Supplement for Lecture 18: Coding Categorical Predictors

Load Data

```
data("NCbirths")

NCB = NCbirths[,c("BirthWeightOz","Weeks","Plural","MomRace")]

str(NCB)
```

```
## 'data.frame':    1450 obs. of  4 variables:
##  $ BirthWeightOz: int  111 116 138 136 121 117 143 113 120 124 ...
##  $ Weeks        : int  40 37 39 39 39 43 39 42 39 40 ...
##  $ Plural       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MomRace      : Factor w/ 4 levels "black","hispanic",..: 4 4 4 4 4 4 4 4 4 4 ...
```

Models Based Only on Mother's Race

```
#Fit Model with Only MomRace as Predictor Variable
mod.race.1 = lm(BirthWeightOz ~ MomRace,data=NCB)

#Notice that Black is the Current Reference Category
summary(mod.race.1)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ MomRace, data = NCB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.872  -10.872    1.482   13.966   63.128
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      110.563      1.215  91.022 < 0.0000000000000002 ***
## MomRacehispanic    7.955      2.112   3.766             0.000173 ***
## MomRaceother       6.583      3.418   1.926             0.054298 .
## MomRacewhite       7.309      1.420   5.147          0.000000301 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.13 on 1446 degrees of freedom
## Multiple R-squared:  0.01938,    Adjusted R-squared:  0.01735
## F-statistic: 9.528 on 3 and 1446 DF,  p-value: 0.000003118
```

```
#Confidence Intervals for Difference in Mean Birth Weight Between Each Other Race and Black
confint(mod.race.1)
```

```
##                       2.5 %    97.5 %
## (Intercept)     108.1805115 112.94599
## MomRacehispanic   3.8112678  12.09881
## MomRaceother     -0.1216373  13.28680
## MomRacewhite      4.5234042  10.09402
```

```
#Representation of Race Groups
table(NCB$MomRace)
```

```
##
##    black hispanic    other    white
##      332      164       48      906
```

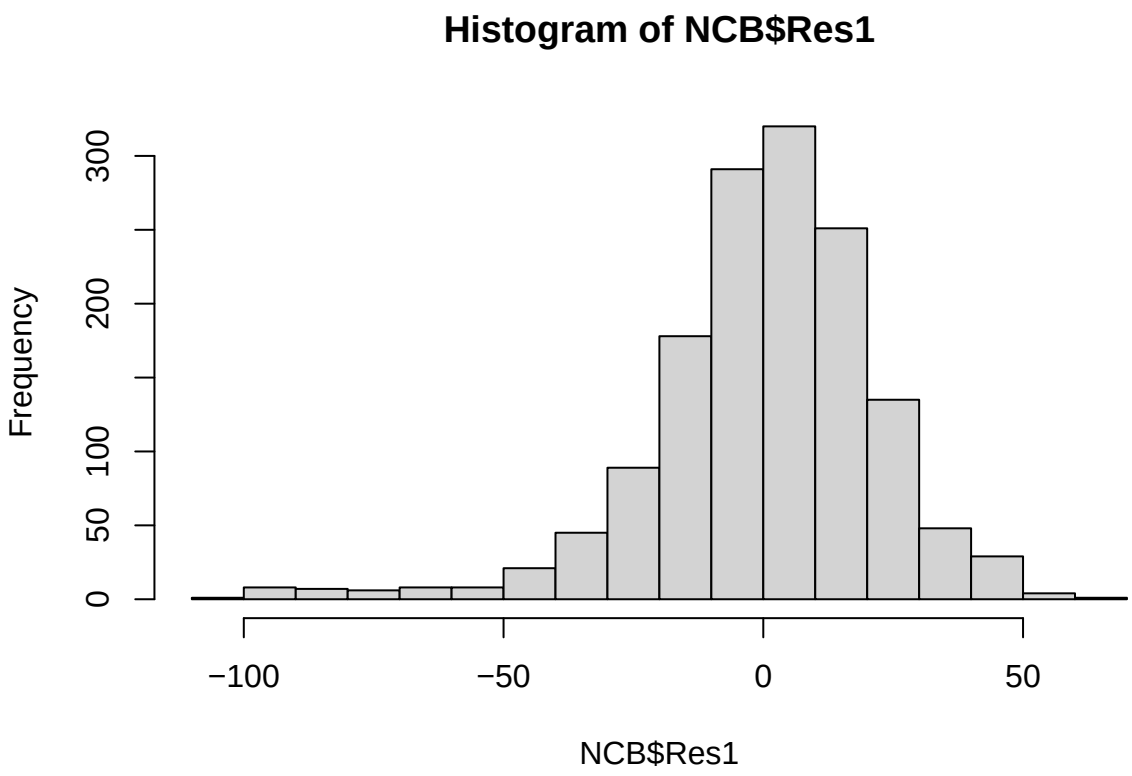
```
#Make Reference Category the Majority: We Can Do This Easily Since MomRace is Factor Variable
NCB$MomRace = relevel(NCB$MomRace,ref=4)

#Refit Model (Compare to Exercise 4.13 in Textbook)
mod.race.2 = lm(BirthWeightOz ~ MomRace,data=NCB)
summary(mod.race.2)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ MomRace, data = NCB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.872  -10.872    1.482   13.966   63.128
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     117.8720     0.7353 160.303 < 0.0000000000000002 ***
## MomRaceblack     -7.3087     1.4199  -5.147          0.000000301 ***
## MomRacehispanic   0.6463     1.8782   0.344                0.731
## MomRaceother     -0.7261     3.2781  -0.222                0.825
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.13 on 1446 degrees of freedom
## Multiple R-squared:  0.01938,    Adjusted R-squared:  0.01735
## F-statistic: 9.528 on 3 and 1446 DF,  p-value: 0.000003118
```

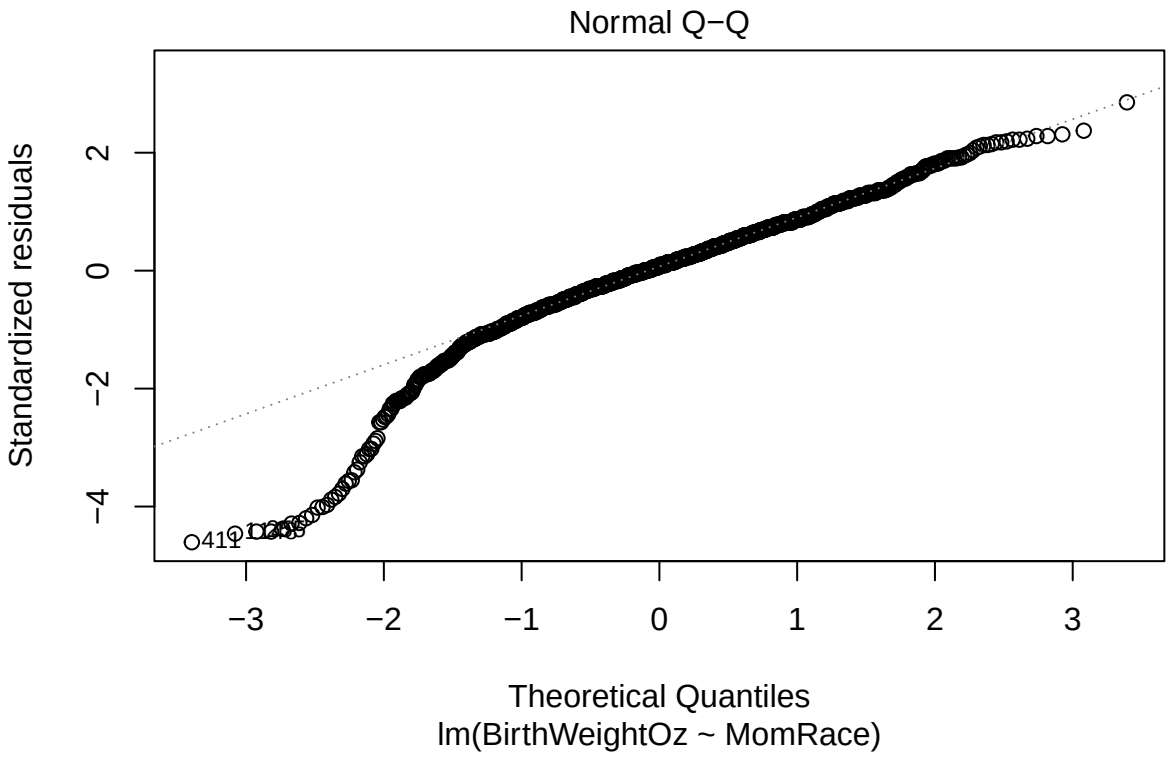
```
confint(mod.race.2)
```

```
##                       2.5 %     97.5 %
## (Intercept)     116.429577 119.314352
## MomRaceblack    -10.094019  -4.523404
## MomRacehispanic  -3.037945   4.330601
## MomRaceother     -7.156494   5.704231
```

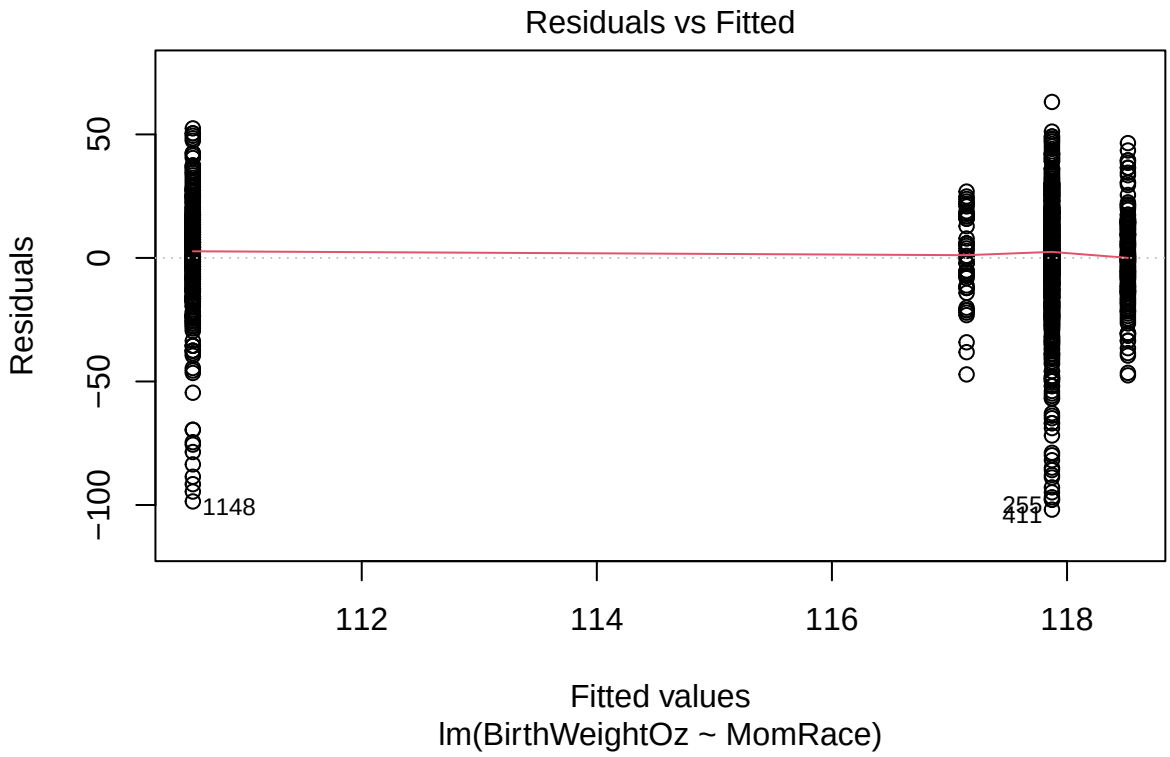
```
#Add Predictions and Residuals to Data
NCB$Pred1 = fitted(mod.race.2)
NCB$Res1 = residuals(mod.race.2)

#Check Assumptions
hist(NCB$Res1) #Normality?
```



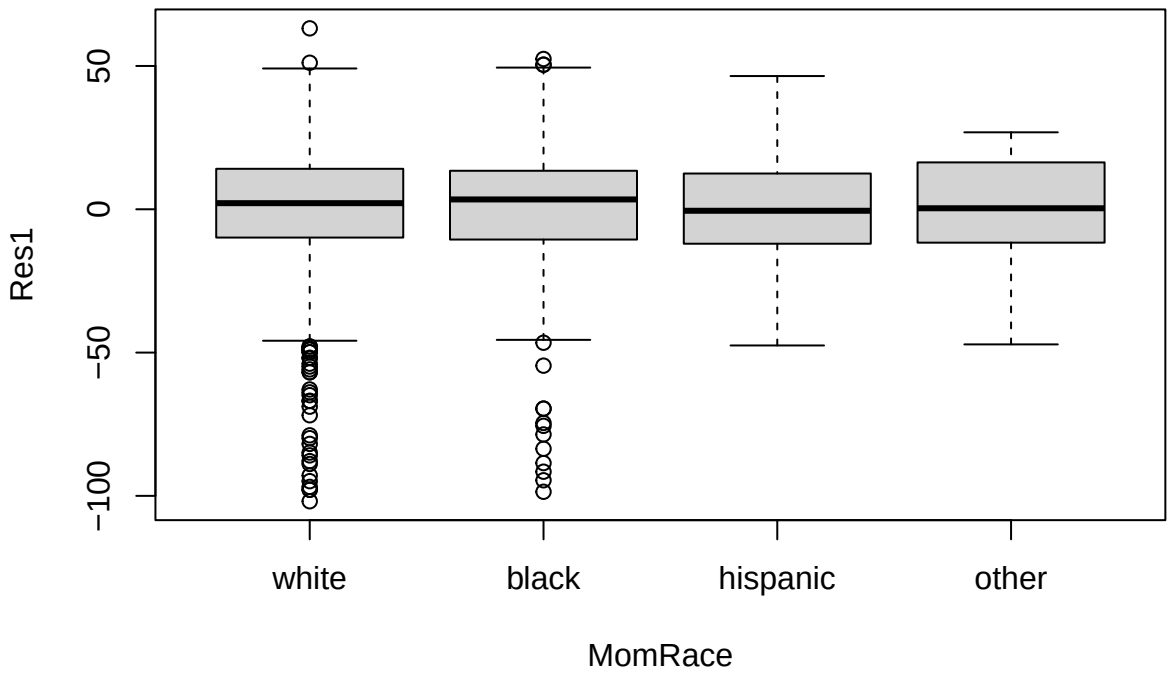

```
plot(mod.race.2,which=c(2)) #Normality?
```



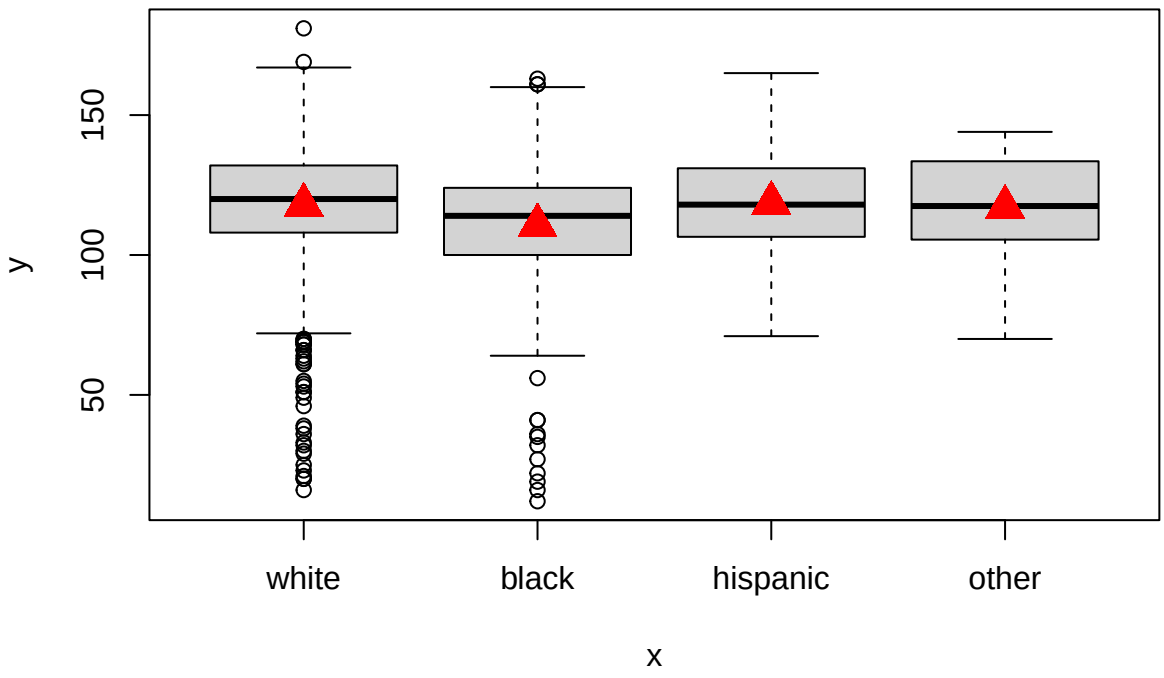

```
plot(mod.race.2, which=c(1)) #Homoscedasticity?
```




```
boxplot(Res1~MomRace,data=NCB) #Homoscedasticity?
```




```
#Plot of Model (Defaults to Boxplots since x is a Factor Variable)
plot(x=NCB$MomRace,y=NCB$BirthWeightOz)
points(NCB$MomRace,NCB$Pred1,col="red",pch=17,cex=2)
```



Model Including Weeks

```
#Fit Model with Only Weeks as Predictor Variable
mod.weeks.1 = lm(BirthWeightOz ~ Weeks,data=NCB)
summary(mod.weeks.1)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Weeks, data = NCB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.480 -11.994  -0.286  11.908  58.006
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -73.1171     6.7817  -10.78 <0.0000000000000002 ***
## Weeks         4.9028     0.1752   27.99 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.99 on 1447 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3512, Adjusted R-squared:  0.3508
## F-statistic: 783.4 on 1 and 1447 DF,  p-value: < 0.00000000000000022
```

```r
NCB$Pred2[!is.na(NCB$Weeks)] = fitted(mod.weeks.1)
NCB$Res2[!is.na(NCB$Weeks)] = residuals(mod.weeks.1)

#Fit Model with Weeks + MomRace as Predictor Variables
mod.weeks.2 = lm(BirthWeightOz~ Weeks + MomRace,data=NCB)
summary(mod.weeks.2)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Weeks + MomRace, data = NCB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.147 -11.578  -0.003  11.429  56.573
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     -69.8163     6.7636 -10.322 < 0.0000000000000002 ***
## Weeks             4.8561     0.1743  27.857 < 0.0000000000000002 ***
## MomRaceblack     -5.7543     1.1482  -5.012          0.000000606 ***
## MomRacehispanic  -0.9935     1.5163  -0.655                0.512
## MomRaceother     -2.1217     2.6449  -0.802                0.423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.85 on 1444 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3624, Adjusted R-squared:  0.3606
## F-statistic: 205.2 on 4 and 1444 DF,  p-value: < 0.00000000000000022
```

```r
NCB$Pred3[!is.na(NCB$Weeks)] = fitted(mod.weeks.2)
NCB$Res3[!is.na(NCB$Weeks)] = residuals(mod.weeks.2)

#Fit Model to Include Interaction Variable for Slope
mod.weeks.3 = lm(BirthWeightOz~ Weeks + MomRace + Weeks*MomRace,data=NCB)
summary(mod.weeks.3)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Weeks + MomRace + Weeks * MomRace,
##     data = NCB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -69.634 -11.625  -0.314  11.380  56.375
##
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           -75.4908     8.7353  -8.642 < 0.0000000000000002 ***
## Weeks                   5.0029     0.2255  22.186 < 0.0000000000000002 ***
## MomRaceblack          -15.5496    14.7501  -1.054                0.292
## MomRacehispanic       108.1458    27.0683   3.995            0.0000679 ***
## MomRaceother           47.8224    43.6001   1.097                0.273
## Weeks:MomRaceblack      0.2569     0.3827   0.671                0.502
```

```
## Weeks:MomRacehispanic  -2.8006     0.6938  -4.036           0.0000571 ***
## Weeks:MomRaceother     -1.2838     1.1180  -1.148               0.251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.75 on 1441 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3709, Adjusted R-squared:  0.3679
## F-statistic: 121.4 on 7 and 1441 DF,  p-value: < 0.00000000000000022
```

```
NCB$Pred4[!is.na(NCB$Weeks)] = fitted(mod.weeks.3)
NCB$Res4[!is.na(NCB$Weeks)] = residuals(mod.weeks.3)

#Visualize mod.weeks.3
library(HelpersMG) #Helpful Function Called plot_add()
```

```
## Warning: package 'HelpersMG' was built under R version 4.2.3
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.2.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Loading required package: rlang
```

```
## Warning: package 'rlang' was built under R version 4.2.3
```

```
## Loading required package: coda
```

```
## Warning: package 'coda' was built under R version 4.2.3
```

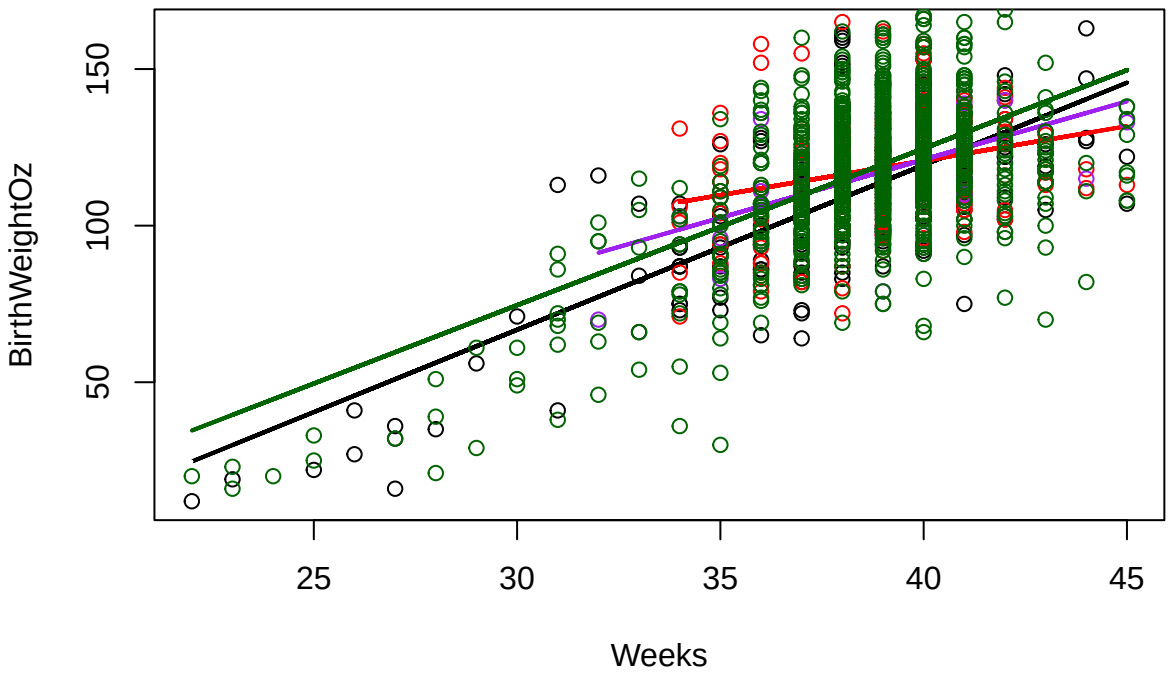
```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.3
```

```
## Welcome in package HelpersMG version 6.0.3
```

```
## No update is available
```

```
plot(BirthWeightOz ~ Weeks , data=subset(NCB,MomRace=="black"))
plot_add(Pred4 ~ Weeks , data=subset(NCB,MomRace=="black"),type="l",lwd=2)
plot_add(BirthWeightOz ~ Weeks , data=subset(NCB,MomRace=="hispanic"),col="red")
plot_add(Pred4 ~ Weeks , data=subset(NCB,MomRace=="hispanic"),type="l",col="red",lwd=2)
plot_add(BirthWeightOz ~ Weeks , data=subset(NCB,MomRace=="other"),col="purple")
plot_add(Pred4 ~ Weeks , data=subset(NCB,MomRace=="other"),type="l",col="purple",lwd=2)
plot_add(BirthWeightOz ~ Weeks , data=subset(NCB,MomRace=="white"),col="darkgreen")
plot_add(Pred4 ~ Weeks , data=subset(NCB,MomRace=="white"),type="l",col="darkgreen",lwd=2)
```



Model Including Plural

```
#Create Indicator Variables Manually for Plural Variable
NCB$Twins = ifelse(NCB$Plural==2,1,0)
NCB$Triplets = ifelse(NCB$Plural==3,1,0)
NCB$Plural=NULL

NCB2 =na.omit(NCB)

#Stepwise Regression on Full Model
mod.full = lm(BirthWeightOz ~ Weeks + MomRace + Twins + Triplets +
                  Weeks*MomRace+Weeks*Twins+Weeks*Triplets,data=NCB2)

mod.empty = lm(BirthWeightOz ~ 1,data=NCB2)

mod.step = step(mod.empty,scope=list(upper=mod.full),scale=summary(mod.empty)$sigma,direction="both")
```

```
## Start:  AIC=30885.25
## BirthWeightOz ~ 1
##
##             Df Sum of Sq    RSS    Cp
## + Weeks      1    253569 468374 19531
## + Twins      1     71651 650293 27678
## + MomRace    3     14261 707683 30253
## + Triplets   1     11581 710362 30369
```

```
## <none>                  721944 30885
##
## Step:  AIC=19531.15
## BirthWeightOz ~ Weeks
##
##            Df Sum of Sq    RSS    Cp
## + Twins     1     16107 452267 18812
## + MomRace   3      8067 460307 19176
## + Triplets  1      1449 466926 19468
## <none>                  468374 19531
## - Weeks     1    253569 721944 30885
##
## Step:  AIC=18811.8
## BirthWeightOz ~ Weeks + Twins
##
##               Df Sum of Sq    RSS    Cp
## + MomRace      3      9051 443216 18412
## + Triplets     1      2006 450262 18724
## + Weeks:Twins  1        60 452207 18811
## <none>                     452267 18812
## - Twins        1     16107 468374 19531
## - Weeks        1    198025 650293 27678
##
## Step:  AIC=18412.45
## BirthWeightOz ~ Weeks + Twins + MomRace
##
##                 Df Sum of Sq    RSS    Cp
## + Weeks:MomRace  3      5018 438198 18194
## + Triplets       1      2406 440810 18307
## <none>                       443216 18412
## + Weeks:Twins    1        17 443200 18414
## - MomRace        3      9051 452267 18812
## - Twins          1     17091 460307 19176
## - Weeks          1    191876 635093 27004
##
## Step:  AIC=18193.7
## BirthWeightOz ~ Weeks + Twins + MomRace + Weeks:MomRace
##
##                 Df Sum of Sq    RSS    Cp
## + Triplets       1    2277.3 435921 18094
## <none>                       438198 18194
## + Weeks:Twins    1      19.6 438178 18195
## - Weeks:MomRace  3    5018.3 443216 18412
## - Twins          1   15943.5 454142 18906
##
## Step:  AIC=18093.71
## BirthWeightOz ~ Weeks + Twins + MomRace + Triplets + Weeks:MomRace
##
##                   Df Sum of Sq    RSS    Cp
## + Weeks:Triplets  1     140.2 435781 18089
## <none>                        435921 18094
## + Weeks:Twins     1       6.6 435914 18095
## - Triplets        1    2277.3 438198 18194
## - Weeks:MomRace   3    4889.3 440810 18307
```

```
## - Twins           1   16624.3 452545 18836
##
## Step:  AIC=18089.43
## BirthWeightOz ~ Weeks + Twins + MomRace + Triplets + Weeks:MomRace +
##     Weeks:Triplets
##
##                  Df Sum of Sq    RSS    Cp
## <none>                        435781 18089
## + Weeks:Twins     1       5.3 435775 18091
## - Weeks:Triplets  1     140.2 435921 18094
## - Weeks:MomRace   3    4877.4 440658 18302
## - Twins           1   16695.5 452476 18835
```

```
summary(mod.step)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Weeks + Twins + MomRace + Triplets +
##     Weeks:MomRace + Weeks:Triplets, data = NCB2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.536 -10.979  -0.273  11.250  56.250
##
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           -52.9607     9.1059  -5.816    0.000000007411705 ***
## Weeks                   4.4428     0.2342  18.970 < 0.0000000000000002 ***
## Twins                 -20.5871     2.7736  -7.422    0.000000000000196 ***
## MomRaceblack          -22.9688    14.5879  -1.575             0.115590
## MomRacehispanic        91.3229    26.6620   3.425             0.000632 ***
## MomRaceother           25.2924    42.8656   0.590             0.555257
## Triplets              -72.5065    71.3313  -1.016             0.309574
## Weeks:MomRaceblack      0.4373     0.3783   1.156             0.247874
## Weeks:MomRacehispanic  -2.3804     0.6833  -3.484             0.000509 ***
## Weeks:MomRaceother     -0.7236     1.0991  -0.658             0.510387
## Weeks:Triplets          1.5244     2.2412   0.680             0.496486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.41 on 1438 degrees of freedom
## Multiple R-squared:  0.3964, Adjusted R-squared:  0.3922
## F-statistic: 94.43 on 10 and 1438 DF,  p-value: < 0.00000000000000022
```