MultipredictorRegMod.Rmd

F. A. BarriosInstituto de Neurobiología, UNAM

2020-11-04

Outline

1

23

```
# Multilevel categorical predictors Chap 4 Vittinghoff
# libraries
#
library(tidyverse)
library(emmeans)
library(multcomp)
library(MASS)
library(car)
library(HSAUR2)
```

```
#
```

Exercise and Glucose (4.1 Vittinghoff)

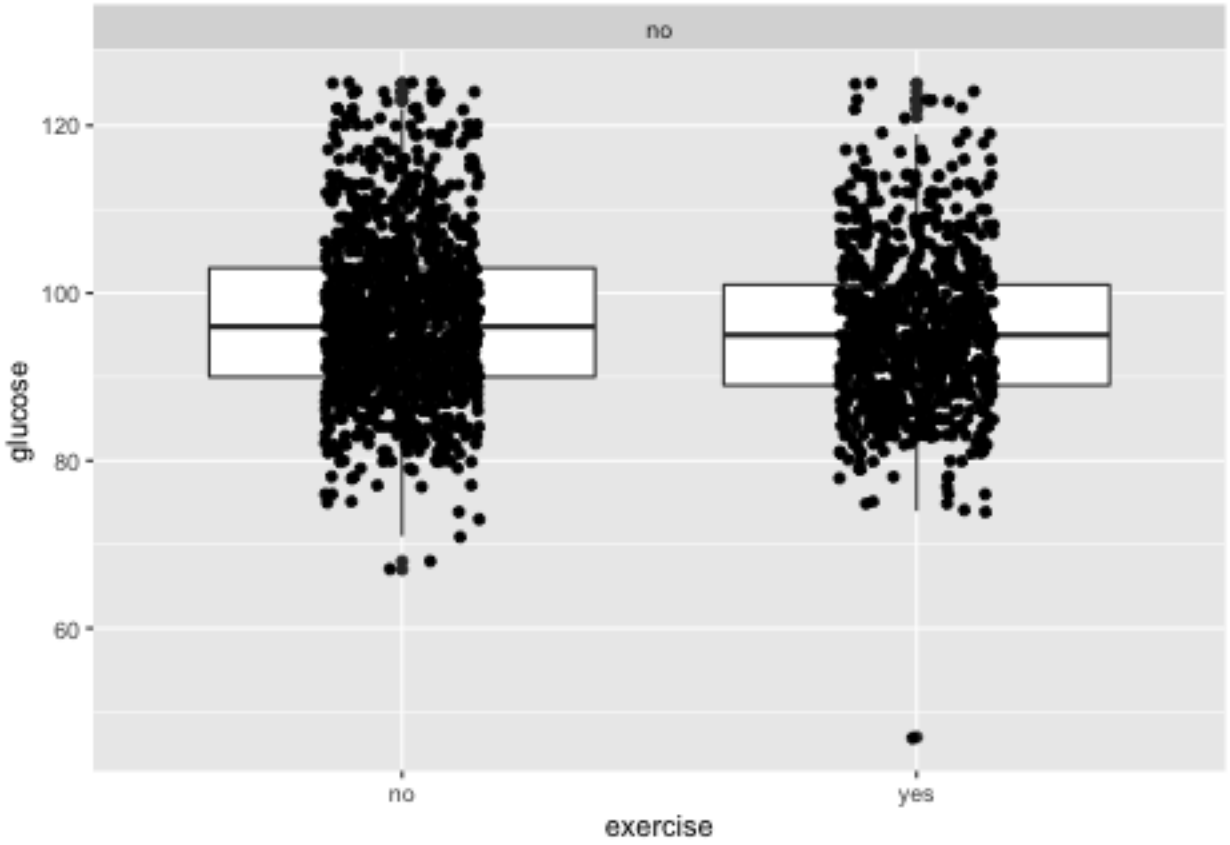
Glucose levels above 125 mg/dL are diagnostic of diabetes, while levels in the range from 100 to 125 mg/dL signal increased risk of progressing to this serious and increasingly widespread condition. So it is of interest to determine whether exercise, a modifiable lifestyle factor, would help people reduce their glucose levels and thus avoid diabetes. The R code shows a simple linear model using a measure of exercise to predict baseline glucose levels among 2,032 participants without diabetes, and boxplot of the data, in the HERS. Women with diabetes are excluded because the research question is whether exercise might help to prevent progression to diabetes among women at risk, and because the causal determinants of glucose may be different in that group.

```
setwd("~/Dropbox/GitHub/Class2020")
# Chapter 4 examples. 4.1
hers <- read_csv("DataRegressBook/Chap3/hersdata.csv")
```

```
hers_nodi <- filter(hers, diabetes == "no")
# Simple linear model with HERS data for women without diabetes
ggplot(data = hers_nodi, mapping = aes(x = exercise, y = glucose)) +
  geom_boxplot(na.rm = TRUE) + facet_grid( . ~ diabetes) + geom_jitter(height = 0.15, width = 0.15)
# The simple linear model adjust the exercise like in table 4.1
hers_nodi_Fit <- lm(glucose ~ exercise, data = hers_nodi)
summary(hers_nodi_Fit)
```

```
Call:
lm(formula = glucose ~ exercise, data = hers_nodi)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-48.668  -6.668  -0.668   5.639  29.332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  97.3610     0.2815 345.848  < 2e-16 ***
exerciseyes  -1.6928     0.4376  -3.868 0.000113 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.715 on 2030 degrees of freedom
Multiple R-squared:  0.007318,  Adjusted R-squared:  0.006829
F-statistic: 14.97 on 1 and 2030 DF,  p-value: 0.000113
```



The "coefficients" for the variable "exerciseyes" shows that average baseline glucose levels were about 1.7mg/dL lower among women who exercised at least three times a week than among women who exercised less. This difference is statistically significant ($t = -3.87, P < 0.000113$). However, women who exercise are slightly younger, a little more likely to use alcohol, and in particular have lower average BMI, all factors associated with glucose levels. This implies that the lower average glucose we observe among women who exercise could be due at least in part to differences in these other predictors (independent variables). Under these conditions, it is important that our estimate of the difference in average glucose levels associated with exercise be "adjusted" for the effects of these potential confounders of the unadjusted association. Ideally, adjustment using a multipredictor regression model provides an estimate of the causal effect of exercise on average glucose levels, by holding the other variables constant.

The multiple regression model (4.2)

In the multiple regression model the expected value $E[y \mid x]$ (expected value of the response function y given the vector x) is

$$E[y \mid x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p$$

where x represents the collection of p predictors $x_1, x_2, \ldots x_p$ in the model, and the β are the corresponding regression coefficients.

```
# Chap 4 4.2 Multiple linear regresor model
# and to obtain the table 4.2 with multiple linear model

hers_nodi_Fit2 <- lm(glucose ~ exercise + age + drinkany + BMI, data = hers_nodi)
S(hers_nodi_Fit2)
```

```
Call: lm(formula = glucose ~ exercise + age + drinkany + BMI, data = hers_nodi)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.96239    2.59284  30.454   <2e-16 ***
exerciseyes -0.95044    0.42873  -2.217   0.0267 *
age          0.06355    0.03139   2.024   0.0431 *
drinkanyyes  0.68026    0.42196   1.612   0.1071
BMI          0.48924    0.04155  11.774   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard deviation: 9.389 on 2023 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.07197
F-statistic: 39.22 on 4 and 2023 DF,  p-value: < 2.2e-16
     AIC      BIC
14845.62 14879.31
```

```
#
```

in a multiple regression model that also includes —that is, adjusts for—age, alcohol use (drinkany), and BMI, average glucose is estimated to be only about 1 mg/dL lower among women who exercise, holding the other three factors constant. The multipredictor model also shows that average glucose levels are about 0.7 mg/dL higher among alcohol users than among nonusers. Average levels also increase by about 0.5 mg/dL per unit increase in BMI, and by 0.06 mg/dL for each additional year of age. Each of these associations is statistically significant after adjustment for the other predictors in the model.

Interpretation of Adjusted Regression Coefficients

The coefficient $\beta_j, j = 1, \ldots, p$ gives the change in $E[y \mid x]$ for an increase of one unit in predictor x_j , holding other factors in the model constant; each of the estimates is adjusted for the effects of all the other predictors. As in the simple linear model, the intercept β_0 gives the value of $E[y \mid x]$ when all the predictors are equal to zero.

Generalization of R-squared and r

The coefficient of determination R^2 is the proportion of the total variability of the outcome that can be accounted for by the predictors. And the multiple correlatio coefficient $r = \sqrt{R^2}$ represents the correlation between the outcome y and the fitted values $\hat{y}$.

Categorical Predictors

Predictors in both simple and multiple predictor regression models can be binary, categorical, or discrete numeric, as well as continuous numeric.

Binary Predictors

Binary predictors, a group with a characteristic and other group with out the characteristic, can be coded with a dummy variable, an indicator or dummy variable that can take value "1" for the group with the chracteristic and "0" for the group without the characteristic. With this coding, the regression coefficient corresponding to this variable has a straightforward interpretation as the increase or decrease in average outcome levels in the group with the characteristic, with respect to the reference group. With this coding for binary variables 1 = yes and 0 = no β_0 is the average of the baseline variable and $\beta_0 + \beta_1$ is related to the value for the "yes" condition ("yes" + average).

Multilevel Categorical Predictors (4.3)

The 2,763 women in the HERS cohort also responded to a question about how physically active they considered themselves compared to other women their age. The five-level response variable "physact" ranged from "much less active" to "much more active," and was coded in order from 1 to 5. This is an example of an ordinal variable. Multilevel categorical variables can also be nominal, in the sense that there is no intrinsic ordering in the categories. Examples include ethnicity, marital status, occupation, and geographic region. With nominal variables, it is even clearer that the numeric codes often used to represent the variable in the database cannot be treated like the values of a numeric variable. Categorical variables are easily accommodated in multipredictor linear and other regression models, using indicator or dummy variables. As with binary variables, where two categories are represented in the model by a single indicator variable, categorical variables with $K \leq 2$ levels are represented by $K - 1$ indicators, one for each of level of the variable except a baseline or reference level.

```
# Chap 4  4.3 Categorical predictors
# we are using the same file hers <- read_csv("DataRegressBook/Chap3/hersdata.csv")
# Multilevel categorical predictors using the linear model for women without diebetes
# IMPORTANT compare with table 4.4 Regression of physical activity on glucose

hers_nodi <- mutate(hers_nodi, physact = factor(physact, levels=c("much less active","somewhat less acti
levels(hers_nodi$physact)
```

```
[1] "much less active"     "somewhat less active" "about as active"
[4] "somewhat more active" "much more active"
```

```
ggplot(data = hers_nodi, mapping = aes(x = physact, y = glucose)) + geom_boxplot(na.rm = TRUE)
glucose_fit_act <- lm(glucose ~ physact, data = hers_nodi)
#
Anova(glucose_fit_act, type="II")
```

```
Anova Table (Type II tests)

Response: glucose
          Sum Sq   Df F value   Pr(>F)
physact     1673    4   4.431 0.001441 **
Residuals 191345 2027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
S(glucose_fit_act)
```

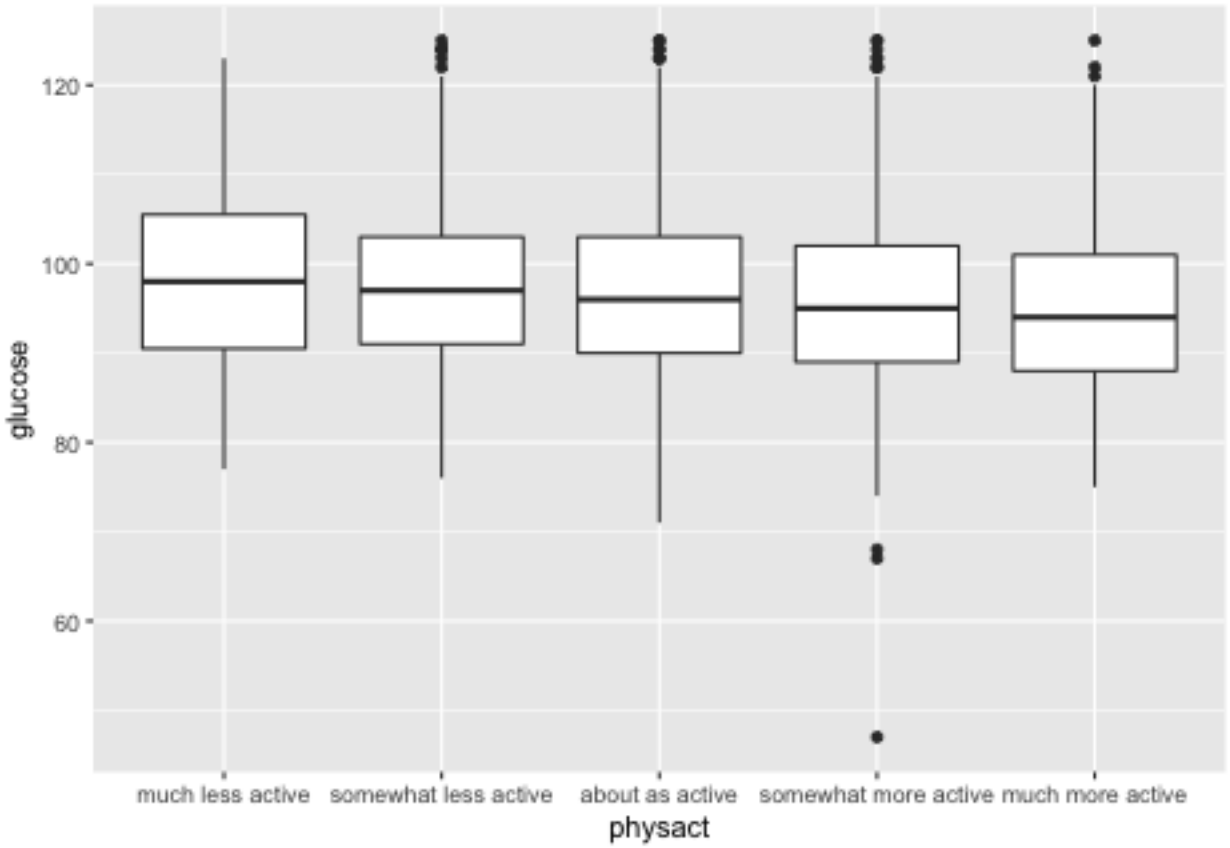
```
Call: lm(formula = glucose ~ physact, data = hers_nodi)
```

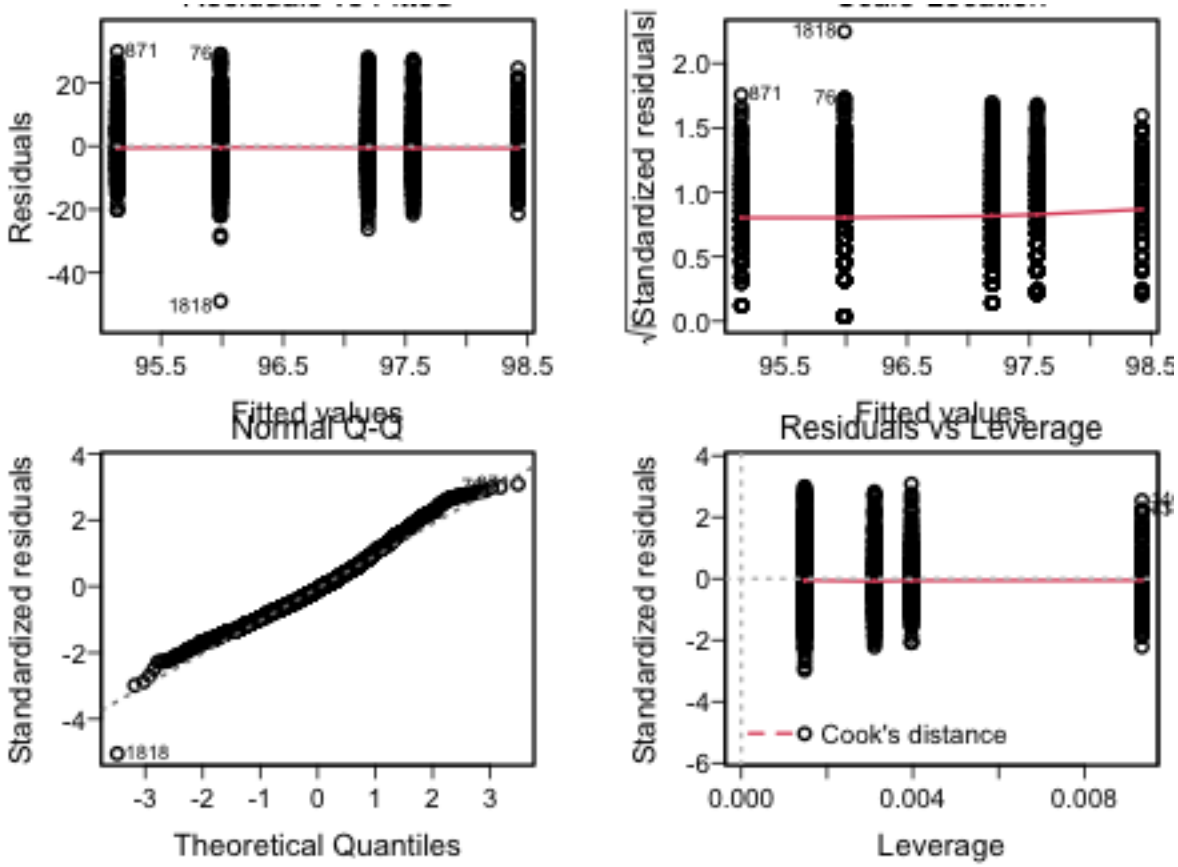
```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  98.4206     0.9393 104.784   <2e-16 ***
physactsomewhat less active  -0.8584     1.0842  -0.792   0.4286
physactabout as active       -1.2262     1.0111  -1.213   0.2254
physactsomewhat more active  -2.4339     1.0108  -2.408   0.0161 *
physactmuch more active      -3.2777     1.1211  -2.924   0.0035 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard deviation: 9.716 on 2027 degrees of freedom
Multiple R-squared: 0.008668
F-statistic: 4.431 on 4 and 2027 DF,  p-value: 0.001441
     AIC      BIC
15014.12 15047.82
```

```r
layout(matrix(1:4, nrow = 2))
plot(glucose_fit_act)
# To compute the estimates marginal means for specified factors or factor combinations in a linear mode
glucose_emmeans <- emmeans(glucose_fit_act, "physact")
summary(glucose_emmeans)
```

```
 physact              emmean    SE   df lower.CL upper.CL
 much less active       98.4 0.939 2027     96.6    100.3
 somewhat less active   97.6 0.541 2027     96.5     98.6
 about as active        97.2 0.374 2027     96.5     97.9
 somewhat more active   96.0 0.373 2027     95.3     96.7
 much more active       95.1 0.612 2027     93.9     96.3

Confidence level used: 0.95
```

```r
#
```


120
100
80
60
glucose
much less active
somewhat less active
about as active
somewhat more active
much more active
physact



the corresponding β_i have a straightforward interpretation. For the moment, consider a simple regression model in which the five levels of "physact" are the only predictors:

$$E[glucose \mid x] = \beta_0 + \beta_2 SomewhatLessActive + \beta_3 AboutAsActive + \beta_4 SomewhatMoreActive + \beta_5 MuchMoreActive$$

Without other predictors, or covariates, the model is equivalent to a one-way ANOVA. The parameters of the model can be manipulated to give the estimated mean in any group, or to give the estimated differences between any two groups. Estimated differences are the contrasts.

Multiple Pairwise Comparisons Between Categories

It is frequently of interest to examine multiple pairwise differences between levels of a categorical predictor, especially when the overall F -test is statistically significant, and in some cases even when it is not. For this case, various methods are available for controlling the familywise error rate (FER) for the wider set of comparisons being made. These methods differ in the trade-off made between power and the breadth of the circumstances under which the type-I error rate is protected. One of the most straightforward is Fisher's least significant difference (LSD) procedure, in which the pairwise comparisons are carried out using t -tests at the nominal type-I error rate, but only if the overall F-test is statistically significant. More conservative procedures that protect the FER (FWE) under partial null hypotheses include setting the level of the pairwise tests required to declare statistical significance equal to α/k (Bonferroni) or$1-(1-\alpha)^{1/k}$ (Sidak), where α is the desired FER and k is the number of preplanned comparisons to be made. The Sidak correction is slightly more liberal for small values of k, but otherwise equivalent.

```
# Contrasts
# contrasts using the adjusted parameters for a categorical variable with several categories
# or multiple ordinals
# R call categorical variables factors and their categories levels so we have factors with different
# levels in these cases we can estimate contrasts of the adjusted parameters.
```

```
Contrast_Table4.6 <- list(gluc_b0 = c(1, 0, 0, 0, 0),
                          gluc_b2 = c(-1, 1, 0, 0, 0),
                          gluc_b3 = c(-1, 0, 1, 0, 0),
                          gluc_b4 = c(-1, 0, 0, 1, 0),
                          gluc_b5 = c(-1, 0, 0, 0, 1))
# For the Bonferroni correction adjust "bonferroni" this is talbe 4.6
contrast(glucose_emmeans, Contrast_Table4.6, adjust="sidak")
```

```
 contrast estimate    SE   df t.ratio p.value
 gluc_b0    98.421 0.939 2027 104.784 <.0001
 gluc_b2    -0.858 1.084 2027  -0.792 0.9391
 gluc_b3    -1.226 1.011 2027  -1.213 0.7211
 gluc_b4    -2.434 1.011 2027  -2.408 0.0781
 gluc_b5    -3.278 1.121 2027  -2.924 0.0174

P value adjustment: sidak method for 5 tests
```

```
contrast(glucose_emmeans, Contrast_Table4.6, adjust="bonferroni")
```

```
 contrast estimate    SE   df t.ratio p.value
 gluc_b0    98.421 0.939 2027 104.784 <.0001
 gluc_b2    -0.858 1.084 2027  -0.792 1.0000
 gluc_b3    -1.226 1.011 2027  -1.213 1.0000
 gluc_b4    -2.434 1.011 2027  -2.408 0.0807
 gluc_b5    -3.278 1.121 2027  -2.924 0.0175

P value adjustment: bonferroni method for 5 tests
```

```
# Same contrasts with multicomp library
```

Testing for Trend Across Categories

The coefficient estimates for the categories of physact from last section, decrease in order, suggesting that mean glucose levels are characterized by a linear trend across the levels of physact. Tests for linear trend are best performed using a contrast in the coefficients corresponding to the various levels of the categorical predictor. Definition: A contrast is a weighted sum of the regression coefficients of the form $a_1\beta_1 + a_2\beta_2 + \cdots + a_p\beta_p$ in which the weights, or contrast coefficients, sum to zero: that is, $a_1 + a_2 + \cdots + a_p = 0$

```
#
Contrasts_glu <- list(MAvsLA          = c(-1, -1, 0,  1,  1),
                      MAvsLAforMuch   = c(-1,  0, 0,  0,  1),
                      MAvsLAforSome   = c( 0, -1, 0,  1,  0),
                      MLAvsC          = c(-1,  0, 1,  0,  0),
                      SLAvsC          = c( 0, -1, 1,  0,  0),
                      SMAvsC          = c( 0,  0,-1,  1,  0),
                      MMAvsC          = c( 0,  0,-1,  0,  1),
                      LinTrend_phys   = c(-2, -1, 0,  1,  2))

# compare the results with emmeans adjusted with Sidak, FWE.
contrast(glucose_emmeans, Contrasts_glu, adjust="bonferroni")
```

```
 contrast      estimate    SE   df t.ratio p.value
 MAvsLA          -4.853 1.300 2027 -3.734  0.0016
 MAvsLAforMuch   -3.278 1.121 2027 -2.924  0.0280
 MAvsLAforSome   -1.575 0.658 2027 -2.395  0.1336
```

```
 MLAvsC          -1.226 1.011 2027 -1.213  1.0000
 SLAvsC          -0.368 0.658 2027 -0.559  1.0000
 SMAvsC          -1.208 0.529 2027 -2.284  0.1796
 MMAvsC          -2.052 0.717 2027 -2.860  0.0343
 LinTrend_phys   -8.131 2.337 2027 -3.480  0.0041

P value adjustment: bonferroni method for 8 tests
```

```
contrast(glucose_emmeans, Contrasts_glu, adjust="sidak")
```

```
 contrast      estimate    SE   df t.ratio p.value
 MAvsLA          -4.853 1.300 2027 -3.734  0.0015
 MAvsLAforMuch   -3.278 1.121 2027 -2.924  0.0276
 MAvsLAforSome   -1.575 0.658 2027 -2.395  0.1260
 MLAvsC          -1.226 1.011 2027 -1.213  0.8703
 SLAvsC          -0.368 0.658 2027 -0.559  0.9990
 SMAvsC          -1.208 0.529 2027 -2.284  0.1661
 MMAvsC          -2.052 0.717 2027 -2.860  0.0338
 LinTrend_phys   -8.131 2.337 2027 -3.480  0.0041

P value adjustment: sidak method for 8 tests
```

```
# With adjust="none", results will be the same as the aov method.

# for other more general examples
# Using the multcomp library

ContrastGlucExa1 <- ("
Contrast.Name     MLA SLA AAA SMA MMA
 MAvsLA           -1  -1   0   1   1
 MAvsLAforMuch    -1   0   0   0   1
 MAvsLAforSome     0  -1   0   1   0
 MLAvsC           -1   0   1   0   0
 SLAvsC            0  -1   1   0   0
 SMAvsC            0   0  -1   1   0
 MMAvsC            0   0  -1   0   1
 LinearTrending   -2  -1   0   1   2
")

ContrastGlucExa1_Matriz = as.matrix(read.table(textConnection(ContrastGlucExa1), header=TRUE, row.names=
ContrastGlucExa1_Matriz
```

```
               MLA SLA AAA SMA MMA
MAvsLA          -1  -1   0   1   1
MAvsLAforMuch   -1   0   0   0   1
MAvsLAforSome    0  -1   0   1   0
MLAvsC          -1   0   1   0   0
SLAvsC           0  -1   1   0   0
SMAvsC           0   0  -1   1   0
MMAvsC           0   0  -1   0   1
LinearTrending  -2  -1   0   1   2
```

```
#
# glucose_fit_act came from lm(glucose ~ physact, data = hers_nodi)
Gluc_GenLinHypoth = glht(glucose_fit_act, linfct = mcp(physact = ContrastGlucExa1_Matriz))
#
```

```
Gluc_GenLinHypoth$linfct
```

```
               (Intercept) physactsomewhat less active physactabout as active
MAvsLA                   0                          -1                       0
MAvsLAforMuch            0                           0                       0
MAvsLAforSome            0                          -1                       0
MLAvsC                   0                           0                       1
SLAvsC                   0                          -1                       1
SMAvsC                   0                           0                      -1
MMAvsC                   0                           0                      -1
LinearTrending           0                          -1                       0
               physactsomewhat more active physactmuch more active
MAvsLA                                   1                       1
MAvsLAforMuch                            0                       1
MAvsLAforSome                            1                       0
MLAvsC                                   0                       0
SLAvsC                                   0                       0
SMAvsC                                   1                       0
MMAvsC                                   0                       1
LinearTrending                           1                       2
attr(,"type")
[1] "User-defined"
```

```
summary(Gluc_GenLinHypoth, test=adjusted("single-step"))
```

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: User-defined Contrasts


Fit: lm(formula = glucose ~ physact, data = hers_nodi)

Linear Hypotheses:
                     Estimate Std. Error t value Pr(>|t|)
MAvsLA == 0           -4.8531     1.2998  -3.734  0.00137 **
MAvsLAforMuch == 0    -3.2777     1.1211  -2.924  0.02042 *
MAvsLAforSome == 0    -1.5754     0.6577  -2.395  0.08767 .
MLAvsC == 0           -1.2262     1.0111  -1.213  0.68330
SLAvsC == 0           -0.3677     0.6582  -0.559  0.97287
SMAvsC == 0           -1.2077     0.5287  -2.284  0.11481
MMAvsC == 0           -2.0515     0.7174  -2.860  0.02469 *
LinearTrending == 0   -8.1308     2.3366  -3.480  0.00324 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Confounding

The model $\texttt{lm}(\texttt{glucose} \sim \texttt{exercise}, \texttt{data} = \texttt{hersnodi})$, the unadjusted coefficient for exercise estimates the difference in mean glucose levels between two subgroups of the population of women with heart disease. But this comparison ignores other ways in which those subgroups may differ. In other words, the analysis does not take account of confounding of the association we see. Although the unadjusted coefficient may be useful for describing differences between subgroups, it would be risky to infer any causal connection

between exercise and glucose on this basis. In contrast, the adjusted coefficient for exercise in the model $\texttt{lm}(\texttt{glucose} \sim \texttt{exercise} + \texttt{age} + \texttt{drinkany} + \texttt{BMI}, \texttt{data} = \texttt{hersnodi})$, takes account of the fact that women who exercise also have lower BMI and are slightly younger and more likely to report alcohol use, all factors which are associated with differences in glucose levels. While this adjusted model is clearly rudimentary, the underlying premise of multipredictor regression analysis of observational data is that with a sufficiently refined model, we can estimate causal effects, free or almost free of confounding. Confounding is when one predictor x_i (independent variable) can depend in some of the same variables that the response function Y.

Adjusted vs. unadjusted

Uncontrolled confounding induces bias in unadjusted estimates of the causal effects that are commonly the focus of our attention. This suggests that unadjusted parameter estimates are always biased and adjusted estimates less so.

Example: BMI and LDL

We turn to a relatively simple example, again using data from the HERS cohort. BMI and LDL cholesterol are both established heart disease risk factors. It is reasonable to hypothesize that higher BMI leads to higher LDL in some causal sense. An unadjusted model for BMI and LDL is obtainde with $\texttt{lm}(\texttt{LDL} \sim \texttt{BMI}, \texttt{data} = \texttt{hers})$. The unadjusted estimate shows that average LDL increases .42mg/dL per unit increase in BMI.

```
# when one predictor (independent variable) can depend in some of the
# same variables that the response function
# Example 4
# Example of BMI (Body mass index) and LDL (cholesterol) using the HERS cohort data Table 4.12

LDL_fit_bmi <- lm(LDL ~ BMI, data = hers)
S(LDL_fit_bmi)
```

```
Call: lm(formula = LDL ~ BMI, data = hers)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 133.1913     3.7939  35.107  < 2e-16 ***
BMI           0.4151     0.1304   3.184  0.00147 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard deviation: 37.75 on 2745 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared: 0.00368
F-statistic: 10.14 on 1 and 2745 DF,  p-value: 0.001468
     AIC      BIC
27747.67 27765.43
```

However, age, ethnicity (nonwhite), smoking, and alcohol use (drinkany) may confound this unadjusted association. These covariates may either represent determinants of LDL or be proxies for such determinants, and are correlated with but almost surely not caused by BMI, and so may confound the BMI–LDL relationship.

```
# second half of Table 4.12
# example of LDL modeled with BMI and other factors
LDL_fit_all <- lm(LDL ~ BMI + age + nonwhite + smoking + drinkany, data = hers)
S(LDL_fit_all)
```

```
Call: lm(formula = LDL ~ BMI + age + nonwhite + smoking + drinkany, data =
         hers)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.3153     9.2564  15.915  < 2e-16 ***
BMI           0.3591     0.1341   2.678  0.00746 **
age          -0.1897     0.1131  -1.678  0.09351 .
nonwhiteyes   5.2194     2.3237   2.246  0.02477 *
smokingyes    4.7507     2.2104   2.149  0.03170 *
drinkanyyes  -2.7224     1.4989  -1.816  0.06944 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard deviation: 37.65 on 2739 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared: 0.01077
F-statistic: 5.966 on 5 and 2739 DF,  p-value: 1.7e-05
     AIC      BIC
27717.03 27758.45
```

After adjustment for these four demographic and lifestyle factors, the estimated increase in average LDL is 0.36mg/dL per unit increase in BMI, an association that remains highly statistically significant. In addition, average LDL is estimated to be 5.2mg/dL higher among nonwhite women, after adjustment for between-group differences in BMI, age, smoking, and alcohol use. The association of smoking with higher LDL is also statistically significant, and there is some evidence for lower LDL among older women and those who use alcohol. In addition, average LDL is estimated to be 5.2mg/dL higher among nonwhite women, after adjustment for between-group differences in BMI, age, smoking, and alcohol use. The association of smoking with higher LDL is also statistically significant, and there is some evidence for lower LDL among older women and those who use alcohol. In this example, smoking is a negative confounder, because women with higher BMI are less likely to smoke, but both are associated with higher LDL.

Mediation

If the predictor is cause of one of the covariates which in turn affects the outcome, this will be an instance of *mediation*. For the adjusted model for $LDL \sim BMI + age + nonwhite + smoking + drinkany$ we assumed that age, race/ethnicity, smoking, and alcohol use might confound the effect of BMI, because they affect both BMI and LDL levels, or are proxies for factors that do. However, if the primary predictor is a cause of one of the covariates, which in turn affects the outcome, this would be an instance of mediation.

Overall and Direct Effects

If the indirect pathway exists, and confounding has been controlled, then the coefficient for the primary predictor before adjustment for the mediator has a causal interpretation as the *overall effect* of the primary predictor on the outcome. The coefficient adjusted for the mediator is interpretable as the so-called *direct effect* of the primary predictor via other pathways that do not involve the mediator. Finally, the *difference* between overall and direct effects of the primary predictor is interpretable as the *indirect effect*.

Percent Explained

The relative difference between the overall and direct effects is sometimes referred to as the percent explained (PE) and used as an additional summary measure of the indirect effect.

Example: BMI, exercise, and Glucose

We examined the extent to which the effects of BMI on glucose levels might be mediated through its effects on likelihood of exercise. Although exercise may in some cases affect BMI, in HERS exercise was weakly associated (p = 0.06) with a small increase in BMI over the first year of the study. As a result, we would argue that in this population of older women with established heart disease, BMI mainly affects likelihood of exercise, with very little feedback. Thus, mediation of the effects of BMI by exercise makes sense in terms of a hypothesized causal framework. We recognize that our simple models might not completely control confounding of the relationships among BMI, exercise, and glucose, and could be improved with expert input.

```
# second half of Table 4.13
# example of LDL modeled with BMI and other factors
Gluc_fit_BMIall <- lm(glucose ~ BMI + age10 + nonwhite + smoking + drinkany + poorfair, data = hers_nodi
summary(Gluc_fit_BMIall)
```

```

Call:
lm(formula = glucose ~ BMI + age10 + nonwhite + smoking + drinkany +
    poorfair, data = hers_nodi)

Residuals:
    Min      1Q  Median      3Q     Max
-48.025  -6.367  -0.993   5.536  31.978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.63278    2.68721  28.890   <2e-16 ***
BMI          0.50256    0.04148  12.115   <2e-16 ***
age10        0.70940    0.32596   2.176   0.0296 *
nonwhiteyes  0.88015    0.76108   1.156   0.2476
smokingyes   0.18126    0.61352   0.295   0.7677
drinkanyyes  0.71373    0.43050   1.658   0.0975 .
poorfairyes -0.20525    0.53942  -0.381   0.7036
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.407 on 2018 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.07042,   Adjusted R-squared:  0.06766
F-statistic: 25.48 on 6 and 2018 DF,  p-value: < 2.2e-16
```

```
# and the second model
Gluc_fit_BMIexer <- lm(glucose ~ BMI + age10 + nonwhite + smoking + drinkany + poorfair + exercise, data
summary(Gluc_fit_BMIexer)
```

```

Call:
lm(formula = glucose ~ BMI + age10 + nonwhite + smoking + drinkany +
    poorfair + exercise, data = hers_nodi)

Residuals:
    Min      1Q  Median      3Q     Max
-47.574  -6.421  -0.879   5.515  32.417

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 78.86342    2.74136  28.768   <2e-16 ***
BMI          0.48597    0.04211  11.540   <2e-16 ***
age10        0.66558    0.32624   2.040   0.0415 *
nonwhiteyes  0.83154    0.76066   1.093   0.2744
smokingyes  -0.06125    0.62260  -0.098   0.9216
drinkanyyes  0.69540    0.43017   1.617   0.1061
poorfairyes -0.33879    0.54225  -0.625   0.5322
exerciseyes -0.97625    0.44020  -2.218   0.0267 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.398 on 2017 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.07268,   Adjusted R-squared:  0.06947
F-statistic: 22.59 on 7 and 2017 DF,  p-value: < 2.2e-16
```