The child.iq folder contains a subset of the children and mother data discussed earlier in the chapter. You have access to children’s test scores at age 3, mother’s education, and the mother’s age at the time she gave birth for a sample of 400 children. The data are a Stata file which you can read into R by saving in your working directory and then typing the following:

library ("foreign")

iq.data <- read.dta ("child.iq.dta")

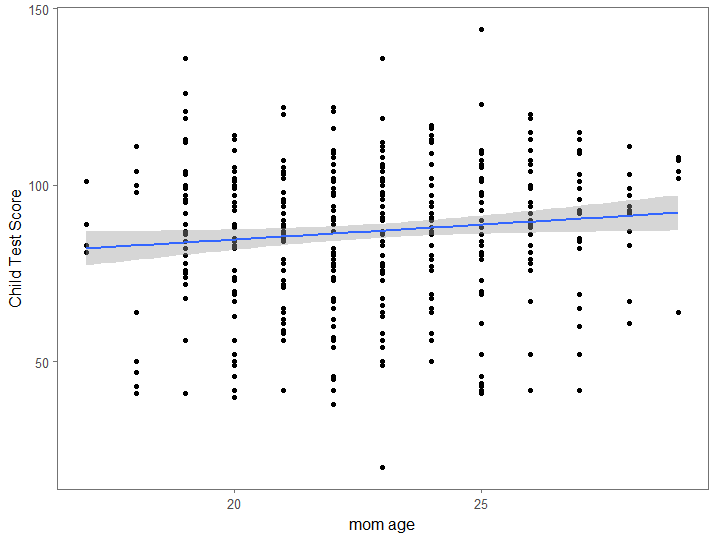
1. **Fit a regression of child test scores on mother’s age, display the data and fitted model, check assumptions, and interpret the slope coefficient. When do you recommend mothers should give birth? What are you assuming in making these recommendations?**

attach(iq.data)

ggplot(data=iq.data, aes(x=momage, y=ppvt))+

geom\_point()+

stat\_smooth(method=lm)



lm1 <- lm(ppvt~momage)

summary(lm1)

Residuals:

Min 1Q Median 3Q Max

-67.109 -11.798 2.971 14.860 55.210

Coefficients:

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

(Intercept) 67.7827 8.6880 7.802 5.42e-14 \*\*\*

momage 0.8403 0.3786 2.219 0.027 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.34 on 398 degrees of freedom

Multiple R-squared: 0.01223, Adjusted R-squared: 0.009743

F-statistic: 4.926 on 1 and 398 DF, p-value: 0.02702

Here we assume that the relationship is actually linear, and we also bin the mothers age into discrete categories that each include a whole year. As a result, it is possible that two mothers in different age groups are closer in age to each other than they are to some mothers in their own age group.

The slope coefficient, .8403, suggests that for each additional year in the mothers age when they give birth, the child will score almost a point higher on this test.

This would suggest that mothers should give birth as late as possible (within the age range of this study, <30). However making this suggestion relies heavily upon the assumption that there is a linear relationship. It is also possible that children’s performance begins to drop off at some point, but this is not reflected in our model. This suggestion also assumes that there are no other detrimental effects from having children at a later age and that the correlation is causal.

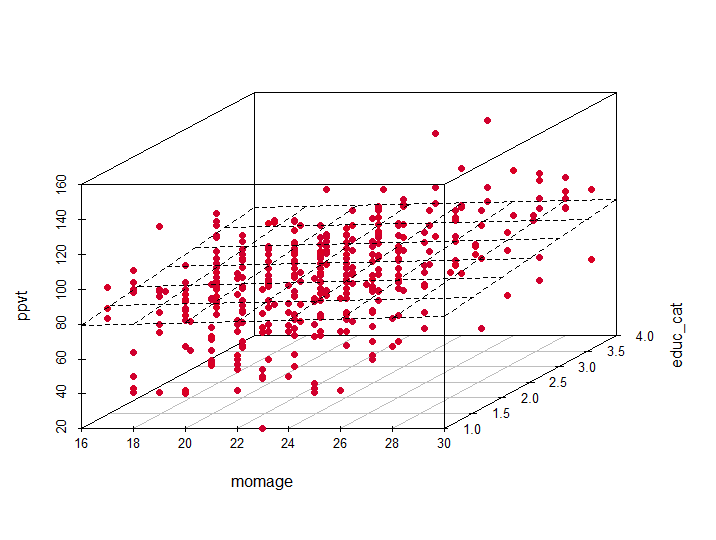
1. **Repeat this for a regression that further includes mother’s education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?**
2. lm(formula = ppvt ~ momage + educ\_cat)
3. Residuals:
4. Min 1Q Median 3Q Max
5. -61.763 -13.130 2.495 14.620 55.610
6. Coefficients:
7. Estimate Std. Error t value Pr(>|t|)
8. (Intercept) 69.1554 8.5706 8.069 8.51e-15 \*\*\*
9. momage 0.3433 0.3981 0.862 0.389003
10. educ\_cat 4.7114 1.3165 3.579 0.000388 \*\*\*
11. ---
12. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
13. Residual standard error: 20.05 on 397 degrees of freedom
14. Multiple R-squared: 0.04309, Adjusted R-squared: 0.03827
15. F-statistic: 8.939 on 2 and 397 DF, p-value: 0.0001594

After including the mothers educational level in the model we get very different results. In this model, the mothers age has no significant effect on the child’s test score since the coefficient for age is very small and the p value is large. However, the mothers educational category appears to have a large effect on the child’s score since the effect size is much larger and the p value is very small. Despite this, the R2 value remains very small. As a result, my conclusion changes in that I would suggest that mothers not worry about their age when considering their child’s intelligence. I maintain my assertion that none of these models are very informative and that there are other variables that will likely have a much larger effect on the child (such as early reading exposure).

colFun <- colorRampPalette(c("red","blue"))

plot3d <- scatterplot3d(ppvt~momage+educ\_cat, pch=16,color=colFun(length(ppvt)))

plot3d$plane3d(lm2)



This graph simply allows us to separately see the relationship between the mom’s age and test score as well as the relationship between the mom’s educational category and the test score (as 2 dimensional cross sections of the 3d fit).

**(c) Now create an indicator variable reflecting whether the mother has completed high school or not. Consider interactions between the high school completion and mother’s age in family. Also, create a plot that shows the separate regression lines for each high school completion status group.**

factorVect <- as.factor(educ\_cat)

tempDF <- model.matrix(~factorVect-1)

isHS <- tempDF[,2]+tempDF[,3]+tempDF[,4] #This is my indicator variable

lm3 <- lm(momage~isHS)

summary(lm3)

lm4 <- lm(ppvt~isHS\*momage)

summary(lm4)

lm(formula = momage ~ isHS)

Residuals:

Min 1Q Median 3Q Max

-6.1143 -2.1143 -0.1143 1.8857 6.4118

Coefficients:

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

(Intercept) 21.5882 0.2841 75.989 < 2e-16 \*\*\*

isHS 1.5261 0.3201 4.767 2.63e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.619 on 398 degrees of freedom

Multiple R-squared: 0.05401, Adjusted R-squared: 0.05163

F-statistic: 22.72 on 1 and 398 DF, p-value: 2.629e-06

Call:

lm(formula = ppvt ~ isHS \* momage)

Residuals:

Min 1Q Median 3Q Max

-56.696 -12.407 2.022 14.804 54.343

Coefficients:

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

(Intercept) 105.2202 17.6454 5.963 5.49e-09 \*\*\*

isHS -38.4088 20.2815 -1.894 0.0590 .

momage -1.2402 0.8113 -1.529 0.1271

isHS:momage 2.2097 0.9181 2.407 0.0165 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.85 on 396 degrees of freedom

Multiple R-squared: 0.06417, Adjusted R-squared: 0.05708

F-statistic: 9.051 on 3 and 396 DF, p-value: 8.276e-06

iq.data <- cbind(iq.data,isHS)

dataHS <- iq.data[iq.data$isHS==1,]

dataNoHS <- iq.data[iq.data$isHS==0,]

ggplot(data=iq.data, aes(x=momage, y=ppvt))+

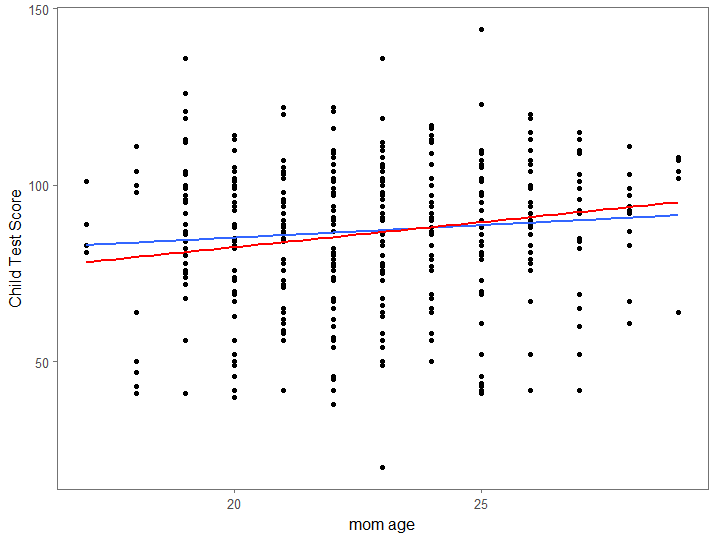
geom\_point()+

stat\_smooth(data=dataHS, method=lm)+

stat\_smooth(data=dataNoHS, method=lm, col="red")+

labs(x="mom age", y="Child Test Score")+

theme\_few()



(d) Finally, fit a regression of child test scores on mother’s age and education level for the first 200 children and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.

#The real data

iq.data.short <- iq.data[1:200,]

attach(iq.data.short)

lm5 <- lm(ppvt~momage, data=iq.data.short)

lm5Coefs <- coef(lm5)

summary(lm5)

#Find stdev of the data from the model

stdevlm5 <- sigma(lm5)

#Use real mom age values for the fake data

xFake <- iq.data$momage[201:400]

#Create fake data based on linear model and its error

y <- lm5Coefs[1]+lm5Coefs[2]\*xFake+rnorm(length(xFake),0,stdevlm5)

#Make data frame with fake data

fake <- rep(1,200)

notFake <- rep(0,200)

isFake <- as.factor(c(notFake,fake))

allX <- c(iq.data.short$momage,xFake)

allY <- c(iq.data.short$ppvt,y)

allDataIQ <- data.frame(allX,allY,isFake)

ggplot(data=allDataIQ, aes(x=allX, y=allY))+

geom\_point(pch=16,size=3,aes(color=isFake))+

labs(x="mom age", y="Child Test Score", color="data")+

scale\_color\_manual(values=c("red","blue"), labels=c("Real Data","Fake Data"))+

stat\_smooth(data=allDataIQ[allDataIQ$isFake==1,], method=lm, se=F)+

stat\_smooth(data=allDataIQ[allDataIQ$isFake==0,], method=lm, col="red",se=F)+

theme\_few()

