Workshop 9 Solutions

Veerasak

null

We’ll begin by loading some packages.

library(MASS)  
library(plyr)

### Interaction terms in regression

# Building up the familiar birthwt data...  
  
# Rename the columns to have more descriptive names  
colnames(birthwt) <- c("birthwt.below.2500", "mother.age", "mother.weight",   
 "race", "mother.smokes", "previous.prem.labor", "hypertension", "uterine.irr",   
 "physician.visits", "birthwt.grams")  
  
# Transform variables to factors with descriptive levels  
birthwt <- transform(birthwt,   
 race = as.factor(mapvalues(race, c(1, 2, 3),   
 c("white","black", "other"))),  
 mother.smokes = as.factor(mapvalues(mother.smokes,   
 c(0,1), c("no", "yes"))),  
 hypertension = as.factor(mapvalues(hypertension,   
 c(0,1), c("no", "yes"))),  
 uterine.irr = as.factor(mapvalues(uterine.irr,   
 c(0,1), c("no", "yes")))  
 )

**(a)** Run a linear regression to better understand how birthweight varies with the mother’s age and smoking status (do not include interaction terms).

# Run regression model  
birthwt.lm <- lm(birthwt.grams ~ mother.age + mother.smokes, data = birthwt)  
  
# Output coefficients table  
summary(birthwt.lm)

##   
## Call:  
## lm(formula = birthwt.grams ~ mother.age + mother.smokes, data = birthwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2119.98 -442.66 52.92 532.38 1690.74   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2791.224 240.950 11.584 <2e-16 \*\*\*  
## mother.age 11.290 9.881 1.143 0.255   
## mother.smokesyes -278.356 106.987 -2.602 0.010 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 717.2 on 186 degrees of freedom  
## Multiple R-squared: 0.04299, Adjusted R-squared: 0.0327   
## F-statistic: 4.177 on 2 and 186 DF, p-value: 0.0168

**(b)** What is the coefficient of mother.age in your regression? How do you interpret this coefficient?

coef(birthwt.lm)["mother.age"]

## mother.age   
## 11.28961

age.coef <- round(coef(birthwt.lm)["mother.age"], 1)

**Note: This solution uses inline code chunks.** The coefficient is 11.3. This means that among mothers with the same smoking status, each additional year of age is on average associated with a 11.3g increase in birthweight.

**(c)** How many coefficients are estimated for the mother’s smoking status variable? How do you interpret these coefficients?

coef(birthwt.lm)["mother.smokesyes"]

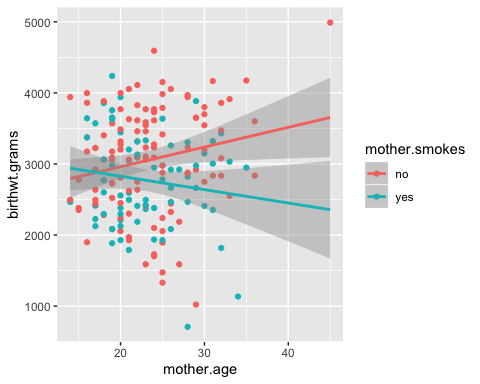
## mother.smokesyes   
## -278.3561

smoke.coef <- abs(round(coef(birthwt.lm)["mother.smokesyes"], 1))

**Note: This solution uses inline code chunks.** There is just one coefficient estimated. This coefficient gives us the average difference in birthweight between mothers that smoke and mother’s that don’t, in a model that adjusts for the effect of mother’s age. That is, after we adjust for the effect of age, smoking leads to an average 278.4 decrease in birthweight.

**(d)** Using ggplot, construct a scatterplot with birthweight on the y-axis and mother’s age on the x-axis. Color the points by mother’s smoking status, and add smoking status-specific linear regression lines using the stat\_smooth layer.

library(ggplot2)  
  
# Note fullrange = TRUE is used here to extend the 'mother.smokes = yes' line beyond the maximum age (35) in this group  
qplot(data = birthwt, x = mother.age, y = birthwt.grams, colour = mother.smokes) + stat\_smooth(method = "lm", fullrange = TRUE)



## Prediction

predict(birthwt.lm, newdata = birthwt[20,])

## 105   
## 2828.977

birthwt[20,"birthwt.grams"]

## [1] 2821

## Confident Interval

predict(birthwt.lm, newdata = birthwt[20,], interval = "confidence")

## fit lwr upr  
## 105 2828.977 2637.25 3020.705

## Prediction Interval

predict(birthwt.lm, newdata = birthwt[20,], interval = "prediction")

## fit lwr upr  
## 105 2828.977 1401.166 4256.789