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)

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

#Representation of Race Groups
table(COMPLETE)

#Make Reference Category the Majority: We Can Do This Easily Since MomRace is Factor Variable
NCB$MomRace = COMPLETE

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

#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?
```

```r
#Plot of Model (Defaults to Boxplots since x is a Factor Variable)
plot(COMPLETE)
points(COMPLETE)
```

Model Including Weeks

```r
#Fit Model with Only Weeks as Predictor Variable
mod.weeks.1 = lm(BirthWeightOz ~ Weeks,data=NCB)
summary(mod.weeks.1)
NCB$Pred2 = fitted(mod.weeks.1) #FIX ERROR
NCB$Res2 = residuals(mod.weeks.1) #FIX ERROR

#Fit Model with Weeks + MomRace as Predictor Variables
mod.weeks.2 = lm(BirthWeightOz~ Weeks + MomRace,data=NCB)
summary(mod.weeks.2)
NCB$Pred3 = fitted(mod.weeks.2) #FIX ERROR
NCB$Res3] = residuals(mod.weeks.2) #FIX ERROR

#Fit Model to Include Interaction Variable for Slope
mod.weeks.3 = lm(BirthWeightOz~ Weeks + MomRace + Weeks*MomRace,data=NCB)
summary(mod.weeks.3)
NCB$Pred4 = fitted(mod.weeks.3) #FIX ERROR
NCB$Res4 = residuals(mod.weeks.3) #FIX ERROR

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

plot(BirthWeightOz ~ Weeks , data=subset(NCB,MomRace=="black"))
plot_add(Pred4 ~ Weeks , data=subset(NCB,MomRace=="black"),type="l",lwd=2)
COMPLETE
COMPLETE
```

Model Including Plural

```r
#Create Indicator Variables Manually for Plural Variable
NCB$Twins = COMPLETE
NCB$Triplets = COMPLETE
NCB$Plural=COMPLETE

NCB2 = COMPLETE

#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")
summary(mod.step)
```