Multiple regression with interaction terms

Luis Castro-de-Araujo[a]

11/07/2022

Virginia Institute for Psychiatric and Behavioral Genetics

[a] Post-doc T32. luis.araujo@vcuhealth.org

Multiple regression recap

Interaction terms

Visualizing interactions

Marginal effects

Multiple regression recap

- Extension of the simple linear regression model to two or more independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- Expression = Baseline + Age + Tissue + Sex + Error
- Partial Regression Coefficients: ieffect on the dependent variable when increasing the *i*th independent variable by 1 unit, **holding all other predictors constant**

Categorical Independent Variables

- Qualitative variables are easily incorporated in regression framework through ***dummy variables***
- Simple example: sex can be coded as 0/1
- What if my categorical variable contains three levels:

$$x_1 = \begin{cases} 0 \text{ if } \mathrm{AA} \\ 1 \text{ if } \mathrm{AG} \\ 2 \text{ if } \mathrm{GG} \end{cases}$$

Categorical Independent Variables

- Previous coding would result in ***colinearity***
- Solution is to set up a series of dummy variable.
- for k levels you need k-1 dummy variables

	x1	x2
AA	1	0
AG	0	1
GG	0	0

$$x_1 = \begin{cases} 1 \text{ if } AA \\ 0 \text{ otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if } AG \\ 0 \text{ otherwise} \end{cases}$$

Assumptions

Validity Does the data we're modeling matches to the problem we're actually trying to solve?

Representativeness Is the sample data used to train the regression model representative of the population to which it will be applied?

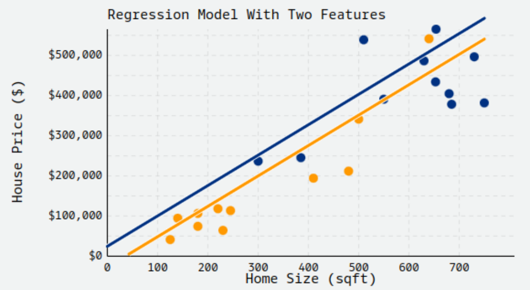
Additivity and Linearity The deterministic component of a regression model is a linear function of the separate predictors: $y = B_0 + B_1 x_1 + \ldots + B_p x_p$

Independence of Errors The errors from our model are independent.

Homoscedasticity The errors from our model have equal variance.

Normality of Errors The errors from our model are normally distributed.

MULTIVARIATE REGRESSION

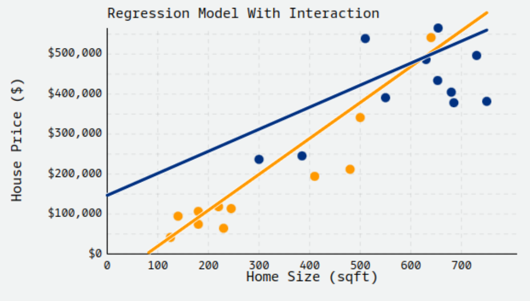


house price=-27154+757*sqft+51867*pool

- In our example, we model home prices as a function of both the size of the house (sqft) and whether or not it has a pool

- intercept: -$27,154, the predicted average housing price for houses with all x_i = 0. Or the cost of houses with no pools and a square-footage of zero.
- coefficient of pool: $51,867, average expected price difference in houses of the same size (in sqft) if they do or do not have a pool. In other words, we expect, on average, houses of the same size to cost $51,867 more if they have a pool than if they do not.
- coefficient of sqft: $757, average expected price difference in housing price for houses that have the same value of pool but differ in size by one square-foot.
- We assume the same slope for sqft.Hence, two lines. This isn't always a valid assumption to make.

Interpretation 4



- house price=-70296+899*sqft+217111*pool-347*(sqft:pool)
- If we believe that the slope for sqft should differ between houses that do have pools and houses that do not, we can add an interaction term to our model, (sqft:pool).

- interaction term: -$347, represents the difference in the slope for sqft, comparing houses that do and do not have pools. Visually, this represents the difference between the slopes of the two lines.
- intercept: -$70,296, represents the predicted housing price for houses with no pools and a square-footage of zero.
- coefficient of pool: $217,111, represents the average expected difference in houses of the same size (0 sqft) that differed in whether or not they had a pool. (It's not super useful since we don't have houses with 0 square-feet).
- coefficient of sqft: $899, represents the average expected difference in housing price for houses that do not have a pool (pool= 0) but differ in size by one square-foot.

Interaction terms

What is an Interaction?

- An interaction is a predictor that is some combination of the other predictors.

- Interactions are often the product of two or more predictors.
- Can be written as,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Conditional vs. Marginal Effects

- Conditional effects: the effect of a predictor on the response, holding all other predictors constant.
- Marginal effects: the effect of a predictor on the response, averaged over all values of the other predictors.

- If the conditional effects of X1 on Y at different levels of X2 are all the same then there is no interaction.

Interpreting Interaction Effects

Parameter	Meaning	Where people (used to) go awry
β_0	Expected value of the DV when X1 and X2 ==0	People get this
β_1	Effect of X1 when X2 == 0	Not marginal effects!
β_2	Effect of X2 when X1 == 0	Not marginal effects!
β_3	The addition to the conditional effect when both X! and X2 are 1	People just look at the significance of the interaction parameter and do not calculate the underlying marginal or conditional effects or standard errors

In the past it was common to see standard errors wrongly calculated

- A common mistake that people make when interpreting interaction models is using the wrong standard errors.
- The standard errors that are printed in every regression table are the positive square roots of the diagonal elements of the variance- covariance matrix of β
- This does not matter anymore because of `margins()`

VISUALIZING INTERACTIONS

Having fun with mtcars - regression of speed on wt*cyl

```r
fit <- glm(qsec ~ wt*cyl, data = mtcars)
summary(fit)
```

```
Call:
glm(formula = qsec ~ wt * cyl, data = mtcars)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-2.1966   -0.8373   0.0499   0.8158   2.1398

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.726      3.118    5.36  0.00001 ***
wt             2.858      1.180    2.42    0.022 *
cyl           -0.542      0.511   -1.06    0.298
wt:cyl        -0.222      0.167   -1.33    0.193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.45)

    Null deviance: 98.988  on 31  degrees of freedom
Residual deviance: 40.636  on 28  degrees of freedom
AIC: 108.5

Number of Fisher Scoring iterations: 2
```

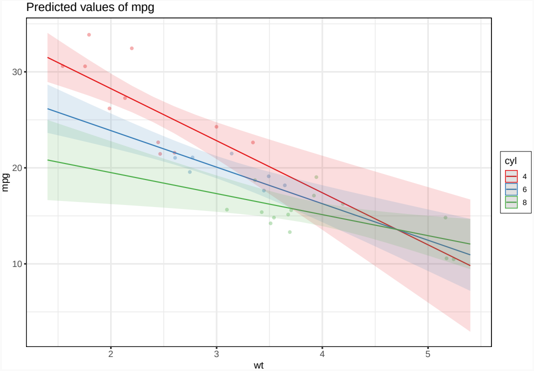
Regression of MPG on WT*CYL

```r
fit <- glm(mpg ~ wt*cyl, data = mtcars)
summary(fit)
```

```
Call:
glm(formula = mpg ~ wt * cyl, data = mtcars)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-4.229  -1.350  -0.504   1.465   5.234

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   54.307      6.128    8.86 0.0000000013 ***
wt            -8.656      2.320   -3.73      0.00086 ***
cyl           -3.803      1.005   -3.78      0.00075 ***
wt:cyl         0.808      0.327    2.47      0.01988 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.61)

    Null deviance: 1126.05  on 31  degrees of freedom
Residual deviance:  156.98  on 28  degrees of freedom
AIC: 151.7

Number of Fisher Scoring iterations: 2
```

Regression of mpg on wt*cyl

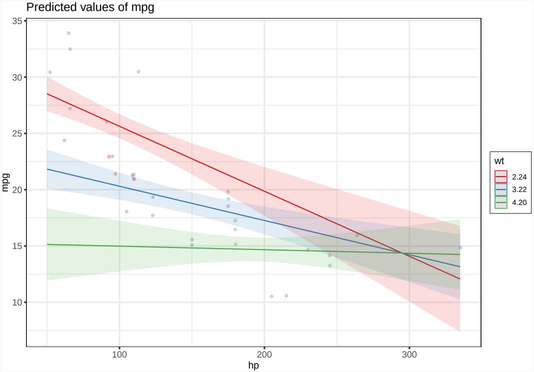
```r
pred <- ggpredict(fit, terms = c("wt", "cyl"))
plot(pred, add.data = TRUE)+
  theme_luis()
```



Now, from what point the slope becomes non signficant?

Changing to mpg ~ hp + wt

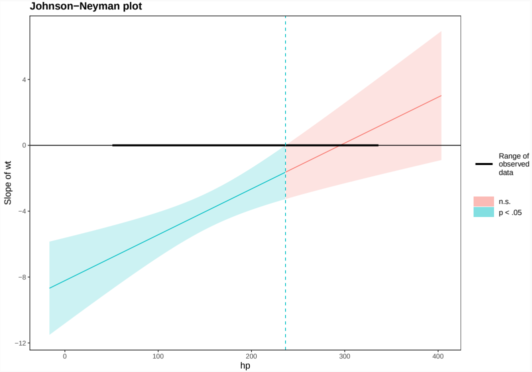
```r
fit <- glm(mpg ~ hp*wt, data = mtcars)
pred <- ggpredict(fit, terms = c("hp", "wt"))

plot(pred, add.data = TRUE) +
  theme_luis()
```



Now, from what point the slope becomes non signficant?

JOHNSON-NEYMAN INTERVAL

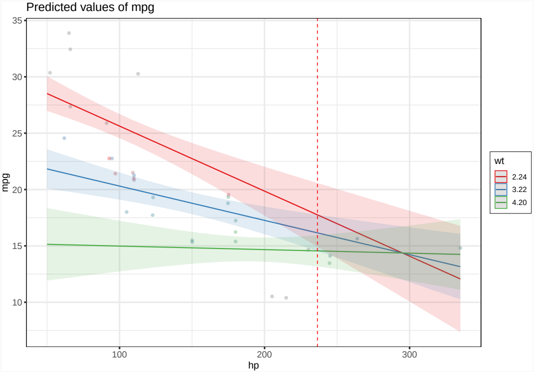
```r
jn <- johnson_neyman(fit, wt, hp , plot = TRUE)
jn
```



Now, from what point the slope becomes non signficant?

JOHNSON-NEYMAN INTERVAL - Overlayed over data

```
fit <- glm(mpg ~ hp*wt, data = mtcars)
pred <- ggpredict(fit, terms = c("hp", "wt"))
jn<-johnson_neyman(fit, wt, hp , plot = TRUE)
jn_bound<-as.numeric(jn$bounds[1])

plot(pred, add.data=T) +
  geom_vline(xintercept = jn_bound, linetype = "dashed",
             color = "red")+
  theme_luis()
```

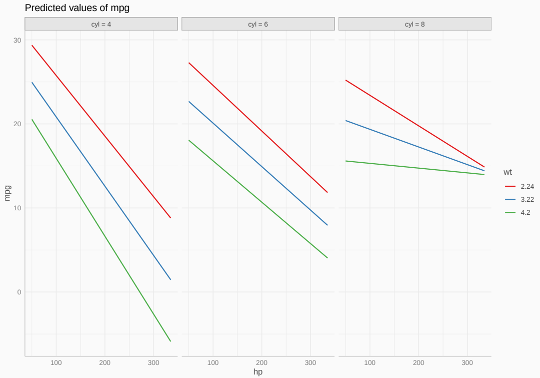


Three-way interactions

```r
fit <- glm(mpg ~ hp*wt*cyl, data = mtcars)
```

```r
dat <- ggpredict(fit, terms = c("hp", "wt", "cyl"))
plot(dat, ci = FALSE)
```

```
Call:
glm(formula = mpg ~ hp * wt * cyl, data = mtcars)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-3.352   -1.464   -0.169    1.345    4.001

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.96543   30.32070    1.45     0.16
hp          -0.02587    0.24000   -0.11     0.92
wt          -2.25515   10.68401   -0.21     0.83
cyl         -0.52189    6.33725   -0.08     0.94
hp:wt       -0.03666    0.09360   -0.39     0.70
hp:cyl      -0.00569    0.03850   -0.15     0.88
wt:cyl      -0.42991    1.99058   -0.22     0.83
hp:wt:cyl    0.00654    0.01375    0.48     0.64
```



Marginal effects

- Marginal effects: the effect of a predictor on the response, averaged over all values of the other predictors.
- It is achieved by..

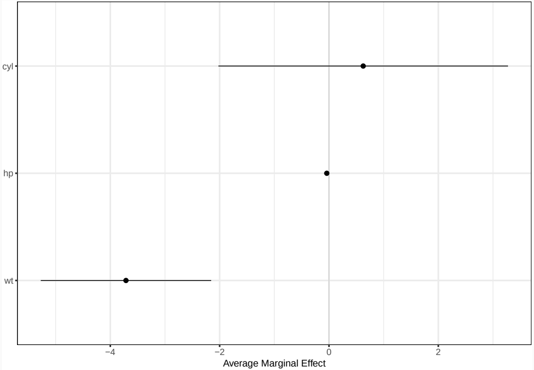
What are the marginal effects of the latest model?

```r
fit_m <- margins(fit)
summary(fit_m)
```

factor	AME	SE	z	p	lower	upper
cyl	0.6261	1.3513	0.4633	0.6431	-2.0224	3.2745
hp	-0.0402	0.0152	-2.6390	0.0083	-0.0700	-0.0103
wt	-3.7134	0.7952	-4.6696	0.0000	-5.2720	-2.1547

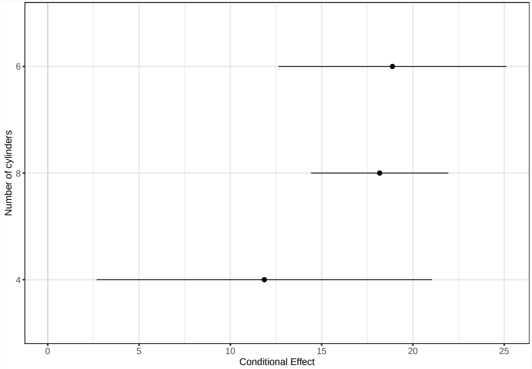
Plotting the marginal effects

```r
fit_mo <- as_tibble(summary(fit_m))
p <- ggplot(data = fit_mo, aes(x = reorder(factor, AME),
                               y = AME, ymin = lower, ymax = upper))

p + geom_hline(yintercept = 0, color = "gray80") +
    geom_pointrange() + coord_flip() +
    labs(x = NULL, y = "Average Marginal Effect") +
    theme_luis()
```



Can we plot the conditional effects too?

```r
glm(mpg ~ hp*wt*as.factor(cyl), data = mtcars) %>%
  cplot(x = "cyl", draw = F) %>%
  ggplot( aes(x = reorder(xvals, yvals),
                    y = yvals, ymin = lower, ymax = upper))+
                    geom_hline(yintercept = 0, color = "gray80") +
                    geom_pointrange() + coord_flip() +
    labs(x = "Number of cylinders", y = "Conditional Effect") +
    theme_luis()
```



- It is sensitive to the scale of the predictors[1]
  - litres/100 km and miles/gallon may generate different results
- At the beginning I said, keep the interaction that is significant, but[2]
  - In the conversion to probabilities (AME) the interaction may not be significant anymore, or worse
  - The interaction may be significant in the AME, but not in the original model

[1] **???**

[2] **???**

Conclusion

- We started reviwing multiple regression
- Then discussed the syntax and interpretation of parameters when an interaction term is included
- Finally, we discussed how to extract the marginal effects of the interaction term
- Luckly the package margins() makes this extremely simple, thus lectures on interaction became much shorter than it used to be (if you want a hands on approach).

Acknowledgements

Team

- Charles Gardner (2015)
- Brad Verhulst (2013)
- Joshua Pritkin.
- Rob Kirkpatrick.

- Michael C Neale.
- NIH grant no R01 DA049867 and 5T32MH-020030

Contact

- Thank you