General Linear Model

F. A. BarriosInstituto de Neurobiología UNAM

2020-10-18

Outline

1

23

```r
setwd("~/Dropbox/Fdo/ClaseStats/RegressionClass/RegressionR_code")
# To set the working directory at the user dir
library(tidyverse)
```

```r
library(multcomp)
```

```r
library(car)
```

```r
library(emmeans)

hers <- read_csv("DataRegressBook/Chap3/hersdata.csv")
```

General Linear Model GLM (Modelo lineal General)

GLM

Response variable Y is a random variable that is measured and has a distribution with expected value $E(Y|x)$ given a set of independent variables x.

$$Y_j(j = 1, ..., J)$$

for a set of x_jl predictor variables (or independent variables) defined as vectors for each j

$$x_jl(l = 1, ..., L)$$

with $L(L < J$, a general linear model with an error function ϵ_j can be expressed:

$$Y_j = x_{j1}\beta_1 + x_{j2}\beta_2 + x_{j3}\beta_3 + ... + x_{jL}\beta_L + \epsilon_j$$

with ϵ_j an independent variable identically distributed to the Normal with mean equal to zero.

$$\epsilon_j \approx N(0, \sigma^2)_{iid}$$

Example of simple linear regression: exercise and glucose Glucose levels above 125 mg/dL are diagnostic of diabetes, while 100-125 mg/dL signal increased risk. Data from HERS (public data) has baseline of glucose levels among 2,032 participants in a clinical trial of Hormone Therapy (HT). Women with diabetes are excluded, to study if the exercise might help prevent progression to diabetes.

```r
# hers data structure
hers_nodi <- filter(hers, diabetes == "no")
hers_nodi_Fit <- lm(glucose ~ exercise, data = hers_nodi)
```

```r
# the linear model results can be printed using summary
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
```

Simple linear regression model shows coefficient estimate (Coef) for exercise 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.

```r
#  Example of the HERS data for diabetic participants
hers_yesdi <- filter(hers, diabetes == "yes")
hers_yesdi <- mutate(hers_yesdi, physact = factor(physact, levels=c("much less active","somewhat less ac
#  Example of ANOVA with HERS data for diabetic participants
#
ggplot(data = hers_yesdi, mapping = aes(x = physact, y = glucose)) + geom_boxplot(na.rm = TRUE)
glucose_yesdi_act <- lm(glucose ~ physact, data = hers_yesdi)
Anova(glucose_yesdi_act, type="II")
```

```
Anova Table (Type II tests)

Response: glucose
           Sum Sq  Df F value Pr(>F)
physact     17992   4   1.925 0.1044
Residuals 1696313 726               
```

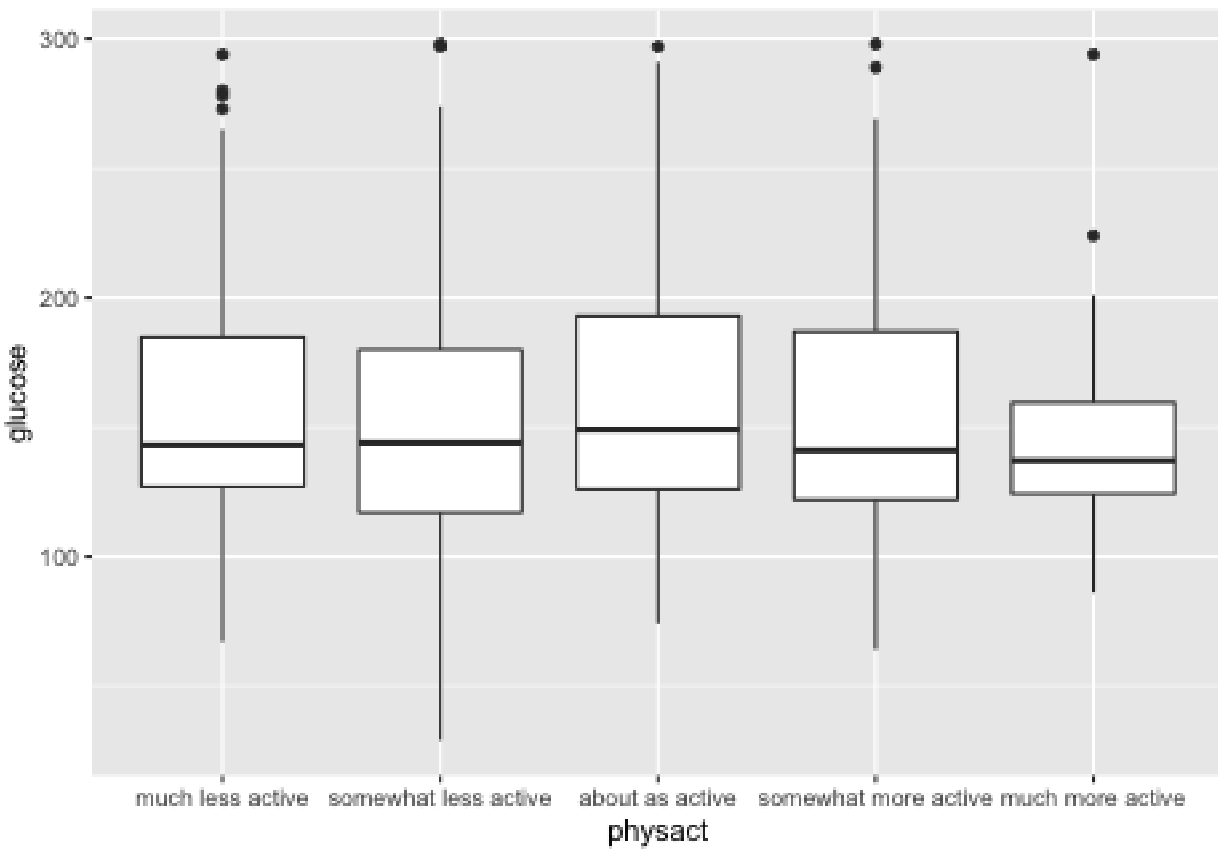
```r
#
S(glucose_yesdi_act)
```

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

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   155.789      5.095  30.575   <2e-16 ***
physactsomewhat less active    -4.590      6.235  -0.736    0.462    
physactabout as active          5.191      5.958   0.871    0.384    
physactsomewhat more active    -1.398      6.362  -0.220    0.826    
physactmuch more active       -11.789      8.320  -1.417    0.157    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard deviation: 48.34 on 726 degrees of freedom
Multiple R-squared: 0.0105
F-statistic: 1.925 on 4 and 726 DF,  p-value: 0.1044
    AIC     BIC
7751.41 7778.98
```

```r
glucose_emmeans <- emmeans(glucose_yesdi_act, "physact")
contrast(glucose_emmeans, adjust="sidak")
```

```
 contrast                    estimate   SE  df t.ratio p.value
 much less active effect         2.52 4.45 726   0.565  0.9856
 somewhat less active effect    -2.07 3.46 726  -0.599  0.9815
 about as active effect          7.71 3.16 726   2.441  0.0722
 somewhat more active effect     1.12 3.60 726   0.311  0.9991
 much more active effect        -9.27 5.50 726  -1.687  0.3830

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

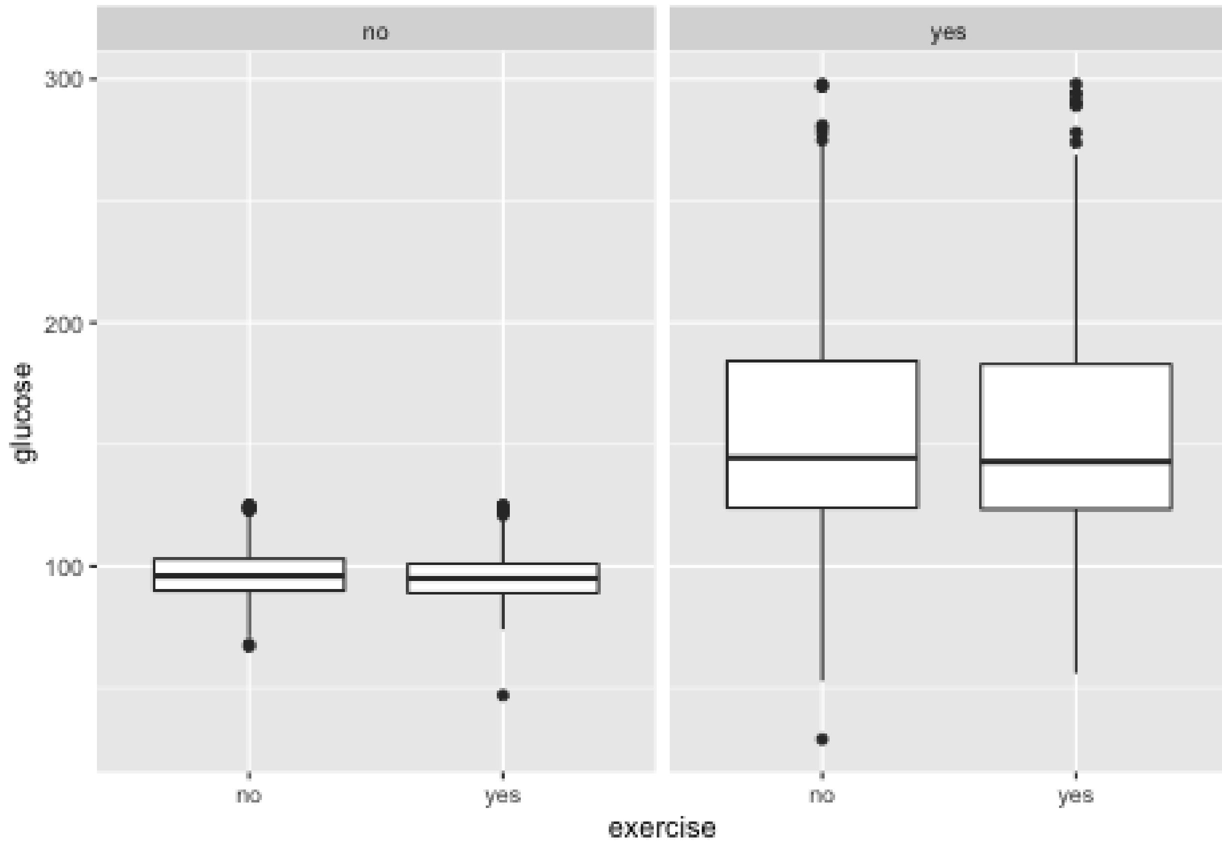


More Box Plots

```r
summary(hers$diabetes)
```

```
   Length     Class      Mode 
     2763 character character 
```

```
ggplot(data = hers, mapping = aes(x = exercise, y = glucose)) + geom_boxplot() + facet_grid( . ~ diabet
```



For a multiple linear model

There are models to regress several predictor variables to relate several random independent variables.

$$y_i = E[y_i|x_i] + \epsilon_i$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Multiple linear regression model coefficients, the betas, give the change in $E[Y|x]$ for an increase of one unit on the 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 (beta zero) gives the value $E[Y|x]$ when all the predictors are equal to zero. Example of multiple linear model estimate is done with: glucose ~ exercise + age + drinkany + BMI.

In general in R we can write:$Y = variable_1 + variable_2 + variable_3$ for a multiple linear model.

```
hers_nodi_multFit <- lm(glucose ~ exercise + age + drinkany + BMI, data = hers_nodi)
# the linear model results can be printed using summary
summary(hers_nodi_multFit)
```

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

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-47.560  -6.400  -0.886   5.496  32.060

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 error: 9.389 on 2023 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.07197,   Adjusted R-squared:  0.07013
F-statistic: 39.22 on 4 and 2023 DF,  p-value: < 2.2e-16
```

Multiple linear model, with interactions

In general in R we can write:$Y = variable_1 + variable_2 + variable_1 : varible_2$ for a multiple linear model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

(The following is a very good link: http://www.sthda.com/english/articles/40-regression-analysis/)