[,c('mwhte','mblck','moth')])*c(1,2,4)),1,sum)])
data$frace <- (race[apply(t(t(data[,c('fwhte','fblck','foth')])*c(1,2,4)),1,sum)])
data$ilbw <- (ilbw[1+apply(data[,c('lbw','vlbw')],1,sum)])
is.factor.var <- c(14,24:26)
is.value.var <- c(1:11,21:23)
```

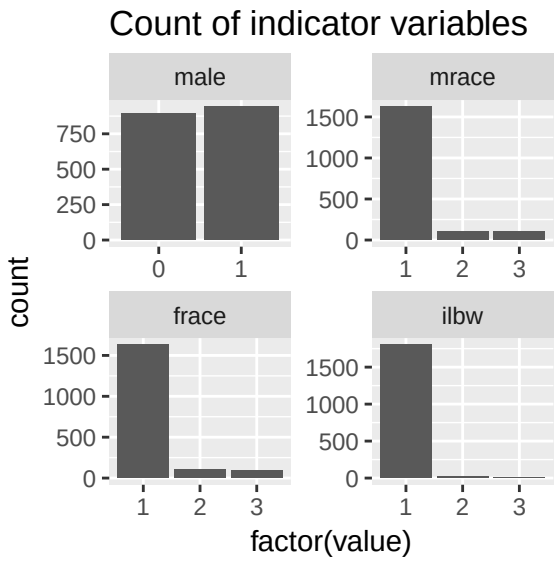
```
# non-indicator vars:
ggplot(melt(data[,is.value.var]), aes(x=value)) +
    geom_histogram(bins=30) +
    facet_wrap(~variable, scales='free') +
    ggtitle('Histograms of non-indicator variables')
```

```
## No id variables; using all as measure variables
```

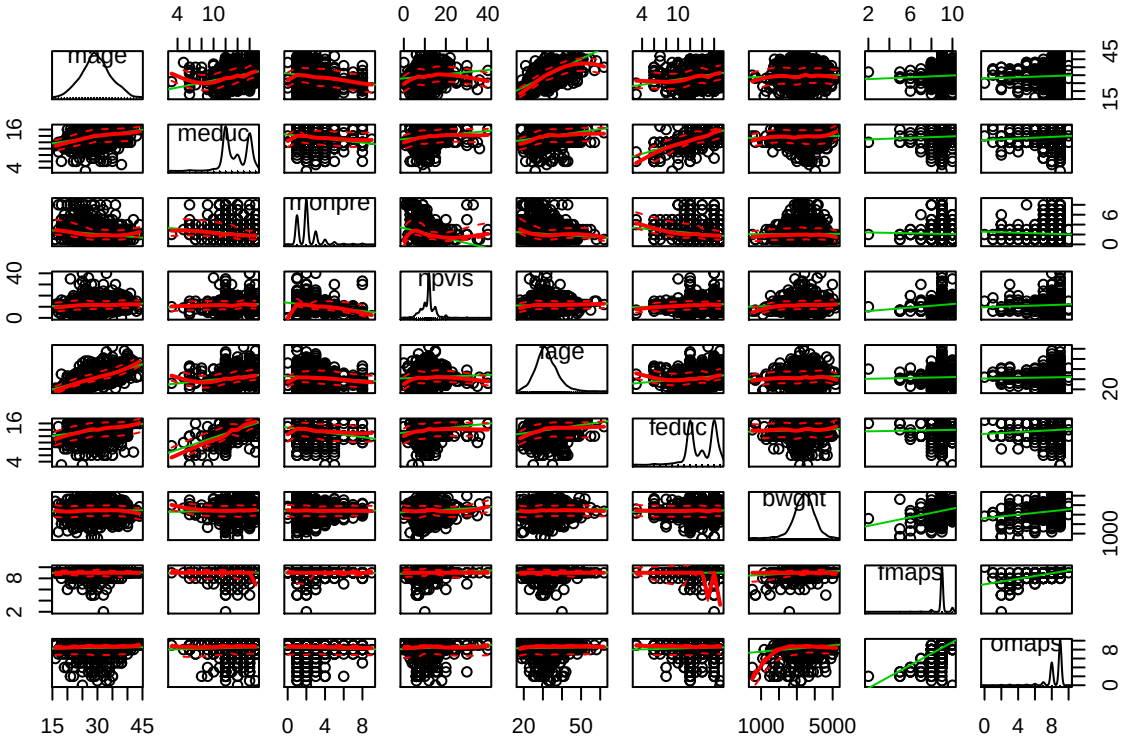



```
# indicator vars
ggplot(melt(data[,is.factor.var]), aes(x=factor(value))) +
    stat_count() +
    facet_wrap(~variable, scales='free') +
    ggtitle('Count of indicator variables')
```

```
## No id variables; using all as measure variables
```




```
# warnings for loess due to discretized variables have been suppressed
scatterplotMatrix(data[,c('mage', 'meduc', 'monpre', 'npvis', 'fage',
                          'feduc', 'bwght', 'fmaps', 'omaps')])
```



```r
bwdt<-as.data.table(data)
bwdt[order(fmaps),.N,by=fmaps]
```

```
##    fmaps    N
## 1:     2    1
## 2:     5    4
## 3:     6    6
## 4:     7   12
## 5:     8   67
## 6:     9 1600
## 7:    10  139
## 8:    NA    3
```

```r
bwdt[order(cigs),.N,by=cigs]
```

```
##     cigs    N
##  1:    0 1575
##  2:    1    2
##  3:    2    5
##  4:    3    5
##  5:    4    7
##  6:    5   11
##  7:    6    6
##  8:    7    1
##  9:    8    2
## 10:    9    1
## 11:   10   50
## 12:   12    1
## 13:   15    7
## 14:   16    1
## 15:   20   39
## 16:   25    1
## 17:   30    6
## 18:   40    2
## 19:   NA  110
```

Variable transformation and aggregation

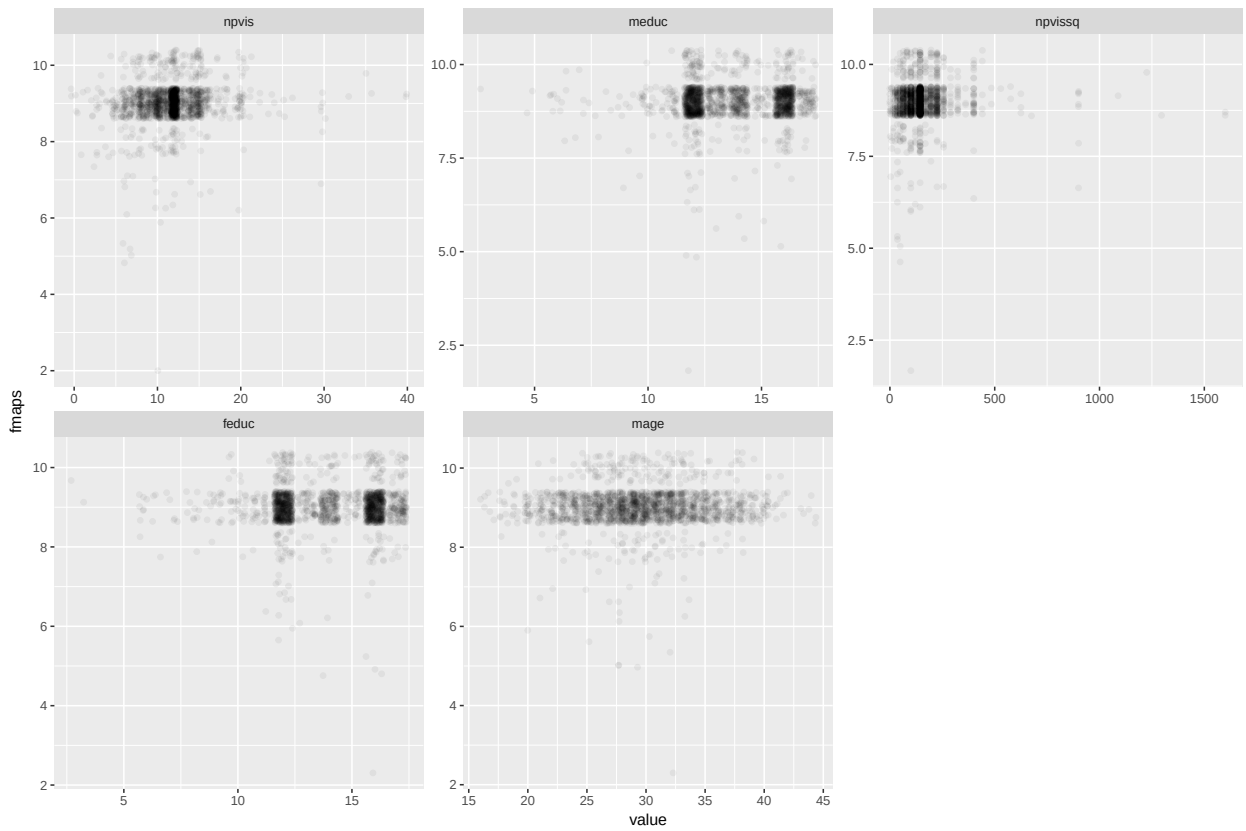
The non-indicator variables don't appear particularly conducive to log transformation. The as-supplied transformed variables `magesq` and `npvissq` appear more positively skewed their originals. The `lbwght` variable is also more negatively skewed than the non-transformed `bwght`. The education variables `meduc` and `feduc` have spikes at the end of traditional highschool and 4-year college educations. Indicator variables for degree of education may be a more appropriate grouping that weighs higher education separately.

Among the indicator variables, there are far more "normal" birthweights than low or very low birthweights, and more white mothers and fathers than black or other.

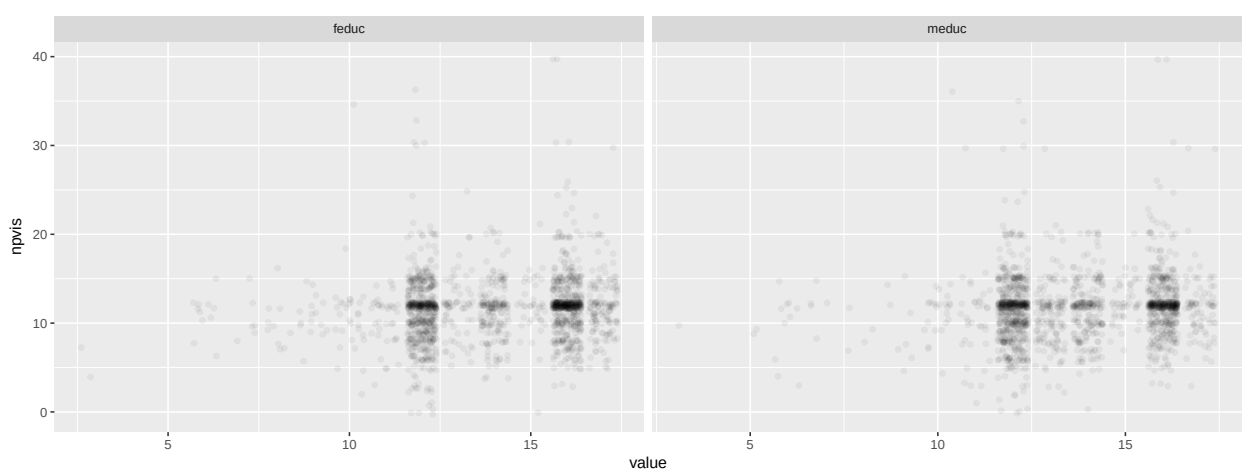
We create variables `mrace` and `frace` to summarize the races.

```r
bwdt[,mrace:=factor(mwhte*1+mblck*2+moth*3,
                    levels=c(1, 2, 3),
                    labels=c('white', 'black', 'other'))]
bwdt[,frace:=factor(fwhte*1+fblck*2+foth*3,
                    levels=c(1, 2, 3),
                    labels=c('white', 'black', 'other'))]
```

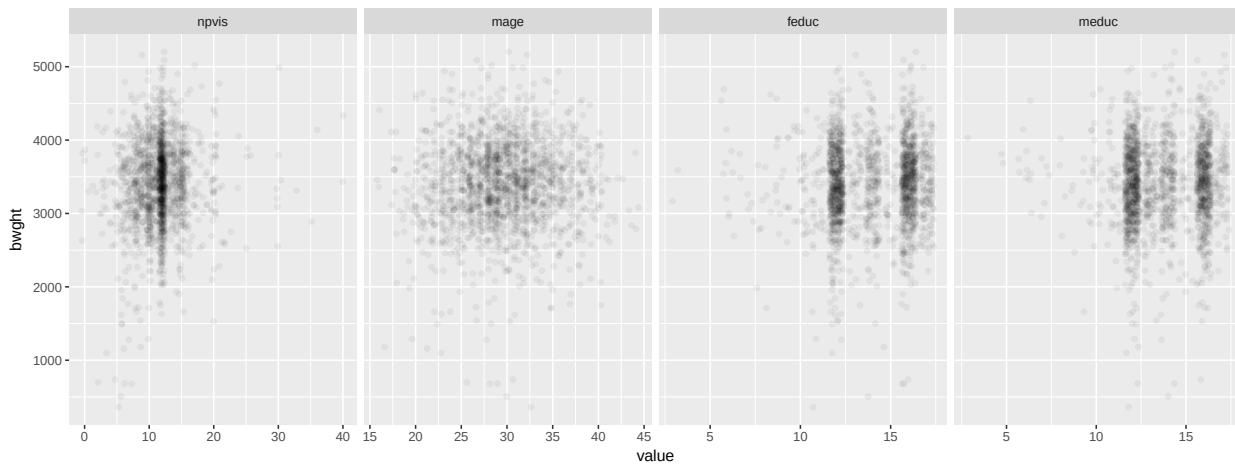
```r
ggplot(melt(bwdt[,.(fmaps,npvis,meduc,npvissq,feduc,mage)], id.vars='fmaps'),
       aes(x=value, y=fmaps)) +
       geom_jitter(alpha=I(1/20)) +
       facet_wrap(~variable, scales='free', ncol=3)
```




```r
ggplot(melt(bwdt[,.(npvis,feduc,meduc)], id.vars='npvis'),
       aes(x=value, y=npvis)) +
       geom_jitter(alpha=I(1/20)) +
       facet_grid(.~variable, scales='free')
```



```r
ggplot(melt(bwdt[,.(bwght,npvis,mage,feduc,meduc)], id.vars='bwght'),
       aes(x=value, y=bwght)) +
       geom_jitter(alpha=I(1/20)) +
       facet_grid(.~variable, scales='free')
```



Variables to include in each model

APGAR score is an aggregate rating of heart rate, respiratory effort, muscle tone, response to stimulation, and skin coloration. The one minute score is used to determine immediate medical response, but a low score doesn't directly translate to an unhealthy baby (premature or c-section births typically score low). The five minute score is a measure of progress and response to medical intervention and is a better indicator of a baby's overall health than the one minute score. The `fmaps` variable seems to be the appropriate dependent variable that describes infant health. However, the distribution of `fmaps` data is poorly suited for regression exercise as the variable is ordinal and an overwhelming majority (close to 90%) is concentrated in `fmaps==9`.

We opt for birthweight instead as the distribution is more well-behaved. Our models will not control for "low" and "very low" birthweight indicator variables. As a proxy for infant health, the ceteris paribus effect of `npvis` on birthweight will be evaluated against the "low" and "very low" birthweight thresholds of 2500 and 1500 grams respectively.

3 Models

3.1 Model with only the explanatory variables of key interest

We are interested in discovering the effect of prenatal care, so for key interest variables we take count of prenatal visits `npvis` and months of prenatal care `monpre`. We additionally take `npvissq` to model diminishing return.

```r
(mod1b<-lm(bwght ~ npvis + npvissq + monpre, data=bwdt))
```

```
##
## Call:
## lm(formula = bwght ~ npvis + npvissq + monpre, data = bwdt)
##
## Coefficients:
## (Intercept)        npvis      npvissq       monpre
##    2916.155       50.703       -1.114       29.008
```

3.2 Model with only covariates without bias

We additionally take average cigarette smoked `cigs`, and add mother's age and squared age `mage` and `magesq` for dimishing return. These variables relate to the mother's health and are expected to correlate with the number of prenatal vists `npvis`. Controlling for these additional variables would improve exogeneity of the `npvis` variable and the OLS coefficient estimates.

```
(mod2b<-lm(bwght ~ npvis + npvissq + monpre + cigs + mage + magesq, data=bwdt))
```

```
##
## Call:
## lm(formula = bwght ~ npvis + npvissq + monpre + cigs + mage +
##     magesq, data = bwdt)
##
## Coefficients:
## (Intercept)        npvis      npvissq       monpre         cigs
##   1830.8827      40.9214      -0.8734      33.2893     -10.0188
##        mage       magesq
##     76.3582      -1.2089
```

3.3 Model with potentially problematic covariates

In model 3 we will consider various indicator variables and the non indicator variable drink also four combination variables for combination variable of type of couple having the baby will also be considered. The types of couple would be $ White Couple$, *blackCouple*, *otherCouple* and *mixedCouple*

```
# Dummy coding factor variables to Dichotomous
bwdt$whtCouple<-as.numeric(bwdt$mwhte & bwdt$fwhte)
bwdt$blkCouple<-as.numeric(bwdt$mblck & bwdt$fblck)
bwdt$othCouple<-as.numeric(bwdt$moth & bwdt$foth)
bwdt$mixedCouple<-as.numeric(!(bwdt$whtCouple | bwdt$blkCouple | bwdt$othCouple))

(mod3b<-lm(bwght ~ npvis + npvissq + monpre + cigs+ mage + magesq + drink + mixedCouple +
                   whtCouple + blkCouple, data=bwdt))
```

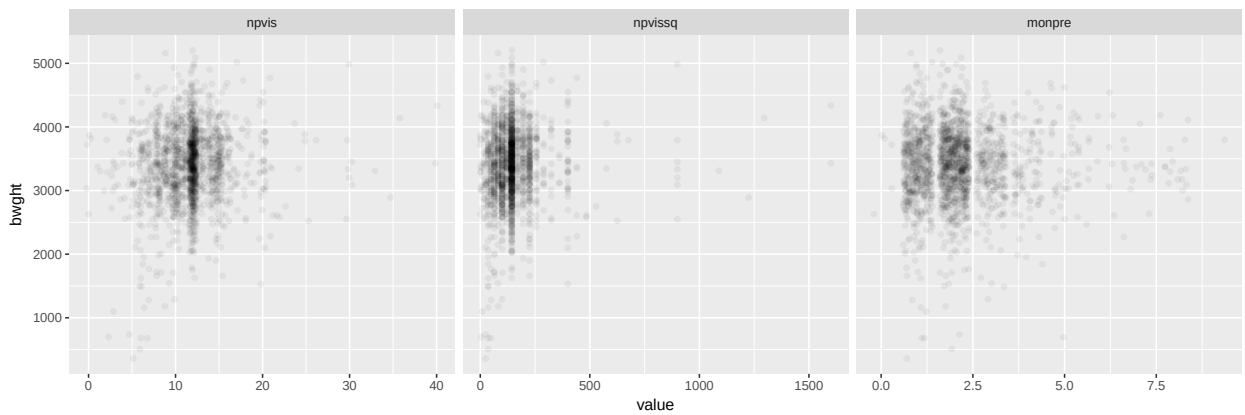
```
##
## Call:
## lm(formula = bwght ~ npvis + npvissq + monpre + cigs + mage +
##     magesq + drink + mixedCouple + whtCouple + blkCouple, data = bwdt)
##
## Coefficients:
## (Intercept)        npvis      npvissq       monpre         cigs
##   1530.2239      42.7848      -0.9019      34.3352     -10.9830
##        mage       magesq        drink  mixedCouple    whtCouple
##     78.2528      -1.2361     -20.1197     350.6606     263.1422
##   blkCouple
##    232.6484
```

4 Diagnostic Assessments

4.1 Model 1

MLR.1: Linear in Parameters

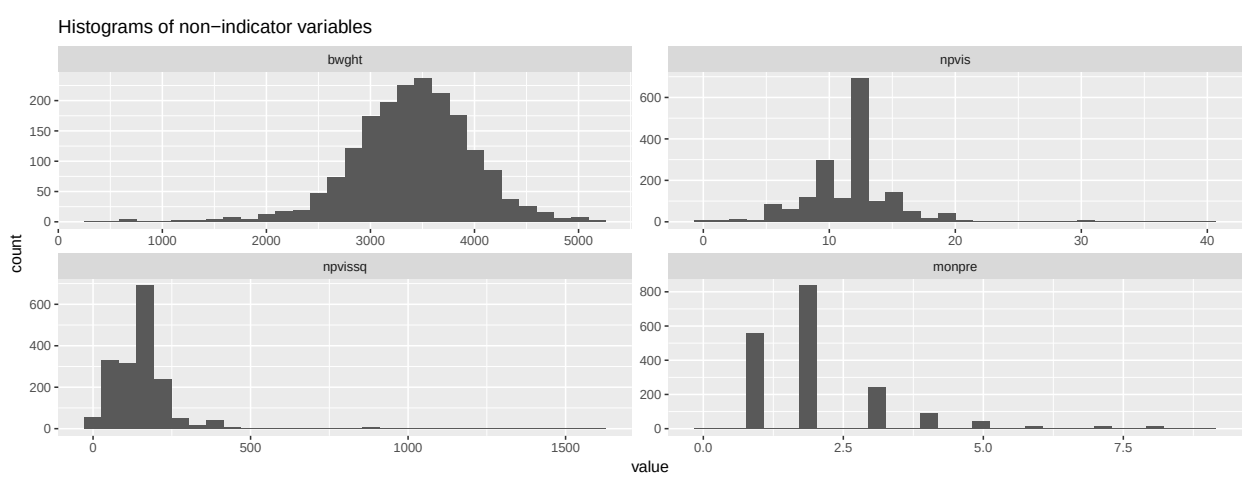
```
ggplot(melt(bwdt[,.(bwght,npvis,npvissq,monpre)], id.vars='bwght'), aes(x=value, y=bwght)) +
    geom_jitter(alpha=I(1/20)) +
    facet_grid(.~variable, scales='free')
```



From the scatterplot matrix, we observe a major cluster of bivariate Gaussian distribution on top left for all predictors. However, in all cases there are also distribution of linearly increasing effect of predictor on birth weight.

MLR.2: Random Sampling

```
ggplot(melt(bwdt[,.(bwght,npvis,npvissq,monpre)]), aes(x=value)) +
    geom_histogram(bins=30) +
    facet_wrap(~variable, scales='free') +
    ggtitle('Histograms of non-indicator variables')
```



From the histograms we observe that the sample has somewhat normally distributed variable values over the ranges.

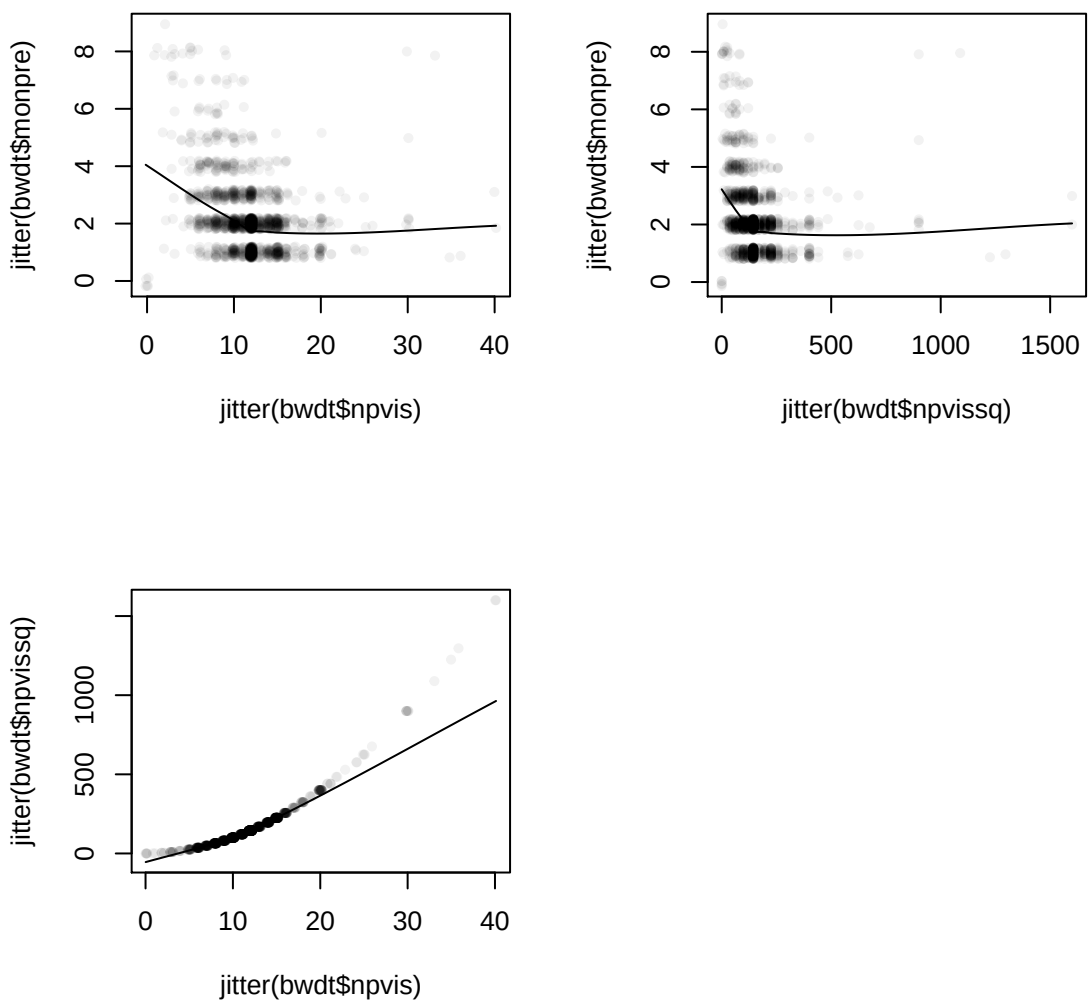
MLR.3: No Perfect Collinearity

```
par.orig <- par(no.readonly=T); par(mfrow=c(2,2)) # plot on 2x2

scatter.smooth(jitter(bwdt$npvis), jitter(bwdt$monpre), pch=20, col=rgb(0,0,0,0.05))
scatter.smooth(jitter(bwdt$npvissq), jitter(bwdt$monpre), pch=20, col=rgb(0,0,0,0.05))
scatter.smooth(jitter(bwdt$npvis), jitter(bwdt$npvissq), pch=20, col=rgb(0,0,0,0.05))

cor(bwdt[,.(npvis,npvissq,monpre)], use='p')
```

```
##              npvis    npvissq     monpre
## npvis    1.0000000  0.9340980 -0.3061006
## npvissq  0.9340980  1.0000000 -0.1799153
## monpre  -0.3061006 -0.1799153  1.0000000
```

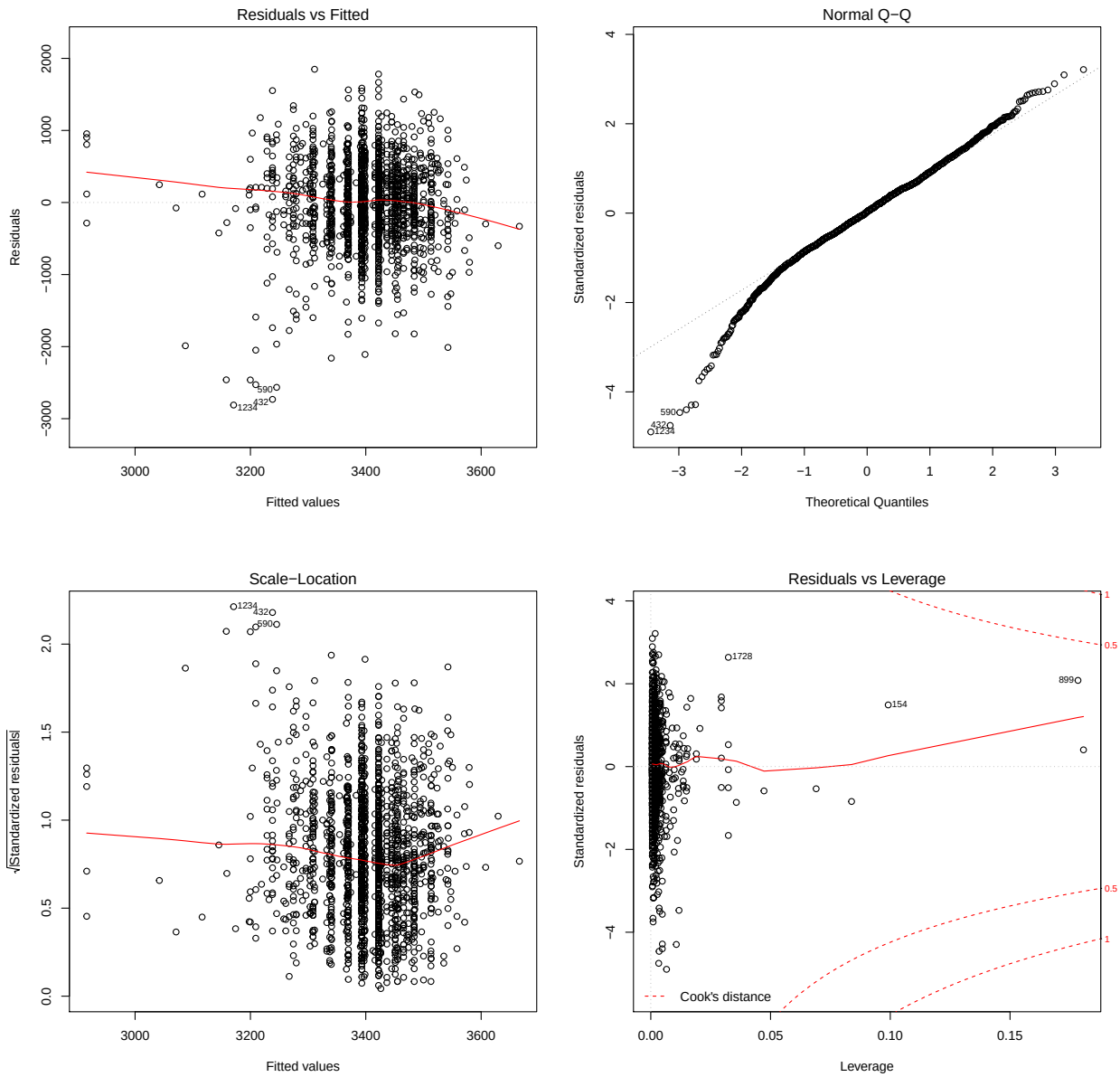


Scatter plot between the predictors show that collinearity is unlikely, though raw values show minor correlation. We ignore the expected `npvis-npvissq` correlation.

MLR.4: Zero Conditional Mean

```r
par(mfrow=c(2,2)) # plot diagnostics on 2x2

plot(mod1b)
```



Looking at residuals v. fitted values plot, we see the residuals are not too well-behaved in terms of being in the proximity of 0. Remarkedly a potential problem is the apparently decreasing loess across increasing fitted values. Although this may also be caused by sparsity of points producing higher and lower fitted values, it may suggest that our model has overestimated the magnitude of some coefficients. However, given that we have large sample size, we argue for zero conditional mean from asymptotics.

MLR.5: Homoskedasticity

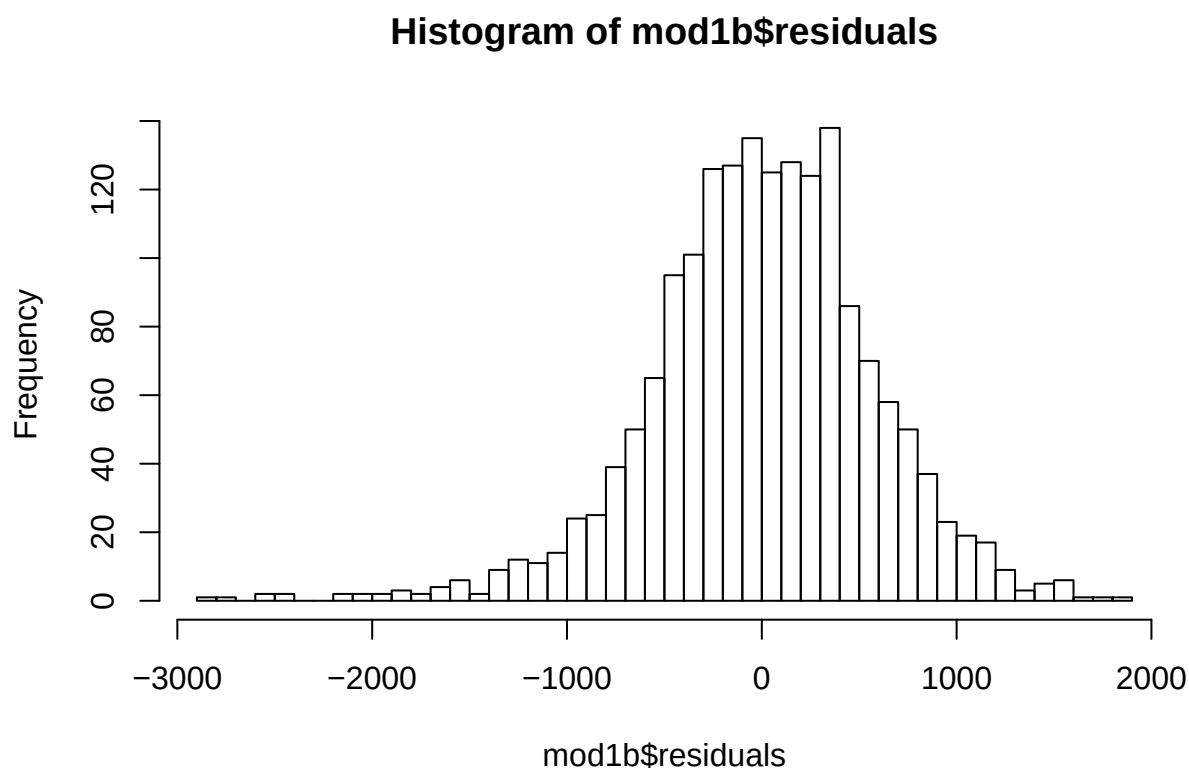
```r
ncvTest(mod1b)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 77.20357    Df = 1     p = 1.542146e-18
```

From residual v. fitted values plot evidently the center of the plot forms a thicker band than the outer parts. Again, this may be caused by sparsity of points in the outer region compared to the center. The Scale-Location plot shows a somewhat straight line within the proximity of $\sqrt{\text{Standardized residuals}} = 1$ but not without a depression around center and some outlying points. Breusch-Pagan test is showing sufficient statistical evidence at $p < 0.05$ to reject homoskedasticity. We should then use heteroskedasticity-robust errors since inferences from variance with homoskedasticity assumption will be biased.

MLR.6: Normality

```r
hist(mod1b$residuals, breaks=50)
```



Points at the either ends have slight departure from the theoretical Q-Q line towards 0 suggesting violation of normality, especially large on the lower end of quantile. Histogram of residuals however appears to be well-behaved and shows approximately normal distribution of residuals, despite the long left tail. To respond to this potential violation we again argue OLS asymptotics due to large sample.

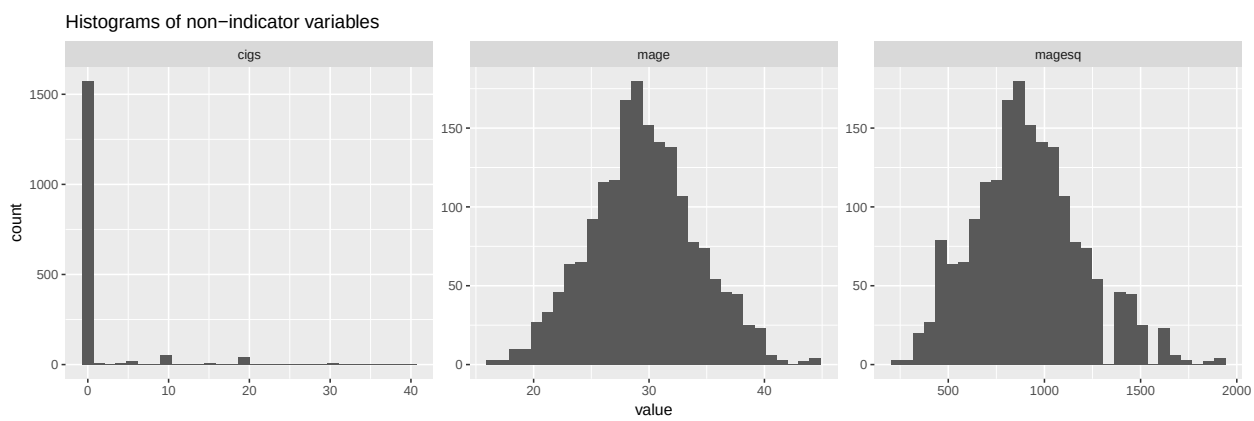
4.2 Model 2

MLR.1: Linear in Parameters

No changes effect MLR.1, as defined by the our linear model construction.

MLR.2: Random Sampling

```
ggplot(melt(bwdt[,.(cigs,mage,magesq)]), aes(x=value)) +
    geom_histogram(bins=30) +
    facet_wrap(~variable, scales='free') +
    ggtitle('Histograms of non-indicator variables')
```




```
bwdt[order(cigs),.N,by=cigs>0]
```

```
##     cigs    N
## 1: FALSE 1575
## 2:  TRUE  147
## 3:    NA  110
```

Histograms of the `mage` and `magesq` variables look approximately normal. The `cigs` variable is mostly 0, with a marginal number of positive cases. Omitting `NA` cases of `cigs` shows about ~10% of mothers smoked at least 1 cigarette per day.

MLR.3: No Perfect Collinearity

```
round(cor(bwdt[,.(bwght,npvis,npvissq,monpre,cigs,mage,magesq)], use='p'), 5)
```

```
##            bwght    npvis  npvissq   monpre     cigs     mage   magesq
## bwght    1.00000  0.10039  0.07268 -0.00101 -0.08499  0.03368  0.02417
## npvis    0.10039  1.00000  0.93410 -0.30610 -0.03872  0.10209  0.09650
## npvissq  0.07268  0.93410  1.00000 -0.17992  0.00517  0.06066  0.05688
## monpre  -0.00101 -0.30610 -0.17992  1.00000  0.09792 -0.18360 -0.16543
## cigs    -0.08499 -0.03872  0.00517  0.09792  1.00000 -0.05656 -0.05208
## mage     0.03368  0.10209  0.06066 -0.18360 -0.05656  1.00000  0.99399
## magesq   0.02417  0.09650  0.05688 -0.16543 -0.05208  0.99399  1.00000
```

```
sqrt(vif(mod2b))
```

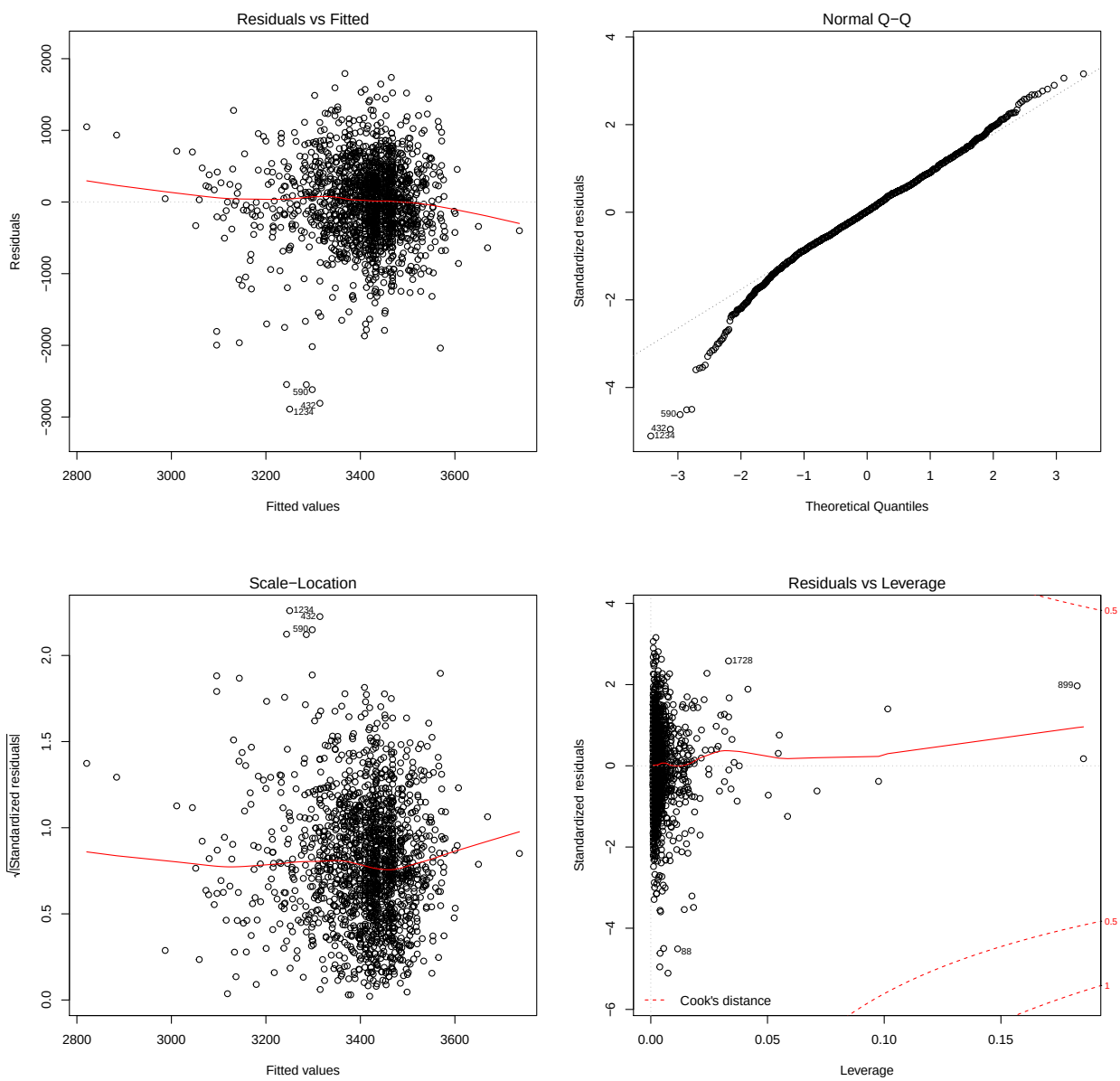
```
##    npvis  npvissq   monpre     cigs     mage   magesq
## 3.086155 2.980702 1.141388 1.010453 9.296427 9.263756
```

Aside from the expected correlation between `mage` and `magesq`, `npvis` and `npvissq`, there's no evidence of strong multicollinearity with the addition of `cigs`, `mage`, and `magesq`. The variance inflation factor of adding `cigs` is ~1%.

MLR.4: Zero Conditional Mean

```
par(mfrow=c(2,2)) # plot diagnostics on 2x2

plot(mod2b)
```



There's no major difference in the residuals vs fitted values plot compared to model 1. The smoothing line still shows an approximate distribution of residuals about zero, with the same small grouping of values toward -3000. The distribution of residuals on the lower and higher ends of fitted values seem to get closer to 0,

which is desirable compared to model 1.

MLR.5: Homoskedasticity

```r
bptest(mod2b)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod2b
## BP = 41.404, df = 6, p-value = 2.41e-07
```

The Scale-Location plot also looks similar to model 1. The smoothing line is fairly consistent with a minor uptick at higher fitted values (x-axis). The Breusch-Pagan test results in $p < 0.05$, rejecting the null hypothesis (homoskedasticity), however the strong statistical significance is likely due to $n = 1832$ observations.

MLR.6: Normality

```r
shapiro.test(mod2b$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod2b$residuals
## W = 0.98046, p-value = 3.088e-14
```

The Normal Q-Q plot shows deviation from normal at lower theoretical quantiles, and similar to the Breusch-Pagan test, the high number of observations influences the Shapiro-Wilks test and rejects normality with very high significance. No changes from model 1 and $n = 1832$ invokes CLT.

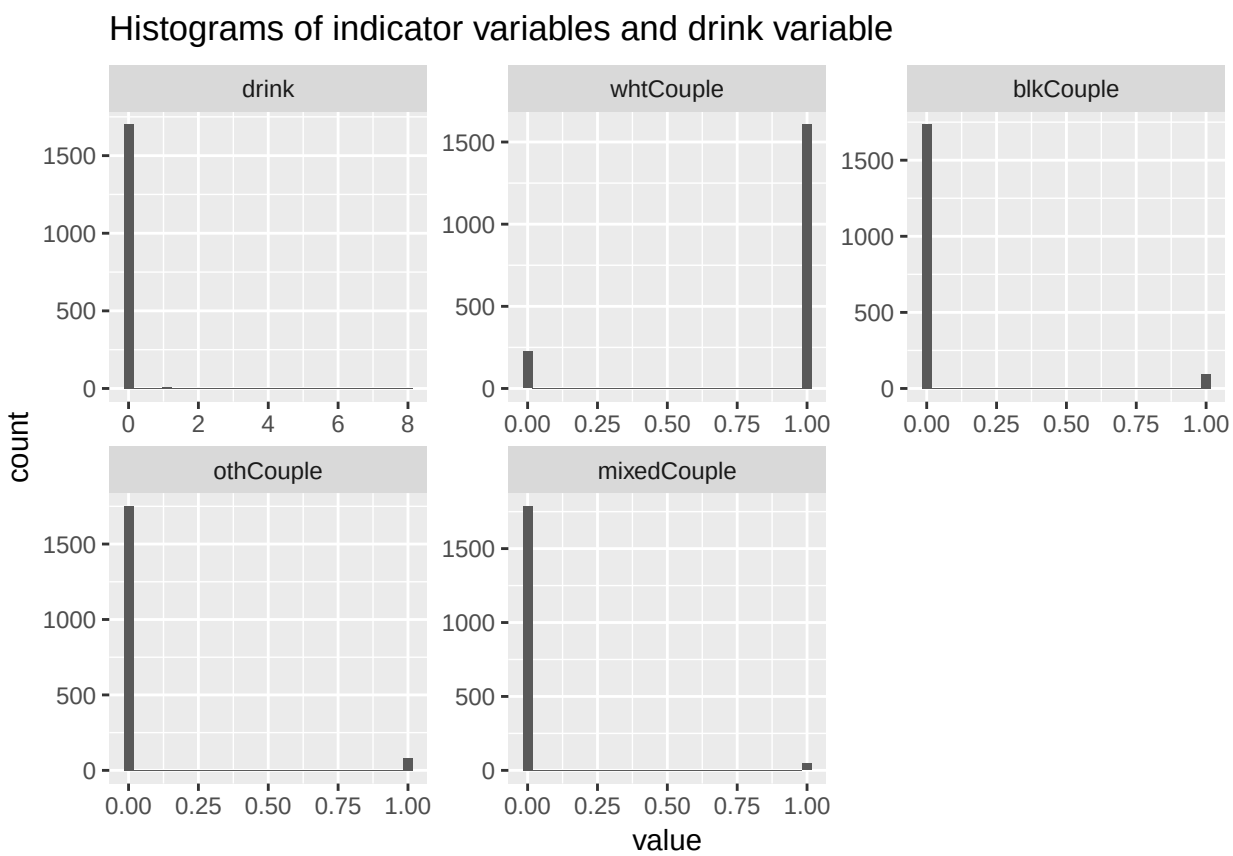
4.3 Model 3

MLR.1: Linear in Parameters

The parameters as observed in model 1 are all linear, the variables for the model are both dummy coded and contineous variables but the model parameters are still linear.

MLR.2: Random Sampling

```
ggplot(melt(bwdt[,.(drink,whtCouple,blkCouple,othCouple,mixedCouple)]), aes(x=value)) +
  geom_histogram(bins=30) +
  facet_wrap(~variable, scales='free') +
  ggtitle('Histograms of indicator variables and drink variable')
```




```
## The distribution of number of drinks is as follows
  table(bwdt$drink)
```

```
##
##    0    1    2    5    6    8
## 1701   11    2    1    1    1
```

```
  data.frame(numWht=sum(bwdt$whtCouple),numMixed=sum(bwdt$mixedCouple),numBlk=sum(bwdt$blkCouple),num
```

```
##   numWht numMixed numBlk numOth
## 1   1607       46     97     82
```

The table shows the number of different types of couples part of the study dataset. We can see that the sampling of different types of couples is extremely skewed. With most of the observations in the dataset of

white couples, while the minorites are extremely under represented.

Also the drinks variables has an extremely skewed sampling. Where most of the mothers are non drinkers. With such a small observation set for drinker, it is very difficult to analyze the effect of drinking on the birthweight and infant's health.

So for this model atleast which includes the race of the infant ($ White$, *black*, *other* and *mixed*) and the drinking habits of the mother(only 16 samples of mother having drinks) the sampling does not seem to be random. One can argue that the sampling of drink variable may be random as most of the mothers stop drinking during pregnancy, but still only 16 observations out of 1832 observations (which are not even normally distributed) is a small number for the model to be able to quantify the effects of drinking to infant's health and birthweight

MLR.3: No Perfect Collinearity

```
x<-data.frame(cor(bwdt[,.(bwght,npvis,npvissq,monpre,cigs,mage,magesq,drink,mixedCouple,
                          whtCouple,blkCouple)], use='pairwise.complete.obs'))
x[,c("drink","mixedCouple","whtCouple","blkCouple")]
```

```
##                    drink mixedCouple   whtCouple   blkCouple
## bwght       -0.019905817  0.01902665  0.05382909 -0.02011734
## npvis        0.052721312 -0.03563146  0.02162652 -0.02724284
## npvissq      0.050408001 -0.03564506  0.02486149 -0.01935900
## monpre      -0.009718353  0.06076286 -0.04877962  0.03595308
## cigs         0.181951798  0.01867295  0.03980284 -0.02190966
## mage         0.004071960  0.03316323  0.01761415 -0.07620584
## magesq       0.007123435  0.04147281  0.01219209 -0.07115074
## drink        1.000000000 -0.01137158  0.02565423 -0.01573810
## mixedCouple -0.011371581  1.00000000 -0.42889884 -0.03794673
## whtCouple    0.025654226 -0.42889884  1.00000000 -0.63190648
## blkCouple   -0.015738104 -0.03794673 -0.63190648  1.00000000
```

```
sqrt(vif(mod3b))
```

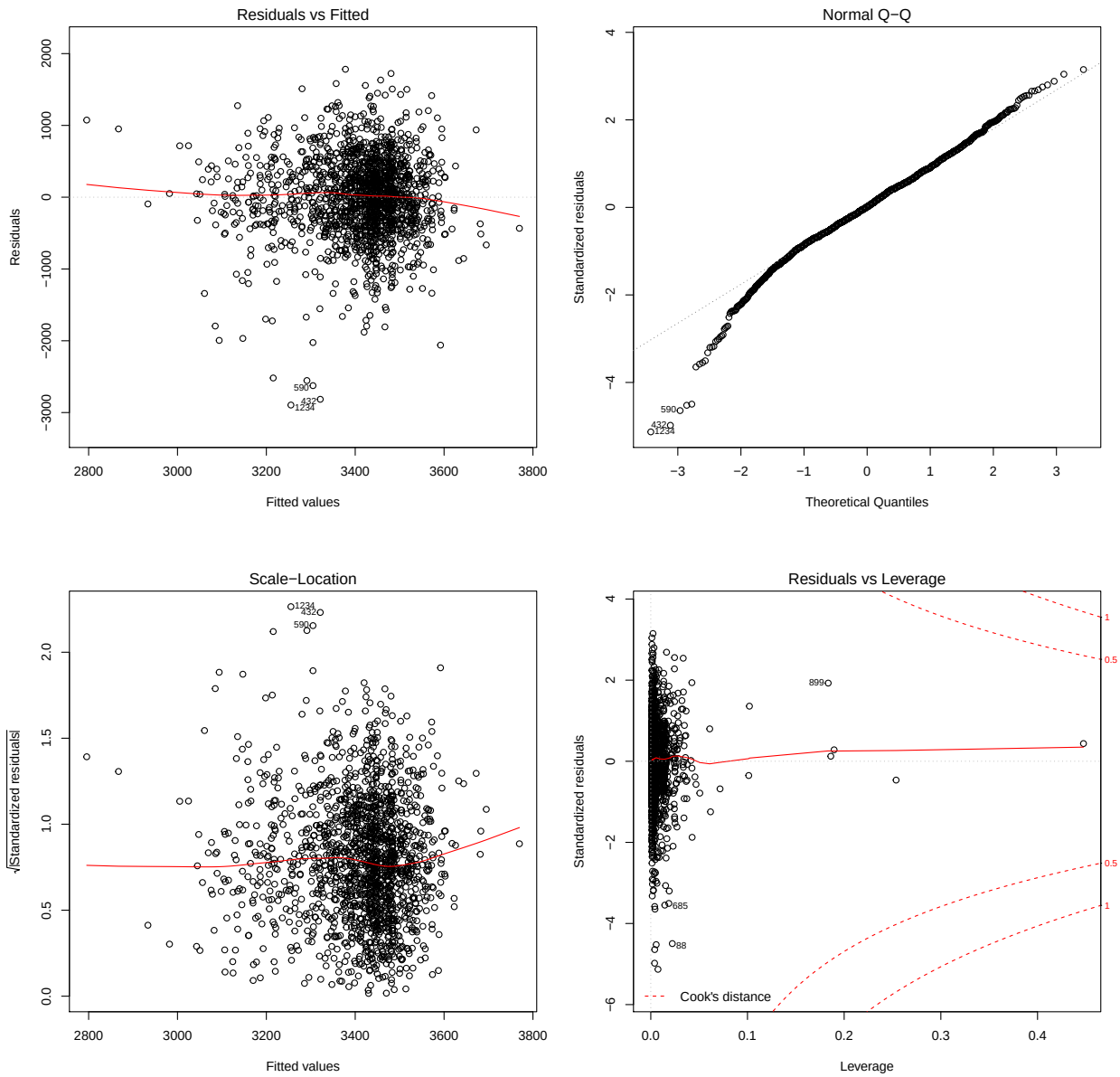
```
##       npvis     npvissq      monpre        cigs        mage      magesq
##    3.086081    2.979162    1.143538    1.031202    9.306584    9.271348
##       drink mixedCouple   whtCouple   blkCouple
##    1.020327    1.236905    1.568068    1.414963
```

As expected the couple variables show high collinearity with respective mother's and father's race as expected as the couple variables are derived from the respective father's and mother's race, as the correlation is greater than 0.9, So,we do not include them in a linear model. But none of the added variables show perfect collinearity with any of the variables.

MLR.4: Zero Conditional Mean

```
par(mfrow=c(2,2)) # plot diagnostics on 2x2

plot(mod3b)
```



There's no major difference in the residuals vs fitted values plot compared to model 1. The smoothing line still shows an approximate distribution of residuals about zero, with the same small grouping of values toward -3000.

MLR.5: Homoskedasticity

```r
bptest(mod3b)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3b
## BP = 41.922, df = 10, p-value = 7.743e-06
```

The Scale-Location plot also looks similar to model 1. The smoothing line is fairly consistent with a minor uptick at higher fitted values (x-axis). The Breusch-Pagan test results in $p < 0.05$, rejecting the null hypothesis (homoskedasticity), however the strong statistical significance is likely due to $n = 1832$ observations. Which is similar to the first and second model. This must be same because the dependent variable is the same for all the models

MLR.6: Normality

```r
shapiro.test(mod3b$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod3b$residuals
## W = 0.9795, p-value = 1.341e-14
```

The Q-Q plot is not on the diagonal and has tails. But the sample size is relatively large and thus assumtions 1 to 5 are sufficient for OLS regression as CLT can be invoked. The shapiro-wilks test usally tests positive (significant) for large samples.

5 Model Summary

```r
se.mod1b = sqrt(diag(vcovHC(mod1b)))
se.mod2b = sqrt(diag(vcovHC(mod2b)))
se.mod3b = sqrt(diag(vcovHC(mod3b)))

stargazer(mod1b, mod2b, mod3b, type = "latex", header=F,
          se = list(se.mod1b, se.mod2b, se.mod3b),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```r
(AIC(mod1b)-AIC(mod2b))/AIC(mod1b)
```

```
## [1] 0.0628228
```

```r
(AIC(mod2b)-AIC(mod3b))/AIC(mod2b)
```

```
## [1] 0.005031949
```

Using heteroskedasticity-robust standard errors, we find $p < 0.05$ statistical significance across all parameters in model 1, and all but `npvissq` coefficient in model 2. The addition of `drink` and parent race variables prevents `vcovHC()` calculation of standard errors for model 3 so we cannot comment on statistical signifiance, other than the relative drop in AIC from model 1 to 2 was 6.2%, whereas from model 2 to 3 was only 0.5%.

In terms of practical signifiance, model 1 shows that the `npvis` parameter has the greatest effect on birthweight (50 grams per prenatal care visit, ignoring the quadratic term for simplicity). The addition of `mage` and `cigs` in model 2 has a drastic effect on both the `npvis` parameter (50 drops to 40 grams per visit) and the intercept term ~2900 grams falls to ~1800 grams with zero prenatal vists. In addition, model 2 shows mother's age

Table 1:

	Dependent variable:		
	bwght		
	(1)	(2)	(3)
npvis	50.703**	40.921*	42.785**
	(16.083)	(16.022)	(15.978)
npvissq	−1.114*	−0.873	−0.902
	(0.512)	(0.499)	(0.494)
monpre	29.008*	33.289*	34.335*
	(13.452)	(13.799)	(13.822)
cigs		−10.019**	−10.983**
		(3.308)	(3.660)
mage		76.358**	78.253**
		(29.551)	(29.673)
magesq		−1.209*	−1.236*
		(0.485)	(0.487)
drink			−20.120
			(31.038)
mixedCouple			350.661***
			(95.192)
whtCouple			263.142***
			(57.016)
blkCouple			232.648**
			(83.270)
Constant	2,916.155***	1,830.883***	1,530.224**
	(139.921)	(466.515)	(470.558)
Observations	1,763	1,655	1,647
R^2	0.017	0.024	0.035
Adjusted R^2	0.015	0.021	0.029
Residual Std. Error	576.052 (df = 1759)	568.062 (df = 1648)	566.493 (df = 1636)
F Statistic	9.925*** (df = 3; 1759)	6.807*** (df = 6; 1648)	5.956*** (df = 10; 1636)
Note:			*p<0.05; **p<0.01; ***p<0.001

mage having a stronger effect than npvis. Proper ceteris paribus interpretation requires accounting for the quadratic term, which will be covered in the causal interpretation. The mage coefficient cannot be zero in reality, so reasonable interpretation of the importance of npvis can't be based on the intercept term alone. The negative value cigs parameter agrees with the expected negative impact on infant health.

The addition of drink and race variables has marginal effect on the npvis, mage, monpre, and cigs parameters, which increase in magnitude slightly. The greater effect on practical significance occurs for the race-modification of the intercept term. The intercept term cannot be interpreted by itself, because one of the *Couple variables will always have a value of 1. The minimum infant weight for white couples is very close to the model 2 estimate of the intercept, likely due to the high majority of observations being white couples. Mixed couples tend to produce infants that are 90 grams heavier than white couples, while black couples' infants are 30 grams lighter.

6 Discussion

6.1 Causal Interpretation

```r
# ceteris paribus effect of npvis on birthweight, model 1
mod1b$coefficients[2]/(2*abs(mod1b$coefficients[3]))
```

```
##    npvis
## 22.74891
```

```r
# ceteris paribus effect of npvis on birthweight, model 2
mod2b$coefficients[2]/(2*abs(mod2b$coefficients[3]))
```

```
##    npvis
## 23.42751
```

```r
# predict the effect of a single prenatal visit on birthweight
# using the lowest sample values of monpre, cigs, mage
predict(mod2b, newdata=data.frame(npvis=1,
                                  npvissq=1,
                                  monpre=min(bwdt$monpre, na.rm=T),
                                  cigs=min(bwdt$cigs, na.rm=T),
                                  mage=min(bwdt$mage, na.rm=T),
                                  magesq=min(bwdt$magesq, na.rm=T)))
```

```
##        1
## 2783.188
```

Including the quadratic term npvissq accounts for the diminishing effect of the number of prenatal vists npvis on birthweight bwght. Model 1 shows a negative return on birthweight after 23 vists, while model 2 is slightly higher at 24 vists. Taken as is, model 1 cannot predict low birthweights because the intercept term ~2900 grams is greater than low birthweight threshold of 2500 grams. While the intercept term in model 2 is below the low birthweight threshold, it was shown that even with minimum values of monpre, cigs, and mage, the effect of one prenatal visit already yields a birthweight value over 2500 grams, i.e. a healthy infant.

Ultimately, we do not have a controlled experiment where the number of prenatal visits can be manipulated as a treatment variable, nor would it be ethical to limit the number of prenatal vists during pregnancy as part of a research design. A causal framework requires an exogenous treatment variable, uncorrelated with all other independent variables, such that only changes in the treatment variable will affect the outcome. For example, the number of prenatal visits may be correlated with omitted variables such as the mother's health or medical conditions not related to cigarette usage or consumption of alcohol.

The lack of randomization yields a significant reservation on the external validity of causal inference. As shown in our earlier EDA of indicator variables, over 87% of observations belong to white male and female

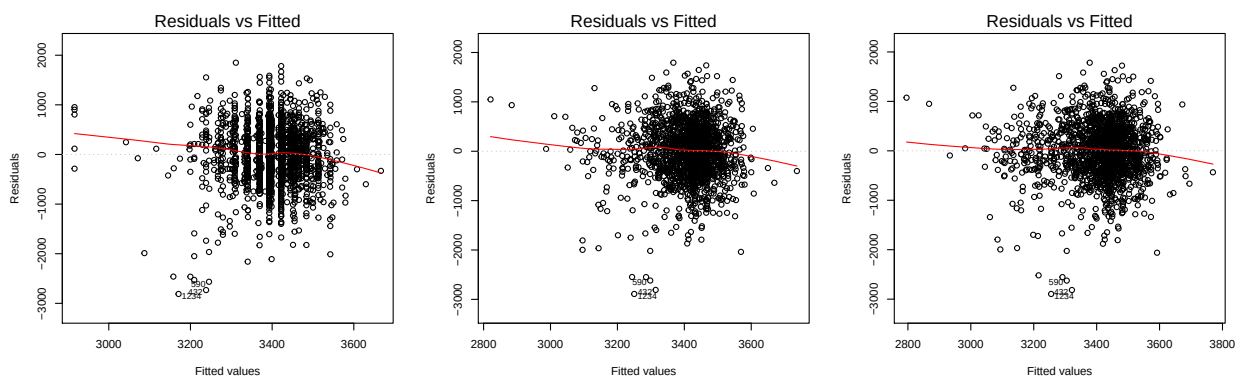
parents. Only 5.3% of observations pertain to black male and female parents, while 4.4% pertain to "other". The generalizability of a causal inference from this sample is limited.

An argument for these models to be causal would only hinge on the fact that we have precluded post-birth variables: the APGAR scores. However, all other evidences as described above override this argument. Given those limitations, the model developed in our analysis is associative rather than causal.

6.2 Omitted Variable Bias

```
par(mfcol=c(1,3))

plot(mod1b, 1)
plot(mod2b, 1)
plot(mod3b, 1)
```




```
#           Model 1                    Model 2                    Model 3
```

A recurring theme throughout all 3 of our models is that the loess in residuals v. fitted values show a somewhat negative slope. It is increasingly difficult to quantitatively analyze and forecast the effect of omitted variable bias as the model goes higher in dimensionality. Since this effect is present in all models, however, we will look at model 1. The coefficients in model 1 are positive with the exception of diminishing return variable `npvissq`, but we are suspicious that the coefficients given by our specification may have been too small on average. This is an indication that we may have omitted variables that are either:

- Negatively affecting birthweight but are positively correlated with number of visits and months between prenatal care start and birth, for example:
  - Parents' blood/physiology test results of relevant markers
  - Count of prior premature delivery or miscarriages
  - Other indicators of presence of hereditary (genetic or epigenetic) disorders or any pre-existing conditions that increased parents' concern of the pregnancy, inducing frequent visits
- Positively affecting birthweight but are negatively correlated with number of visits and months between prenatal care start and birth, for example:
  - Count of mother's previous successful births
  - Average prior births' birthweight
  - Parents' APGAR/birthweights
  - Indicators of earlier successes or inherently healthy pregnancy

Other factors could also fall into the category above, or has less straightforward interaction with the variables that we included.

Infant's health can be dependent on various factors other than race and smoking/drinking habits. Some of

the most important variables which would affect the infant's health may be regarding the mother's health condition. If the mother has been diagnosed with lifestyle diseases in particular like high blood pressure and diabetes.

Similarly the daily work hour, if the mother is working and number of hours spent in travelling could be some minor but essential variables which could improve the model performance, and could also aid in causal interpretation of the model.

6.3 Included Variable Bias

Considering model 1, the only additional variable besides `npvis` is `monpre` the month prenatal care began. We expect `monpre` to correlate negatively with `npvis` since higher values of `monpre` equates to less time available for prenatal visits. The correlation of `monpre` with `bwght`, our proxy for infant health, is unclear. Late start to prenatal care could be due to neglect, in which case the correlation is negative, or due to a scarcity of issues during pregnancy, in which case the correlation is positive. Executing lm() on the restricted model (omitting `monpre`) reveals a negative bias on `npvis` and a small negative correlation between `monpre` and `bwght`.

In model 2, the `cigs` variable should correlate negatively with `bwght` and positively with `npvis` as smoking is a detriment to the mother's health and the health of the infant. Unhealthy mothers may encounter complications which prompt more prenatal visits. Consequently, `cigs` should positively correlate with `npvis`. The bias on `npvis` due to omitting `cigs` might positive. However this is not the case evaluating the restricted model. The bias due to `cigs` is actually positive because `cigs` negatively correlates with `npvis`. This result is reasonable considering that smoking mothers may be less educated or more ignorant about their health and that of the baby's, so they may attend fewer prenatal visits.

As for `mage`, it's not intuitive that the mother's age would correlate positively or negatively with either `npvis` or `bwght`. Evaluating the restricted model without `mage` shows little change in the coeffcient for `npvis`.

7 Conclusion

The supplied dataset is quite limited in the variation of APGAR scores, which we believe is the appropriate metric of infant health. As a proxy, we consider birthweight relative to the low and very low birthweight thresholds. The models we've constructed are associative in relating the number of prenatal vists to birthweight, due to endogeneity of the `npvis` variable with factors in the error term such as mother's general health and stress levels. Although we've leverged the greater variation in birthweight to better fit the OLS assumptions, it's still the case that the sample provided in this dataset consists mainly of healthy infants. The number of prenatal vists has a positive association with birthweight, but the mother's age is also positive with even greater magnitude. There are significant ethical considerations to performing a causality study of the effect of prenatal vists on infant health, which parallels the Tuskegee Syphillis debacle. We also found some pointers for possible improvements of our models, by acquiring data on currently unavailable variables that we have discussed in more detail in our report.