**جلسه نهم  
گردآورنده: پریناز خسروانی**

**مباحث جلسه:**

* Covariance analysis
* Correlation analysis
* Regression analysis
* General linear modeling
* Interaction modeling

مرور جلسات گذشته:  
طی جلسات قبل درباره مقایسه دو و چند گروهی و اشکال مختلف آن ها صحبت شد. که در تمام آن ها متغیر وابسته ما از نوع کمّی پیوسته بود، مثل “FIQ”، و یا از نوع کیفی ترتیبی (Ordinal)، مثل نمرات“ADOS” (که در این حالت و برای متغیر ترتیبی از آمار ناپارامتریک استفاده شد).   
و متغیر مستقل ما از نوع کیفی اسمی (nominal) بود و نشان دهنده یک گروه خاص بود، مثل سایت “NYU”. و ما متغیر وابسته خود را بین گروه های مختلف مقایسه کردیم.

در این جلسه:  
 به مدلی از آمار میپردازیم که در آن هر دو متغیر وابسته و مستقل از نوع کمّی پیوسته (ratio or interval) هستند.   
مثال: مقایسه سن و IQ. بطور کلی در این حالت هر دو متغیر ما مقادیر مختلفی از عدد هستند.

در این جلسه از پکیج های (ggplot2) و (GGally) استفاده میکنیم.   
در صورت نیاز میتوانید برای نصب پیش نیاز های هر پکیج، در کنسول Rstudio این کد را ران کنید:

>install.packages(“GGally”, dependencies = T)  
>install.packages(“ggplot2”, dependencies = T)

بعد از انجام مقدمات، سابجکت های مد نظر خود را از فایل phenotypic.csv انتخاب میکنیم.   
در این جلسه ما تنها از سابجکت های بیمار“ASD” استفاده میکنیم:

##ASD subjects  
  
pheno <- read.csv("Phenotypic.csv")  
pheno[pheno == "-9999"] <- NA  
pheno <- pheno[pheno$DX\_GROUP == 1,]

متغیر های مورد بررسی ما در این سابجکت ها شامل AGE\_AT\_ASCAN، FIQ ، ADOS\_TOTAL و SEX میشوند.  
هر یک از این متغیر ها یک ستون در دیتای فنوتیپیک هستند که ما ان ها را جداگانه تبدیل به متغیر هایی با نام جدید میکنیم:

#Selecting interesting variables  
  
x1 <- pheno$AGE\_AT\_SCAN #var type: ratio   
x2 <- pheno$FIQ #var type: interval  
x3 <- pheno$ADOS\_TOTAL #var type: ordinal  
x4 <- as.factor(pheno$SEX) #var type: nominal

**نکته:** متغیر SEX به صورت اعداد 1 و 2 تعریف شده اما ما میدانیم که ذات این متغیر اسمی است. برای همین از دستور as.factor استفاده میکنیم.

در ادامه قرار است از این متغیر های x1 تا x4 برای ساختن دیتا فریم های جدید برای مقایسه با هم استفاده کنیم. (سه متغیر اول پیوسته هستند)

* Covariance & Correlation analysis:

وقتی متغیر های پیوسته داریم میتوانیم ارتباط بین آن هارا دو به دو با هم بررسی کنیم. (یعنی یک متغیر مستقل و دیگری وابسته). در این حالت ما میتوانیم از آنالیز های آماری زیر استفاده کنیم:

Covariance  
Correlation  
Regression

در این جلسه به بررسی این سه مدل آنالیز می پردازیم.

تعریف همبستگی کوواریانس: وقتی مقدار افت و خیز در یک متغیر مرتبط است با افت و خیز در متغیری دیگر. این ارتباط میتواند در جهت مثبت یا منفی (یا صفر) باشد.

در اینجا جمعیت مورد بررسی ما، که افراد دارای اتیسم هستند، باتوجه به متغیر های منتخب بطور درون گروهی باهم مقایسه میشوند. یعنی بطور مثال میخواهیم بدانیم که آیا در افراد دارای اتیسم کم یا زیاد بودن سن افراد با میزان هوش آن ها همبستگی دارد یا خیر. توجه کنید که این متغیر ها حتما بین سابجکت هایی از یک گروه خاص سنجیده میشوند.

تعریف همبستگی کوریلیشن: کوریلیشن حالت نرمال شده (normalized) کوواریانس است. که عددی بین منفی یک و یک میباشد.

* به نمودار های correlation زیر توجه کنید:

A diagram of scatter plots and correlations

Description automatically generated

تصور کنید هر نقطه نشان دهنده یک سابجکت یا فرد است. ما از هر یک از این افراد دو متغیر xوy داریم (فرض میکنیم متغیر x ما سن است و متغیر y نشان دهنده هوش افراد). پس هر نقطه روی این نمودار همزمان نشان دهنده میزان سن و هوش فرد میباشد.

همانطور که گفتیم مقدار کوریلیشن یک عدد بین 1- تا 1+ میشود.  
هرگاه اسکتر پلاتی(نمودار نقطه ای) داشتیم که در آن همه نقاط بطور بی نقصی روی یک خط مورب قرار گرفتند، آنگاه کوریلیشن ما برابر است با 1- یا 1+ .  
وقتی هر دو متغیر xوy باهم افزایش یا کاهش پیدا کنند این **همبستگی مثبت (ارتباط مستقیم)** است. اما اگر با افزایش یک متغیر دیگری کاهش یابد این **همبستگی منفی (ارتباط غیرمستقیم)** است.

البته دیتای تجربی عموما مثل نمودار های کوریلیشن بی نقص سمت چپ و راست تصویر نیستند. و معمولا پراکندگی بیشتری دارند.

حالا اگر هیچ گونه الگویی بین دو متغیر ما وجود نداشته باشد، مثل نمودار قرمز وسط، آنگاه **همبستگی صفر (بدون ارتباط)** میباشد.

**پرسش و پاسخ:**آیا میتوانیم با استفاده از کوریلیشن و نمودار نقطه ای داده های پرت (outlier) را پیدا کرده و آن ها را حذف کنیم؟  
خیر، معمولا داده های پرت را در مرحله قبل از انالیز همبستگی جدا میکنیم. یعنی در مرحله ای که پراکندگی هر یک از متغیر های xوy بطور جداگانه بررسی میشود.   
مثلا نمودار هیستوگرام متغیر x (سن) را رسم کرده و سپس داده هایی که به اندازه 3 انحراف معیار (standard deviation) متفاوت هستند را حذف میکنیم. در ادامه همین روند را برای متغیر y (هوش) نیز انجام داده و در نهایت میتوانیم از نمودار نقطه ای برای بررسی همبستگی این دو متغیر استفاده کنیم.   
اما اگر در نمودار نقطه ای هم داده ای مشاهده شد که در یک گوشه دور افتاده از بقیه نقاط بود، میتوان حدس زد که آن داده پرت است.

حال به بررسی کوورایانس میان دو متغیر x1 و x2 که به ترتیب سن و هوش جمعیت مد نظر مارا نشان میدهند میپردازیم:

#Covariance analysis  
cov(x1,x2)  
[1] NA

دقت کنید که اگر بار اول این کد را بدون حذف کردن داده های ناموجود (NA) اجرا کنید به شما نتیجه NA میدهد. پس برای محاسبه correlation و covariance باید کد را بشکل زیر تغییر دهیم که این داده های ناموجود را از بردار های x و y حذف کند:

> cov(x1,x2,use = “na.or.complete”)

[1] 19.67828

> cor(x1,x3,use = “na.or.complete”)

[1] 0.05438882

**یادآوری:** برای محاسبه میانگین نیز از دستور na.remove برای حذف NA استفاده میکردیم.

توضیح اعداد بدست امده برای cor و cov:   
عدد 19.67828 برای کوواریانس نشان میدهد که بین متغیر x1 (سن) و متغیر x2 (هوش) یک رابطه مستقیم یا همبستگی مثبت وجود دارد. (ولی هنوز نمیتوانیم استدلال دقیق تری از میزان این رابطه مثبت داشته باشیم تا زمانی که از کورریلیششن استفاده کنیم)

در خط بعدی که با دستور cor کوریلیشن بین x1 و x3 (نمره کل ADOS) محاسبه شد و عدد 0.05438882 بدست آمد. این عدد بسیار نزدیک به صفر است، پس با اینکه همبستگی مثبت است اما تقریبا برابر صفر است و میتوان ان را به عنوان نبود همبستگی تفسیر کرد.   
سپس با اجرای کد های زیر میتوانیم نمودار نقطه ای متغیر هارا دو به دو ببینیم:

> plot(x1,x2) # AGE & FIQ

A graph of dots and numbers

Description automatically generated with medium confidence

> plot(x2,x3) # FIQ & ADOS\_TOTAL

A graph of numbers and circles

Description automatically generated with medium confidence

همانطور که گفتیم، هر کدام از نقطه ها مربوط به یک سابجکت است.  
در نمودار x1 و x2 هیچ ترند خاصی دیده نمیشود (بطوری که بتوان یک خط اریب را روی ان فیت کرد). و این نشان دهنده کورریلیشن نزدیک صفر است.

در نمودار x2 و X3 نیز ترند واضحی دیده نمیشود اما میتوان حدس زد احتمالا یک کورریلیشن منفی نزدیک به صفر داریم. این را با کد زیر بررسی میکنیم:

> cor(x2,x3,use = “na.or.complete”)

[1] -0.1874356

همانطور که حدس زدیم عدد کوریلیشن ما منفی است و نشان دهنده رابطه غیر مستقیم بین این دو متغیر است. هرچند بسیار کوچک است.

**تمرین 1:**اگر بخواهیم میزان معنا داری عدد کوریلیشن را بدست بیاوریم نیازمند محاسبه اعداد effect size و   
 p-valueهستیم. برای این تمرین شما باید نحوه محاسبه این دو عدد و تفسیرشان را در جهت محاسبه میزان معناداری کوریلیشن پیدا کنید. پس در نهایت باید بتوانید از روی مقادیر افکت سایز و p-value میزان قابل استناد بودن یک کوریلیشن را بیان کنید.   
معناداری و قابل استناد بودن کوریلیشن های زیر را گزارش کنید:

> cor(x1,x2,use = “na.or.complete”)

[1] 0.1423377

> cor(x1,x3,use = “na.or.complete”)

[1] 0.05438882

> cor(x2,x3,use = “na.or.complete”)

[1] -0.1874356

**پرسش و پاسخ:**

در نمودار اسکتر پلات، فاصله هر نقطه (سابجکت) با خط نشان دهنده کوریلیشن در نمودار نشان دهنده چیست؟

به این فواصل residual گفته میشود. اگر کوریلیشن بین این دو متغیر واقعا وجود داشته باشد، این residual ها میتوانند نشان دهنده وجود متغیر یا متغیر های تاثیرگذار دیگری بر رابطه بین x1 و x2 باشند. به عبارت دیگر، فاصله هر سابجکت با خط رگرسیون یا کوریلیشن میتواند حاکی از وجود مداخله کننده هایی در این معادله باشد که سبب شده رابطه بین دو متغیر مد نظر کاملا به صورت یک به یک بروز نکند (یعنی بطوری که ضریب همبستگی 1 یا -1 شود).  
در این موقعیت ما میتوانیم بطور multi-variate داده ها را تحلیل کنیم. یعنی روابط سه یا چند متغیر را بطور همزمان با هم بررسی کنیم تا به پیش بینی یا توضیح دقیق تری از داده ها برسیم.

**یادآوری:** هنگامی که متغیر های پیوسته (ratio or interval) داشته باشیم و تابع توزیع آن ها نیز نرمال باشد از آمار پارامتریک استفاده میکنیم.

حالا در این بخش از درس میخواهیم کوریلیشن و کوواریانس را در کانتکست آمار ناپارامتریک بررسی کنیم.

* وقتی حتی یکی از متغیر های مورد بررسی از نوع ترتیبی باشد، یا تابع توزیع ان نرمال نباشد (یا هر دو)، ما از امار ناپارامتریک در محاسبه انواع همبستگی استفاده میکنیم.

همانطور که در جلسات گذشته دیدیم میتوانیم از shapiro.test برای بررسی نرمالیتی داده ها استفاده کنیم. مثل خطوط زیر:

> shapiro.test(x1)

Shapiro-Wilk normality test

data: x1

W = 0.82229, p-value < 2.2e-16

> shapiro.test(x2)

Shapiro-Wilk normality test

data: x2

W = 0.99462, p-value = 0.07372

تفسیر: متغیر x1 که سن سابجکت هاست دارای p-value < 2.2e-16 که بسیار کوچک تر از 0.05 است و این یعنی فرض صفر ما (نرمال بودن توزیع) رد شده و داده های سن بیماران توزیع غیر نرمال دارد و از امار ناپارامتریک استفاده میکنیم.  
برای متغیر x2، هوش سابجکت ها، نتیجه تست بزرگ تر از 0.05 است. این یعنی فرض صفر رد نشده و توزیع نمونه ها نرمال است.   
متغیر x3 هم از نوع ترتیبی است در نتیجه ناپارامتریک است.  
منتها وقتی به انالیز اماری بین این متغیر ها میرسیم باید از امار ناپارامتریک استفاده کنیم زیرا x1 و x3 نرمال نیستند.

در این حالت برای محاسبه از متد ("method") اسپیرمن ("spearman ") استفاده میکنیم:

cor(x2,x3,use = "na.or.complete", method = "spearman") #non-parametric

اما اگر هر دو متغیر از نوع پیوسته بودند و توزیع نرمال داشتند، انگاه از متد پیرسون ("pearson") استفاده میکنیم. (که در حالت عادی نیز این متد به شکل دیفالت در فانکشن cor استفاده میشود.)

* Regression analysis:

همانطور که گفته شد در رگرسیون خطی نیز تنها از متغیر های پیوسته استفاده میشود.  
رگرسیون خطی هم ارتباط بین دو متغیر را بررسی میکند اما ذاتا یک نوع تست اماری است (برخلاف کوریلیشن). یعنی یک سری پارامتر به ما میدهد و سعی میکند یک رابطه خطی بین دو متغیر ایجاد کند و ما ان رابطه خطی را بررسی میکنیم.

در ادامه یک دیتا فریم از 4 متغیر قبلی درست میکنیم و برای هر ستون آن که نشان دهنده یکی از این متغیر هاست نام گذاری میکنیم:

#Regression analysis

data <- data.frame(x1,x2,x3,x4)

colnames(data) <- c("AGE","FIQ","ADOS","SEX")

سپس یک مدل خطی از ارتباط دو متغیر میسازیم. بطور مثال در اینجا متغیر مستقل ما (که همیشه سمت راست علامت ~ است) AGE است و متغیر وابسته ما FIQ ، به عبارتی یعنی از این فانکشن میخواهیم سن افراد را گرفته و بر اساس ان پیش بینی کند هوش کل سابجکت چقدر است (پیش بینی هوش بر اساس سن افراد روی دیتای داده شده). در این صورت متغیر AGE در محور x و متغیر FIQ در محور y قرار میگیرد:

model <- lm(FIQ ~ AGE , data) #lm = linear model

summary(model)

با ران کردن کد summary(model) ما به خروجی تست اماری رگرسیون دست میابیم:

> summary(model)

Call:

lm(formula = FIQ ~ AGE, data = data)

Residuals:

Min 1Q Median 3Q Max

-64.259 -11.739 -0.767 11.362 45.091

Coefficients:

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

(Intercept) 100.36986 1.71538 58.512 < 2e-16 \*\*\*

AGE 0.29773 0.09232 3.225 0.00134 \*\*

---

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

Residual standard error: 16.85 on 503 degrees of freedom

(34 observations deleted due to missingness)

Multiple R-squared: 0.02026, Adjusted R-squared: 0.01831

F-statistic: 10.4 on 1 and 503 DF, p-value: 0.001341

**نکته:**کورریلیشن و کوواریانس نشان دهنده association (ارتباط) هستند. اما رگرسیون خطی causality (علیت) را بررسی میکند.

پس هنگام تفسیر و گزارش انالیز خود باید دقت کنیم تا عبارت مناسب را انتخاب کنیم.   
بطور مثال وقتی نتیجه آنالیز یک مدل خطی بین دو متغیر x و y را گزارش میکنیم میگوییم:  
 x نمیتواند/میتواند متغیر y را پیش بینی کند.  
و به این شکل نتیجه آنالیز کوریلیشن را گزارش میکنیم: بین متغیر x و y رابطه وجود دارد/ندارد.

**تفسیر:**  
همانطور که میبینید عدد p-value معنا دار است: (0.00134). یعنی ارتباط AGE با FIQ از لحاظ اماری قابل استدلال است و AGE میتواند FIQ را پیش بینی کند.  
عدد Estimate نیز شیب خط فیت شده روی داده را نشان میدهد (که همان خط ابی پلات رسم شده پایین است). اگر این عدد مثبت باید یعنی رابطه مستقیم است اما اگر منفی باشد رابطه غیر مستقیم است.  
بطور کلی وقتی یک مدل خطی ران میکنیم باید اعداد Estimate و p-value را گزارش کنیم.

**پرسش و پاسخ:**اگر p-value معنا دار نشود، انگاه ایا Estimate نیز صفر میشود؟ بله به سمت صفر میل میکند.

حالا برای اینکه درک بهتری از مدل داشته باشیم پلات آن را رسم میکنیم:

ggplot(data, aes(x = AGE, y = FIQ)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, col = "blue") +

labs(x = "Age",

y = "Full IQ") +

theme\_minimal()

A graph with a line and dotted line

Description automatically generated

* خط آبی که روی داده فیت شده است نماینده ی همان linear model ماست.

حالا دو متغیر ADOS و AGE را بررسی میکنیم:

model <- lm(ADOS ~ AGE , data)

> summary(model)

Call:

lm(formula = ADOS ~ AGE, data = data)

Residuals:

Min 1Q Median 3Q Max

-9.9097 -2.8053 -0.1346 2.3165 10.2012

Coefficients:

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

(Intercept) 11.43926 0.45984 24.877 <2e-16 \*\*\*

AGE 0.02622 0.02466 1.063 0.288

---

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

Residual standard error: 3.797 on 381 degrees of freedom

(156 observations deleted due to missingness)

Multiple R-squared: 0.002958, Adjusted R-squared: 0.0003412

F-statistic: 1.13 on 1 and 381 DF, p-value: 0.2884

**تفسیر:**

p-value معنا دار نیست: (0.288). یعنی AGE نمیتواند ADOS را پیش بینی کند.  
در نتیجه عدد Estimate نیز به صفر میل میکند.

به نمودار این مدل دقت کنید:

ggplot(data, aes(x = AGE, y = ADOS)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, col = "blue") +

labs(x = "Age",

y = "ADOS") +

theme\_minimal()

A graph with a line and a line

Description automatically generated

\* همانطور که میبینید خط آبی فیت شده روی نمودار شیب بسار کمی دارد و داده ها نیز پراکندگی بیشتری دارند.

* General linear modeling:

تا اینجا گفتیم که در مدل خطی (linear model) دو تا متغیر داریم که پیوسته هستند (یکی از ان ها میتواند ترتیبی هم باشد) و در آن ما متغیر وابسته را از روی متغیر مستقل پیش بینی میکنیم.

در مدل خطی جنرال ما همچنان فقط یک متغیر وابسته داریم، اما آن را از روی دو یا چند متغیر مستقل پیش بینی میکنیم.

در کد زیر میخواهیم بدانیم که آیا میتوانیم نمرات ADOS را از روی AGE و(+) FIQ پیش بینی کنیم:

> model <- lm(ADOS ~ AGE + FIQ , data)

> summary(model)

Call:

lm(formula = ADOS ~ AGE + FIQ, data = data)

Residuals:

Min 1Q Median 3Q Max

-10.148 -2.894 -0.077 2.516 11.358

Coefficients:

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

(Intercept) 15.79627 1.29344 12.213 < 2e-16 \*\*\*

AGE 0.05018 0.02572 1.951 0.051854 .

FIQ -0.04459 0.01170 -3.811 0.000162 \*\*\*

---

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

Residual standard error: 3.717 on 366 degrees of freedom

(170 observations deleted due to missingness)

Multiple R-squared: 0.04506, Adjusted R-squared: 0.03984

F-statistic: 8.635 on 2 and 366 DF, p-value: 0.0002166

**تفسیر:**

p-value :AGE معنا دار نیست: (0.051854). یعنی AGE نمیتواند ADOS را پیش بینی کند.  
p-value :FIQ معنا دار است: (0.000162). یعنی FIQ می تواند ADOS را پیش بینی کند.  
:Estimate FIQ منفی است (-0.04459). یعنی FIQ یک رابطه غیر مستقیم با ADOS دارد (با افزایش هوش افراد میتوانیم پیش بینی کنیم که عدد ADOS ان ها کاهش میابد).

در نظر داشته باشید با اینکه این مدل اثر هر یک از متغیر های مستقل در پیش بینی متغیر وابسته را بطور جداگانه گزارش کرده اما اثر هر یک از این متغیر ها متاثر از وجود متغیر مستقل دیگر نیز میباشد. در نتیجه اگر بطور جداگانه یک مدل خطی ساده بین هر یک از این متغیر ها با ADOS داشته باشید (مثل قسمت قبلی درس) خواهید دید که اعدادی متفاوت با اعداد اینجا برای معنا دار بودن هر یک به شما خواهد داد.

حالا با استفاده از دستور ggpair ، که از لایبری GGally استفاده میکند، برای رسم پلات این مدل استفاده میکنیم.   
در این مدل چون 3 تا متغیر داریم که هر یک یک ستون از دیتا فریم ما هستند، ابتدا ان ها را با دستور زیر سلکت میکنیم:

> ggpairs(data[,1:3])

A screenshot of a graph

Description automatically generated

این نمودار به ما 3 دسته داده میدهد:

یک: هیستوگرام هر یک از متغیر های موجود در مدل. (قطر نمودار)  
 دو: اسکتر پلات متغیر ها بشکل دو به دو. (مثلا گوشه پایین سمت چپ اسکتر پلات AGE و ADOS است)  
سه: مقدار کوریلیشن هر دو متغیر با یکدیگر.

**تفسیر:**

در نمودار بالا میبینیم که کوریلیشن بین AGE و FIQ معنادار است و همینوطور کوریلیشن بین FIQ و ADOS . این را میتوانیم از روی ستاره های جلوی این اعداد در نمودار متوجه شویم.

حالا میخواهیم نقش جنسیت را در این آنالیز ها متوجه شویم. یعنی میخواهیم بدانیم هر متغیر و ارتباط بین این متغیر ها بین دو جنس مختلف و بطور کلی به چه شکل است. برای این منظور با کد زیر در نموداری که رسم میکنیم به هر جنسیت یک رنگ جداگانه میدهیم:

ggpairs(data,

mapping = aes(color = SEX),

upper = list(continuous = wrap("cor", size = 4)),

lower = list(continuous = wrap("smooth", method = "lm", se = TRUE))

)

خط اخر (lower = …) برای هر جنسیت روی نمودار های اسکتر پلات بطور جداگانه یک خط فیت میکند.

A screenshot of a graph

Description automatically generated**تفسیر:**  
در این نمودار هیستوگرام ها به ازای هر جنسیت رسم شده است و بطور جداگانه برای هرکدام خطی فیت میکند.  
مثلا میتوانیم ببینیم برای دو متغیر FIQ و AGE بین دو جنس تفاوتی در شیب خط ها وجود دارد. اعداد کوریلیشن نیز نشان میدهد برای خانم ها (کد 2) همبستگی بیشتری بین این دو متغیر نسبت به اقایان وجود دارد. این مارا ترغیب میکند که یک آنالیز تکمیلی برای مقایسه دیتا بین دو جنس هم انجام دهیم.

* Interaction modeling:

در این مدل میخواهیم متغیر وابسطه را از روی متغیر های مستقل، و همچنین اینترکشن(\*) بین این متغیر ها پیش بینی کنیم.

اینترکشن به زبان ساده یعنی دوتا متغیر را در هم ضرب میکنیم. یعنی اگر جفت متغیر ها با هم زیاد باشن عدد بزرگتری بدست می آید و برعکس.

> model <- lm(ADOS ~ AGE \* FIQ , data)

> summary(model)

Call:

lm(formula = ADOS ~ AGE \* FIQ, data = data)

Residuals:

Min 1Q Median 3Q Max

-10.1744 -2.9523 -0.0713 2.5404 11.2825

Coefficients:

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

(Intercept) 13.632654 2.955999 4.612 5.53e-06 \*\*\*

AGE 0.184859 0.167425 1.104 0.270

FIQ -0.024714 0.027082 -0.913 0.362

AGE:FIQ -0.001232 0.001513 -0.814 0.416

---

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

Residual standard error: 3.719 on 365 degrees of freedom

(170 observations deleted due to missingness)

Multiple R-squared: 0.04679, Adjusted R-squared: 0.03896

F-statistic: 5.972 on 3 and 365 DF, p-value: 0.0005536

**تفسیر:**

میبینیم که به ما سه گزارش جداگانه میدهد. هیچ کدام از متغیر های مستقل(main effect) و نیز اینترکشن این دو با یکدیگر تاثیری روی ADOS ندارند و نمیتوانیم ان را پیش بینی کنیم.

در حالتی دیگر از انالیز ممکن است بخواهیم تنها تاثیر اینترکشن بین دو متغیر را روی متغیر وابسته بررسی کنیم و main effect متغیر های مستقل برایمان مهم نباشد. اینجا از این ترم (:) استفاده میکنیم. مثال:

> model <- lm(ADOS ~ AGE:FIQ , data)

> summary(model)

Call:

lm(formula = ADOS ~ AGE:FIQ, data = data)

Residuals:

Min 1Q Median 3Q Max

-9.9070 -2.8769 0.0484 2.1536 10.0953

Coefficients:

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

(Intercept) 1.178e+01 4.363e-01 27.007 <2e-16 \*\*\*

AGE:FIQ 6.771e-05 2.206e-04 0.307 0.759

---

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

Residual standard error: 3.798 on 367 degrees of freedom

(170 observations deleted due to missingness)

Multiple R-squared: 0.0002567, Adjusted R-squared: -0.002467

F-statistic: 0.09422 on 1 and 367 DF, p-value: 0.7591

که اینجا هم پی ولیو ما برای اینترکشن این دو متغیر معنا دار نیست.

حالا برای کشیدن پلات کوریلیشن این مدل، نیازمند ستونی جداگانه برای به تصویر کشیدن اینترکشن بین دو متغیر AGE و FIQ نیز هستیم. برای این منظور یک دیتا فریم میسازیم:

data <- data.frame(x1,x2,x3,x1\*x2,x4)

colnames(data) <- c("AGE","FIQ","ADOS","AGE\*FIQ","SEX")

ggpairs(data,

mapping = aes(color = SEX),

upper = list(continuous = wrap("cor", size = 4)),

lower = list(continuous = wrap("smooth", method = "lm", se = TRUE))

)

A screenshot of a graph

Description automatically generated

**تفسیر:**

دقت کنید که با اینکه مدل ما پیش بینی کرد که هیچ کدام از متغیر های مستقل(main effect) و نیز اینترکشن این دو با یکدیگر تاثیری روی ADOS ندارند، اما اینجا میبینیم که کوریلیشن بین ADOS و FIQ معنا دار است که این ربطی به خود مدل ندارد زیرا در مدل، همانطور که قبلا گفته شد، برهم کنش متغیر ها بر روی هم در نظر گرفته میشود. حتی اگر به دنبال main effect متغیر ها باشیم. به عبارتی وقتی متغیر ها در مدل بررسی میشوند به شکل یک پکیج کامل مورد بررسی اند.

**پرسش و پاسخ:**

برای تفسیر نهایی از کدام نتایج استفاده میکنیم؟ مدل یا کوورایانس؟  
بستگی به سوال پژوهشی مطالعه خود گزارش میکنید. مثلا وقتی اثرات تعاملی یا اثر چند متغیر با هم را بررسی میکنید باید از نتایج مدل استفاده کنیم.

* Analysis of covariance (ANCOVA):

در این انالیز هم یک متغیر پیوسته را از روی یک متغیر پیوسته دیگر پیش بینی میکنیم با این تفاوت که اینترکشن یک متغیر اسمی بر روی متغیر مستقل را نیز بررسی میکنیم.  
مثلا پیش بینی ADOS از روی AGE را در نظر بگیرید:

model <- lm(ADOS ~ AGE , data)

حالا وقتی یک متغیر اسمی داشته باشیم (مثل جنسیت) و بخواهیم تاثیر ان را بر پیش بینی ADOS از روی AGE بدانیم، (البته باید ابتدا SEX را به as.factor تبدیل کنیم تا از کد عددی به یک فاکتور که متغیر اسمی است تبدیل شود) این متغیر را به مدل خطی خود اضافه میکنیم و به یک انالیز کوواریانس دست میابیم.

این یعنی میخواهیم بدانیم آیا پیش بینی یک متغیر پیوسته (ADOS) از روی متغیر پیوسته دیگر (AGE) متاثر از جنسیت است یا نه، ایا این پیش بینی بین دو جنس فرق دارد یا نه:

model <- lm(ADOS ~ AGE \* SEX , data)

> summary(model)

.  
.

Coefficients:

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

(Intercept) 11.38173 0.48657 23.392 <2e-16 \*\*\*

AGE 0.03281 0.02599 1.263 0.208

SEX2 0.67829 1.50206 0.452 0.652

AGE:SEX2 -0.07337 0.08329 -0.881 0.379

---

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

.  
.

در این انالیز دقیقا نتیجه ترم “AGE:SEX2” بیان کننده این است که ایا این پیش بینی بین دو جنس فرق دارد یا نه. که اینجا معنا دار نیست.

حالا پلات این مدل را رسم میکنیم:

p <- ggplot(data, aes(x = AGE , y = ADOS, color = SEX)) +

geom\_point() +

geom\_smooth(method = "lm", formula = y ~ x, se = FALSE) +

labs(x = "AGE", y = "ADOS") +

theme\_minimal()

print(p)

A graph with red and blue dots

Description automatically generated

شکلی که میبینیم نیز نشان دهنده معنادار نبودن این ANCOVA است زیرا دو خط فیت شده روی نمودار باهم تقاطع دارند.   
زمانی که دو یا چند خط فیت شده روی نمودار باهم فاصله مناسبی داشته باشند و به شکلی تفکیک شده بنظر برسند میتوان گفت انالیز کوواریانس معنا دار بوده است.