Lab 8

This week, we'll start with a dataset from the 2012 US census. For now, we're going to focus on only three variables: income, gender, and education level. So, to make things simple, we'll subset the data to only include complete cases (no NAs) for these three variables.

```
require("openintro")
require("dplyr")
data(acs12)

#using some dplyr today! ...because, why not?!
acs12 = acs12 %>% select(income, gender, edu) %>% na.omit()

#note: this code would do the same as the dplyr code above:
#acs12 = acs12[,c("income","gender","edu")]
#acs12 = na.omit(acs12)
```

Now, suppose that we want to use regression in order to estimate mean income by gender and education level. We will start by estimating these means directly, and then we can see how a regression model matches up (or doesn't).

```
#First, note that gender is a factor with 2 levels and edu is a factor with 3 levels
str(acs12)
```

```
## 'data.frame':    1623 obs. of  3 variables:
##  $ income: int  60000 0 0 0 1700 45000 8600 0 0 33500 ...
##  $ gender: Factor w/ 2 levels "male","female": 2 1 1 2 2 1 2 1 1 1 ...
##  $ edu   : Factor w/ 3 levels "hs or lower",..: 2 1 1 1 1 1 1 2 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int  3 7 8 9 11 15 19 40 52 55 ...
##   ..- attr(*, "names")= chr  "3" "7" "8" "9" ...
```

```
#If they were numerics and we wanted to treat them as factors, we would have to use as.factor()
# acs12$gender = as.factor(acs12$gender)

#Look at factor variable levels
levels(acs12$edu)
```

```
## [1] "hs or lower" "college"     "grad"
```

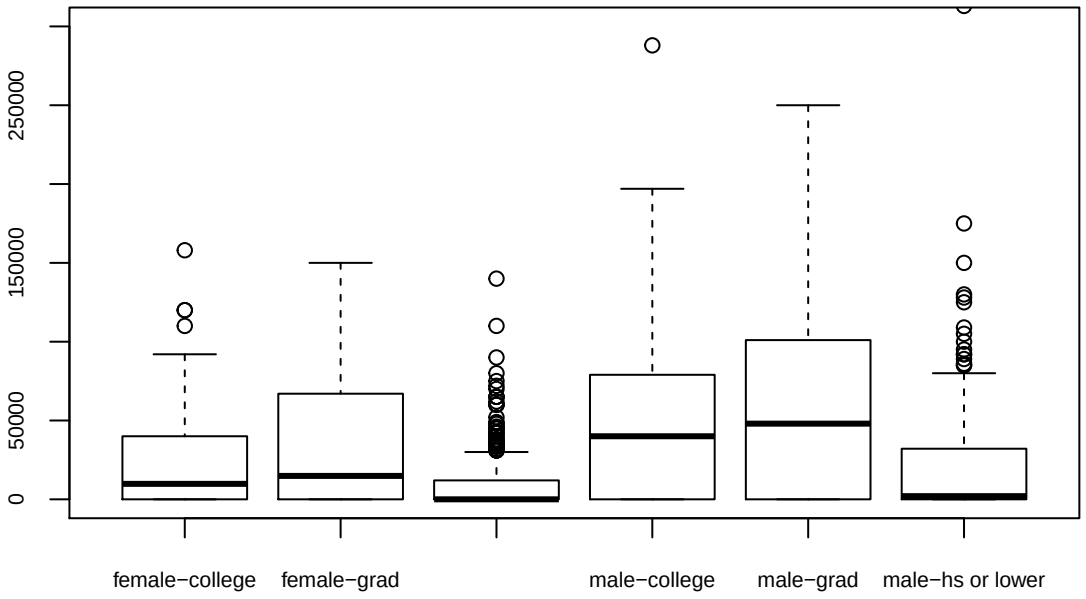
```
levels(acs12$gender)
```

```
## [1] "male"   "female"
```

```
#more dplyr!
acs12 %>% group_by(gender, edu)  %>% summarize(grp_mean=mean(income))
```

```
## # A tibble: 6 x 3
## # Groups:   gender [2]
##   gender edu         grp_mean
##   <fct>  <fct>          <dbl>
## 1 male   hs or lower   19367.
## 2 male   college       48719.
## 3 male   grad          91412.
## 4 female hs or lower    8685.
## 5 female college       22598.
## 6 female grad          39408.
```

```
#we can also look at box plots
grp = as.factor(paste0(acs12$gender,"-", acs12$edu))
boxplot(split(acs12$income, grp), names=levels(grp), cex.axis=.7, ylim=c(0,3e+05))
```



Now, lets fit an additive model, predicting income as a function of gender and ed level.

```
lm_add = lm(income ~ gender + edu, data=acs12)
summary(lm_add)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   22844.47   1650.663 13.839573 2.976116e-41
## genderfemale -17724.85   2121.157 -8.356216 1.374991e-16
## educollege    21570.96   2589.940  8.328750 1.718070e-16
## edugrad       53332.06   3781.463 14.103549 1.088970e-42
```

How do we interpret these coefficients?

```
#intercept = estimated grp mean for the reference category groups
#i.e., mean income when gender=male, edu=hs or lower
summary(lm_add)$coef[1]
```

```
## [1] 22844.47
```

```
#then, the coefficient for educollege is the difference in grp means
#for the college group as compared to the reference group
#so to calculate mean income when gender=male and edu=college:
summary(lm_add)$coef[1] + summary(lm_add)$coef[3]
```

```
## [1] 44415.44
```

```
#to calculate mean income when gender=male and edu=grad:
summary(lm_add)$coef[1] + summary(lm_add)$coef[4]
```

```
## [1] 76176.53
```

```
#The coefficient on genderfemale is the mean difference when gender =female compared to gender=male
#so to calculate grp means for gender = female and each of the college levels, we add this
#coefficient to each of the estimates above:
```

```
#i.e., mean income when gender=female, edu=hs or lower
summary(lm_add)$coef[1] + summary(lm_add)$coef[2]
```

```
## [1] 5119.627
```

```
#to calculate mean income when gender=female and edu=college:
summary(lm_add)$coef[1] + summary(lm_add)$coef[3] + summary(lm_add)$coef[2]
```

```
## [1] 26690.59
```

```
#to calculate mean income when gender=female and edu=grad:
summary(lm_add)$coef[1] + summary(lm_add)$coef[4] + summary(lm_add)$coef[2]
```

```
## [1] 58451.68
```

Compare these estimates to the true group means. Why are they not the same? What assumptions did we make in this model?

```
acs12 %>% group_by(gender, edu)  %>% summarize(grp_mean=mean(income))
```

```
## # A tibble: 6 x 3
## # Groups:   gender [2]
##   gender edu         grp_mean
##   <fct>  <fct>          <dbl>
## 1 male   hs or lower   19367.
## 2 male   college       48719.
## 3 male   grad          91412.
## 4 female hs or lower    8685.
## 5 female college       22598.
## 6 female grad          39408.
```

Now, let's try the same thing, but interacting gender and edu:

```
lm_interact = lm(income ~ gender * edu, data=acs12)
summary(lm_interact)$coef
```

```
##                            Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 19366.68   1775.326 10.908803 8.742645e-27
## genderfemale               -10681.22   2526.531 -4.227624 2.493795e-05
## educollege                  29352.17   3655.620  8.029328 1.868837e-15
## edugrad                     72044.94   5048.767 14.269810 1.331692e-43
## genderfemale:educollege    -15439.97   5129.076 -3.010283 2.650454e-03
## genderfemale:edugrad       -41322.59   7526.251 -5.490462 4.647790e-08
```

Now, let's estimate the same group means:

```
#to calculate mean income when gender=male and edu=hs or lower:
summary(lm_interact)$coef[1]
```

```
## [1] 19366.68
```

```
#to calculate mean income when gender=male and edu=college:
summary(lm_interact)$coef[1] + summary(lm_interact)$coef[3]
```

```
## [1] 48718.86
```

```
#to calculate mean income when gender=male and edu=grad:
summary(lm_interact)$coef[1] + summary(lm_interact)$coef[4]
```

```
## [1] 91411.63
```

```
#to calculate mean income when gender=female, edu=hs or lower
summary(lm_interact)$coef[1] + summary(lm_interact)$coef[2]
```

```
## [1] 8685.461
```

```
#to calculate mean income when gender=female and edu=college:
summary(lm_interact)$coef[1] + summary(lm_interact)$coef[2] +
   summary(lm_interact)$coef[3] + summary(lm_interact)$coef[5]
```

```
## [1] 22597.66
```

```
#to calculate mean income when gender=female and edu=grad:
#TRY THIS ONE ON YOUR OWN!!!
```

Check this against the true group means:

```
acs12 %>% group_by(gender, edu)  %>% summarize(grp_mean=mean(income))
```

```
## # A tibble: 6 x 3
## # Groups:   gender [2]
##   gender edu         grp_mean
##   <fct>  <fct>          <dbl>
## 1 male   hs or lower   19367.
## 2 male   college       48719.
## 3 male   grad          91412.
## 4 female hs or lower    8685.
## 5 female college       22598.
## 6 female grad          39408.
```

Another way to do this:
Create a new group variable with all 6 combinations of gender/edu. Then run a regression on group as a factor variable. If you include "-1" in the equation, then you are telling R not to estimate the intercept and instead estimate group level expected means when all of the other factor levels are equal to 0. This directly estimates group means without using the interaction terms.

```
acs12$grp = paste0("-", acs12$gender, "-", acs12$edu)
lm_allgrps = lm(income~as.factor(grp)-1, data=acs12)
summary(lm_allgrps)$coef
```

```
##                                   Estimate Std. Error   t value
## as.factor(grp)-female-college     22597.663   3116.454  7.251081
## as.factor(grp)-female-grad        39407.812   5284.205  7.457661
## as.factor(grp)-female-hs or lower  8685.461   1797.658  4.831542
## as.factor(grp)-male-college       48718.857   3195.587 15.245667
## as.factor(grp)-male-grad          91411.625   4726.337 19.340903
## as.factor(grp)-male-hs or lower   19366.684   1775.326 10.908803
##                                       Pr(>|t|)
## as.factor(grp)-female-college     6.382841e-13
## as.factor(grp)-female-grad        1.429466e-13
## as.factor(grp)-female-hs or lower 1.483125e-06
## as.factor(grp)-male-college       3.871551e-49
## as.factor(grp)-male-grad          3.889862e-75
## as.factor(grp)-male-hs or lower   8.742645e-27
```

#Practice
Lastly, let's look at a new dataset. We'll return to the ncbirths dataset, which includes data for 2000 births in North Carolina. It is also located in the openintro package.

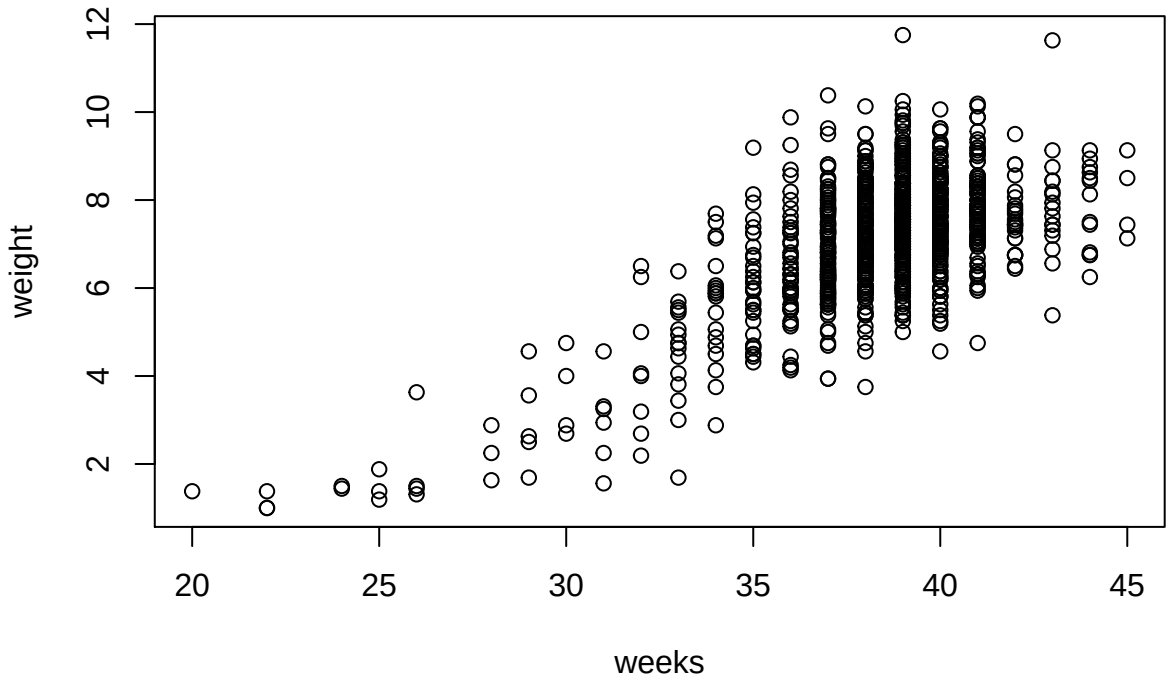
```
data("ncbirths")
```

Try these questions on your own first; then we will discuss the answers together!
1. Plot the weight variable (y-axis) against the weeks variable (x axis). What do you think is the relationship between the number of weeks of pregnancy and expected weight of babies?
2. Fit a linear model of weight ~ weeks. Then fit a quadratic model of weight ~ weeks + weeks^2 and a cubic model of weight ~ weeks + weeks^2 + weeks^3. Superimpose the regression lines from each model onto the plot from question 1. Which looks like a better fit?
3. Test whether the cubic, quadratic, or linear model fits the data better using an F test.
4. Does it make sense to model log(weight) as a function of weeks? Why or why not? Try it and plot the regression line on top of the data. Comment on possible trajectories for lines plotted with this model.

Answers

1. Plot the weight variable (y-axis) against the weeks variable (x axis). What do you think is the relationship between the number of weeks of pregnancy and expected weight of babies?

```
plot(ncbirths$weeks, ncbirths$weight, xlab="weeks", ylab="weight")
```



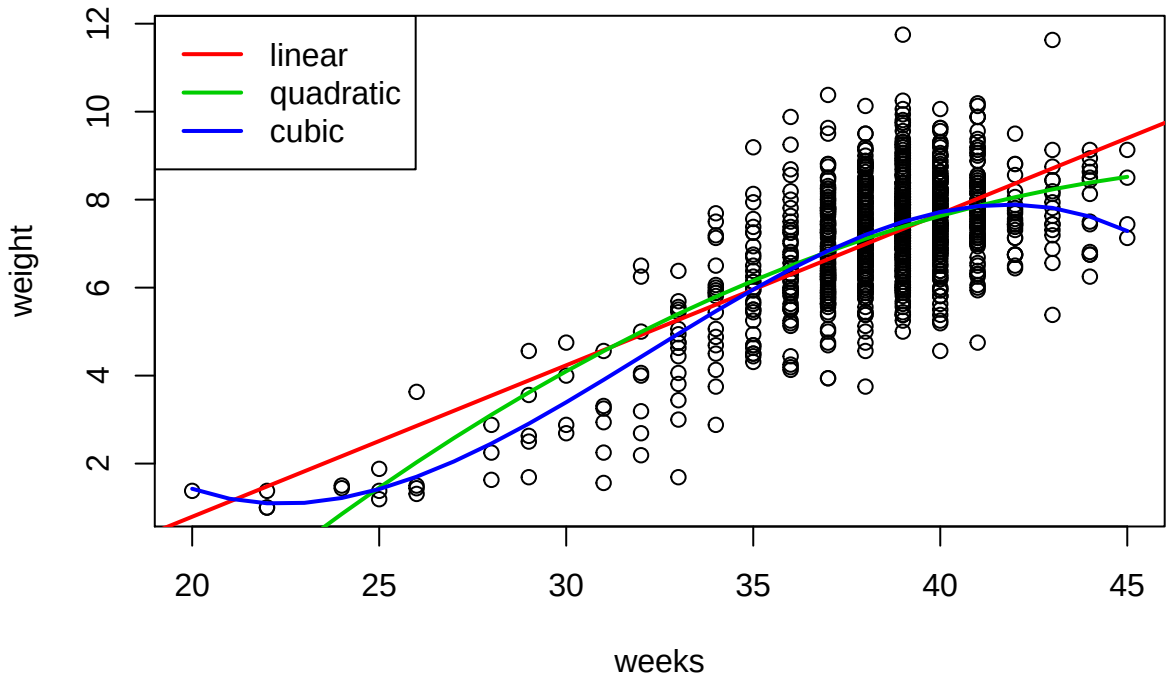
It looks like expected score increases with motheriq, but then either decreases or levels off for large motheriq values.

2. Fit a linear model of weight ~ weeks. Then fit a quadratic model of weight ~ weeks + weeks^2 and a cubic model of weight ~ weeks + weeks^2 + weeks^3. Superimpose the regression lines from each model onto the plot from question 1. Which looks like a better fit?

```
linmod = lm(weight ~ weeks, data=ncbirths)
quadmod = lm(weight ~ weeks + I(weeks^2), data=ncbirths)
cubemod = lm(weight ~ weeks + I(weeks^2) + + I(weeks^3), data=ncbirths)

weeksindata = seq(min(ncbirths$weeks, na.rm=T), max(ncbirths$weeks, na.rm=T), 1)
quadscores = predict(quadmod, newdata = data.frame(weeks = weeksindata))
cubescores = predict(cubemod, newdata = data.frame(weeks = weeksindata))

plot(ncbirths$weeks, ncbirths$weight, xlab="weeks", ylab="weight")
abline(linmod, col=2, lwd=2)
lines(weeksindata, quadscores, col=3, lwd=2)
lines(weeksindata, cubescores, col=4, lwd=2)
legend("topleft",c("linear", "quadratic", "cubic"), col=c(2,3,4), lty=1, lwd=2)
```



The quadratic model looks like it fits the model slightly better, and the cubic model looks even better.

3. Test whether the cubic, quadratic, or linear model fits the data better using an F test.

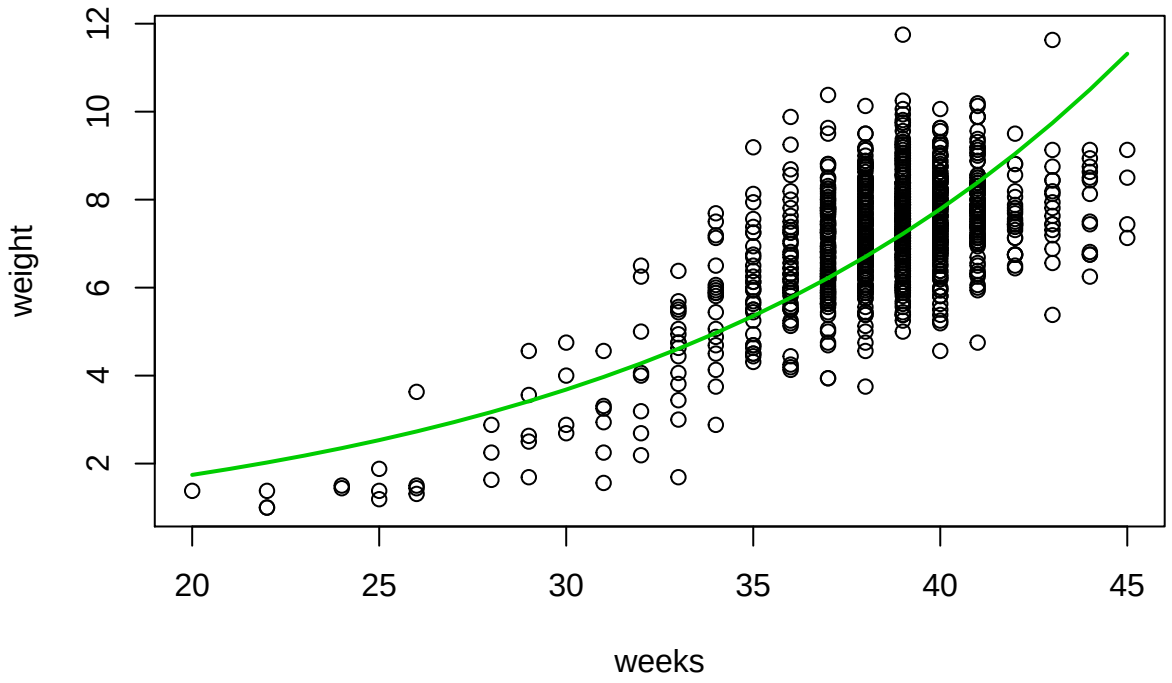
```
anova(linmod, quadmod, cubemod) #cubic is best
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ weeks
## Model 2: weight ~ weeks + I(weeks^2)
## Model 3: weight ~ weeks + I(weeks^2) + +I(weeks^3)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    996 1246.5
## 2    995 1189.8  1    56.673 49.821 3.158e-12 ***
## 3    994 1130.7  1    59.069 51.927 1.138e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Does it make sense to model log(weight) as a function of weeks? Why or why not? Try it and plot the regression line on top of the data. Comment on possible trajectories for lines plotted with this model.

```
logmod = lm(I(log(weight)) ~ weeks, data=ncbirths)
logweights = predict(logmod, newdata = data.frame(weeks = weeksindata))
weight_preds = exp(logweights)

plot(ncbirths$weeks, ncbirths$weight, xlab="weeks", ylab="weight")
lines(weeksindata, weight_preds, col=3, lwd=2)
```



If we model log weight as a function of weeks, we won't be able to capture this curve that seems to increase and then level out. The model log(weight) ~ weeks, assumes that that weight=exp(b0) * exp(b1)^weeks. Thus, for each additional week, baby weights are expected to be exp(b1) times higher. If exp(b1) is less than 1, then we're estimating that baby weights always decrease as the number of weeks increase, but by smaller and smaller amounts for each additional week. Conversely, if exp(b1) is greater than 1, we're estimating that baby weights are expected to increase by larger and larger amounts for every additional week. We can never model a relationship where the direction of the difference in expected baby weights for each additional week changes from positive to negative (or negative to positive); nor can we estimate a curve that increases and then levels out.