Week 04: ANOVA, categorical predictors, and interactions

General Linear Model revisited

Bartosz Maćkiewicz

This week (Week 4)

In previous we focused on regression with numerical predictors. This week we extend the same model family to:

- group comparisons (categorical predictors),
- how coding/contrasts change interpretation,
- and interactions (when the effect of one predictor depends on another).

ANOVA as regression: same model, different language

One-way ANOVA:

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

Regression view:

$$Y = X\beta + \varepsilon$$

It is the same underlying model, but with different language and a different emphasis on how the predictors are represented in the design matrix.

Example: Giancola and Corman (2007)

We will begin with a study by Giancola and Corman (2007). They were interested in studying the effects of a distracting task on aggressive behavior of subjects who had consumed a significant amount of alcohol. It is well known that alcohol often leads to aggressive behavior, but why?

Giancola and Corman began by assuming that alcohol facilitated aggression by focusing attention on more salient provocative cues rather than on less salient inhibitory ones. They reasoned that if they presented their subjects with a distracting task, attention would be focused on the task rather than on provocative cues, thus limiting aggression. However, they also reasoned that if the task became too complex, its distracting effects would disappear and aggression would take over.

Example: Giancola and Corman (2007)

Giancola and Corman asked their subjects to consume alcohol in an amount that raised their average blood alcohol level to about .10%. Subjects then participated in a task that required them to remember the order in which squares in a 3 x 3 matrix were illuminated.

The attentional demands of the task were varied by manipulating the number of squares that subjects had to keep in memory. Subjects played against a fictitious opponent who either delivered mild shocks to the subject or received mild shocks from the subject, dependent on supposed task performance. The dependent variable (aggression) was based on the severity and duration of shocks that subjects delivered to opponents when they had the opportunity.

There were five groups in this study, varying in task difficulty. Subjects had to remember the pattern of either 0, 2, 4, 6, or 8 squares.

A few facts about ANOVA...

- ANOVA, or Analysis of Variance, is a statistical method used to compare multiple group means.
- It was developed by Ronald Fisher in the 1920s.
- ANOVA facilitates the examination of the impacts of one or more experimental factors simultaneously (e.g., one-way, two-way, or three-way ANOVA).
- Examples of such factors include treatment types, gender, and types of fertilizer (a nod to Fisher).
- Each factor can have several levels (e.g., types of fertilizer such as A, B, and C; or varying degrees of humidity).

Theoretical model

Here is a structural model that underlies ANOVA:

$$X_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- X_{ij} - individual observation (i observation from j group)
- μ - grand mean in the population
- τ_j - specific treatment effect of being in Group j
- ϵij - error associated with a specific individual

Assumptions: homogeneity of variance

$$\sigma_1^2 = \sigma_2^2 = ... = \sigma_3^2 = \sigma_\epsilon^2$$

The subscript "ϵ" stands for error, and this variance is the error variance – the variance unrelated to any treatment differences, which is the variability of scores within the same condition. Homogeneity of variance would be expected to occur if the effect of a treatment is to add a constant to everyone's score – if, for example, everyone who was trying to recall the position of eight stimuli scored an extra point above the others on the aggression scale.

Assumptions: normality

A second assumption of the analysis of variance is that the observations for each condition are normally distributed around their group mean.

Because σ_e^2 represents the variability of observations around the mean of that condition, a more correct way to write the assumption is to write that error is normally distributed within conditions.

Thus you will often see the assumption stated in terms of “the normal distribution of error.” or “normally distributed residuals,” which means the same thing.

Moderate departures from normality are not usually fatal. We said much the same thing when looking at the t test for two independent samples, which is really just a special case of the analysis of variance.

Assumptions: independence

The third important assumption is that the observations are independent of one another.

Thus for any two observations within an experimental treatment, we assume that knowing how one of these observations stands relative to the treatment (or population) mean tells us nothing about the other observation.

This is one of the important reasons why subjects are randomly assigned to groups. Violation of the independence assumption can have serious consequences for an analysis.

Hypotheses

Here are a null hypothesis and an alternative hypothesis for ANOVA:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n$$

$$H_A : \neg H_0$$

The null hypothesis could be false in a number of ways (e.g., all means could be different from each other, the first two could be equal to each other but different from the last three, and so on), but for now we are going to be concerned only with whether the null hypothesis is completely true or is false. This is frequently referred to as the **omnibus null hypothesis**.

The Logic of the Analysis of Variance

```r
giancola <- read.csv("giancola2007.csv")
library(tidyverse)
head(giancola)
```

```
  Group   dv
1     0 1.28
2     0 1.35
3     0 3.31
4     0 3.06
5     0 2.59
6     0 3.25
```

The Logic of the Analysis of Variance: two estimators of variance

Consider for a moment the effect of our three major assumptions - **normality**, **homogeneity of variance**, and the **independence of observations** (or residuals).

By making the first two of these assumptions we have said that the five distributions have the same shape and dispersion. As a result, the only way left for them to differ is in terms of their means.

The Logic of the Analysis of Variance: mean square error

We will begin by making no assumption concerning H_0 – it may be true or false. For any one treatment, the variance of the 12 scores in that group would be an estimate of the variance of the population from which the scores were drawn. Because we have assumed that all populations have the same variance, it is also one estimate of the common population variance. We can thus construct a following estimator of population variance, which is equivalent to calculating the variance for each level of the factor separately and taking a mean:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

$$MSE = \frac{SSE}{n-k}$$

The Logic of the Analysis of Variance: mean square error

It is important to note that this estimate does not depend on the truth or falsity of H_0, because s_j^2 is calculated on each sample separately.

```
giancola %>%
  group_by(Group) %>%
  summarize(vars = var(dv)) %>%
  summarize(MSE = mean(vars)) %>%
  pull(MSE)
```

```
[1] 2.262875
```

The Logic of the Analysis of Variance: mean square of treatments

Now let us assume that H_0 is true. If this is the case, then our five samples of $n = 12$ cases can be thought of as five independent samples from the same population (or, equivalently, from five identical populations), and we can produce another possible estimate of the population variance.

Recall that the Central Limit Theorem states that:

$$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Therefore:

$$\sigma^2 = n s_{\bar{X}}^2$$

The Logic of the Analysis of Variance: mean square of treatments

```
giancola %>%
  group_by(Group) %>%
  summarize(means = mean(dv), n = n()) %>%
  summarize(MST = var(means) * 12) %>%
  pull(MST)
```

```
[1] 15.61506
```

The Logic of the Analysis of Variance: two estimates

We now have two estimates of the population variance (σ_ϵ^2). One of these estimates ($\mathrm{MS}_{\text{error}}$) is independent of the truth or falsity of H_0. The other ($\mathrm{MS}_{\text{treatment}}$) is an estimate of σ_ϵ^2 only as long as H_0 is true (only as long as the conditions of the central limit theorem are met; namely, that the means are drawn from one population or several identical populations).

Thus, if the two estimates agree, we will have support for the truth of H_0, and if they disagree, we will have support for the falsity of H_0.

The Logic of the Analysis of Variance: degrees of freedom

With $59\,df$ overall (i.e., $N-1$), four of these are associated with differences among treatment means and the remaining 55 are associated with variability within the treatment groups.

The total degrees of freedom is always $N-1$, where N is the total number of observations. The number of degrees of freedom between treatments is always $k-1$, where k is the number of treatments. The number of degrees of freedom for error is most easily thought of as what is left over. However, it can be calculated more directly as the sum of the degrees of freedom within each treatment.

The Logic of the Analysis of Variance: F ratio and F distribution

F is obtained by dividing $\mathrm{MS}_{\text{treat}}$ by $\mathrm{MS}_{\text{error}}$.

As noted earlier, $\mathrm{MS}_{\text{error}}$ is an estimate of the population variance (σ^2_ϵ). Moreover $\mathrm{MS}_{\text{treat}}$ is an estimate of the population variance if H_0 is true, but not if it is false. If H_0 is true, then $\mathrm{MS}_{\text{error}}$ and $\mathrm{MS}_{\text{treat}}$ are both estimating the same thing, and as such they should be approximately equal. If this is the case, the ratio of one to the other will be approximately 1, give or take a certain amount for sampling error.

Thus, all we have to do is to compute the ratio and determine whether it is close enough to 1 to indicate support for the null hypothesis.

```r
f_stat <- 15.6 / 2.2629
f_stat
```

```
[1] 6.893809
```

```r
pf(f_stat, 5-1, # k - 1 df
   nrow(giancola) - 5, # n - 1 df
   lower.tail = F)
```

```
[1] 0.0001427235
```

The Logic of the Analysis of Variance: using R

We can of course do it using R:

```
fit <- aov(dv ~ factor(Group), data = giancola)
summary(fit)
```

```
              Df Sum Sq Mean Sq F value   Pr(>F)    
factor(Group)  4  62.46  15.615   6.901 0.000142 ***
Residuals     55 124.46   2.263                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-hoc tests

When comparing the means for the levels of a factor in an analysis of variance, a simple comparison using t-tests will inflate the probability of declaring a significant difference when it is not in fact present.

One way to deal with this problem is to use one of the specialized *post-hoc* test, that take it into account.

Post-hoc tests: Tukey Hones Significant Differences (HSD) test

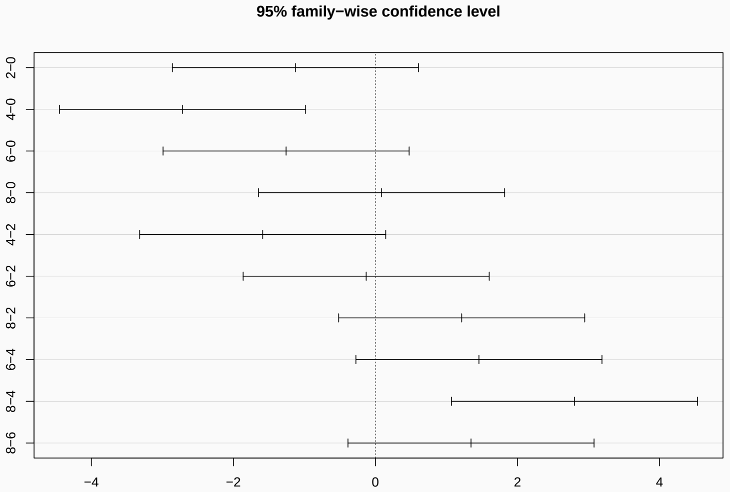
```r
TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = dv ~ factor(Group), data = giancola)

$`factor(Group)`
           diff        lwr        upr     p adj
2-0 -1.12750000 -2.8595252  0.6045252 0.3639370
4-0 -2.71500000 -4.4470252 -0.9829748 0.0004352
6-0 -1.25833333 -2.9903586  0.4736919 0.2570413
8-0  0.08666667 -1.6453586  1.8186919 0.9999078
4-2 -1.58750000 -3.3195252  0.1445252 0.0872492
6-2 -0.13083333 -1.8628586  1.6011919 0.9995272
8-2  1.21416667 -0.5178586  2.9461919 0.2906677
6-4  1.45666667 -0.2753586  3.1886919 0.1385559
8-4  2.80166667  1.0696414  4.5336919 0.0002702
8-6  1.34500000 -0.3870252  3.0770252 0.1987677
```

Post-hoc tests: a visualization

Visualizing results from the Tukey HSD test can help in understanding the differences between group means more intuitively.

```r
plot(TukeyHSD(fit))
```



Effect size

For ANOVA the most commonly used measure of effect size is η^2.

Field (2013) gives the following rules for interpreting η^2

- $\eta^2 < 0.01$ - Very small
- $0.01 \leq \eta^2 < 0.06$ - Small
- $0.16 \leq \eta^2 < 0.14$ - Medium
- $\eta^2 \geq 0.14$ - Large

```r
effectsize::eta_squared(fit)
```

```
# Effect Size for ANOVA

Parameter     | Eta2 |       95% CI
-----------------------------------
factor(Group) | 0.33 | [0.14, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

Categorical predictors: how to think about them

Until now, our focus has been primarily on numerical variables and their relationships. We will now broaden our scope to include linear models that incorporate categorical predictors. Examples of such predictors include *Gender*, *Education Level*, *Political Preferences*, *Religion*, and *Experimental Condition* (e.g., treatment group).

We will use the Salaries dataset from the carData library.

```
library(carData)
head(Salaries)
```

```
       rank discipline yrs.since.phd yrs.service  sex salary
1      Prof          B            19          18 Male 139750
2      Prof          B            20          16 Male 173200
3  AsstProf          B             4           3 Male  79750
4      Prof          B            45          39 Male 115000
5      Prof          B            40          41 Male 141500
6 AssocProf          B             6           6 Male  97000
```

Two levels of categorical predictor

Suppose we want to determine if there is a difference in earnings between men and women. We can include `sex` as a predictor in our linear regression model:

```r
fit <- lm(salary ~ sex, data = Salaries)
summary(fit)
```

```
Call:
lm(formula = salary ~ sex, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-57290 -23502  -6828  19710 116455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101002       4809  21.001  < 2e-16 ***
sexMale        14088       5065   2.782  0.00567 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30030 on 395 degrees of freedom
```

Two levels of categorical predictor

You might wonder why we do not simply perform a t-test for two independent samples. It is possible to conduct such a test, and it would yield results equivalent to those from fitting a linear model:

```
pander::pander(t.test(salary ~ sex, data = Salaries, var.equal = T))
```

Table 1: Two Sample t-test: `salary` by `sex` (continued below)

Test statistic	df	P value	Alternative hypothesis
-2.782	395	0.005667 * *	two.sided

mean in group Female	mean in group Male
101002	115090

Simple (treatment) contrasts (*dummy coding*)

By default, R employs *dummy coding* for categorical variables in regression analysis. This method involves the following steps:

- **Contrast Selection**: Select a baseline category.
- **Binary Variables**: Create a new binary variable for each level of the categorical variable, except the baseline category.

```r
Salaries$isMale <- ifelse(Salaries$sex == "Male", 1, 0)
fit <- lm(salary ~ isMale, data = Salaries)
summary(fit)
```

```

Call:
lm(formula = salary ~ isMale, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max 
-57290 -23502  -6828  19710 116455 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   101002       4809  21.001  < 2e-16 ***
isMale         14088       5065   2.782  0.00567 ** 
```

Simple (treatment) contrast (*dummy coding*)

Understanding Dummy Variables. Generally, you need $k - 1$ dummy variables to code a categorical factor with k levels. Consider the `rank` variable in the `Salaries` dataset. This variable has three levels: `Prof` (Professor), `AsstProf` (Assistant Professor), and `AssocProf` (Associate Professor).

Simple (treatment) contrast (*dummy coding*)

```
fit <- lm(salary ~ rank, data = Salaries)
broom::tidy(summary(fit)) %>%
  mutate(p.value = scales::pvalue(p.value)) %>% knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	80775.99	2887.313	27.976183	<0.001
rankAssocProf	13100.45	4130.850	3.171369	0.002
rankProf	45996.12	3230.540	14.237906	<0.001

```
contrasts(Salaries$rank) %>% knitr::kable()
```

	AssocProf	Prof
AsstProf	0	0
AssocProf	1	0
Prof	0	1

Regression equation for models with categorical variables

Formulating the Equation. The general form of the regression equation for our model is:

$$\hat{Y} = B_1C_1 + B_2C_2 + B_0$$

We have three categories: `AsstProf`, `AssocProf`, and `Prof`. To code them we created two dummy variables: C_1 and C_2.

`AsstProf` is a **baseline group** or a **reference level**. It is coded as $(C_1 = 0, C_2 = 0)$. `AssocProf` is coded as $(C_1 = 1, C_2 = 0)$ and `Prof` as $(C_1 = 0, C_2 = 1)$.

For `AsstProf` we can thus reduce the equation to:

$$\hat{Y} = B_0$$

for `AssocProf` to:

$$\hat{Y} = B_1C_1 + B_0$$

and for `Prof` to:

$$\hat{Y} = B_2C_2 + B_0$$

Regression equation for models with categorical variables

Interpretation

- The intercept (B_0) represents the mean salary for the baseline group (AsstProf).
- The coefficients B_1 and B_2 represent the salary differences relative to `AsstProf` for `AssocProf` and `Prof`, respectively.

Problems

- How to choose the right reference group?
- When it is useful to use such a coding?

Deviation (sum to zero) contrasts

Deviation contrasts, also known as *sum to zero* contrasts, offer an alternative method for coding categorical predictors. This approach modifies the interpretation of both the regression coefficients and the intercept.

```r
contr.sum(levels(Salaries$rank))
```

```
          [,1] [,2]
AsstProf     1    0
AssocProf    0    1
Prof        -1   -1
```

Deviation (sum to zero) contrasts

```r
fit <- lm(salary ~ rank,
          data = Salaries,
          contrasts = list(
            rank = standardize::named_contr_sum(Salaries$rank)))
summary(fit)
```

```

Call:
lm(formula = salary ~ rank, data = Salaries, contrasts = list(rank = standardize::n

Residuals:
   Min     1Q Median     3Q    Max
-68972 -16376  -1580  11755 104773

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     100475       1459  68.855  < 2e-16 ***
rankAssocProf   -19699       2215  -8.892  < 2e-16 ***
rankAsstProf     -6598       2245  -2.940  0.00348 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Deviation (sum to zero) contrasts

Interpretation

- `Intercept` (B_0) is a *grand mean* (mean of means)

```
mean(tapply(Salaries$salary, Salaries$rank, mean))
```

```
## [1] 100474.8
```

- Regression coefficients (B_n) represent deviations from the grand mean

```
tapply(Salaries$salary, Salaries$rank, mean) -
  mean(tapply(Salaries$salary, Salaries$rank, mean))
```

```
##   AsstProf  AssocProf       Prof
## -19698.859  -6598.406  26297.265
```

- Null hypotheses for regression coefficients state that the deviations are equal to 0

Other contrasts

Many other contrasts ara available in `codingMatrices` and `MASS` librares.

Helmert contrasts

```r
contr.helmert(levels(Salaries$rank))
fit <- lm(salary ~ rank, data = Salaries, contrasts =
            list(rank = contr.helmert))
summary(fit)
```

Successive difference contrasts

```r
contr.sum(levels(Salaries$rank))
fit <- lm(salary ~ rank, data = Salaries, contrasts =
            list(rank = MASS::contr.sdif))
summary(fit)
```

The concept of interaction

By interactions, we mean an interplay among predictors that produces an effect on the outcome Y that is different from the sum of the effects of the individual predictors.

No interaction

Consider as an example how ability (X) and motivation (Z) impact achievement in graduate school (F). One possibility is that their effects are additive. The combined impact of ability and motivation on achievement equals the sum of their separate effects; there is no interaction between X and Z. We might say that the whole equals the sum of the parts.

The concept of interaction

Synergy

A second alternative is that ability and motivation may interact *synergistically*, such that graduate students with both high ability and high motivation achieve much more in graduate school than would be expected from the simple sum of the separate effects of ability and motivation. Graduate students with both high ability and high motivation become "superstars"; we would say that the whole is greater than the sum of the parts.

The concept of interaction

Compensation

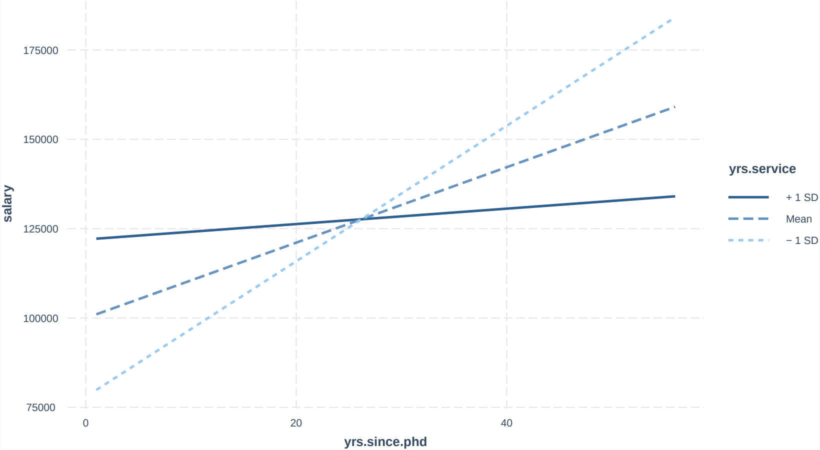
A third alternative is that ability and motivation compensate for one another. For those students who are extremely high in ability, motivation is less important to achievement, whereas for students highest in motivation, sheer native ability has less impact. Here we would say that the whole is less than the sum of the parts; there is some partial trade-off between ability and motivation in the prediction of achievement.

Interactions between two continuous predictors

```r
# Model with no interactions
fit <- lm(salary ~ yrs.service + yrs.since.phd, data = Salaries)

# Model with an interaction
fit_i <- lm(salary ~ yrs.service * yrs.since.phd, data = Salaries)

#summary(fit) # run this code
#summary(fit_i) # run this code
```

Interactions between two continuous predictors

```r
interactions::interact_plot(fit_i,
                            pred = "yrs.since.phd",
                            modx = "yrs.service"
                            )
```

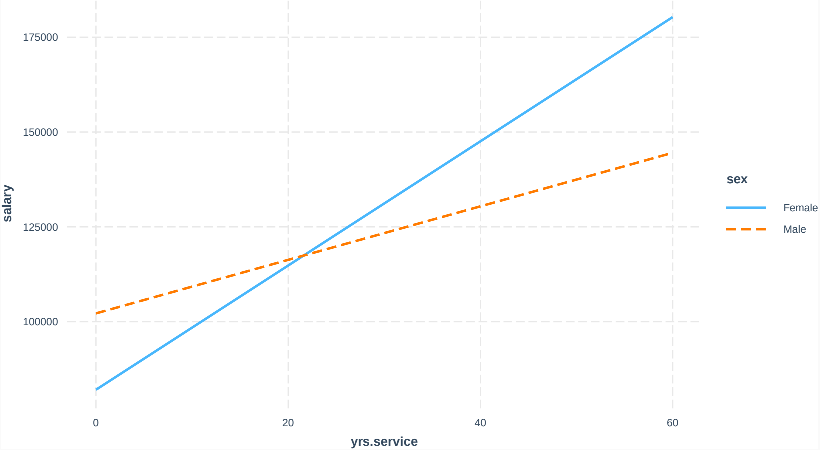


Interactions between categorical and continuous predictors

```
# Model with no interactions
fit <- lm(salary ~ yrs.service + sex, data = Salaries)

# Model with an interaction
fit_i <- lm(salary ~ yrs.service * sex, data = Salaries)

#summary(fit) # run this code
#summary(fit_i) # run this code
```

Interactions between categorical and continuous predictors

```r
interactions::interact_plot(fit_i,
                            pred = "yrs.service",
                            modx = "sex"
                            )
```

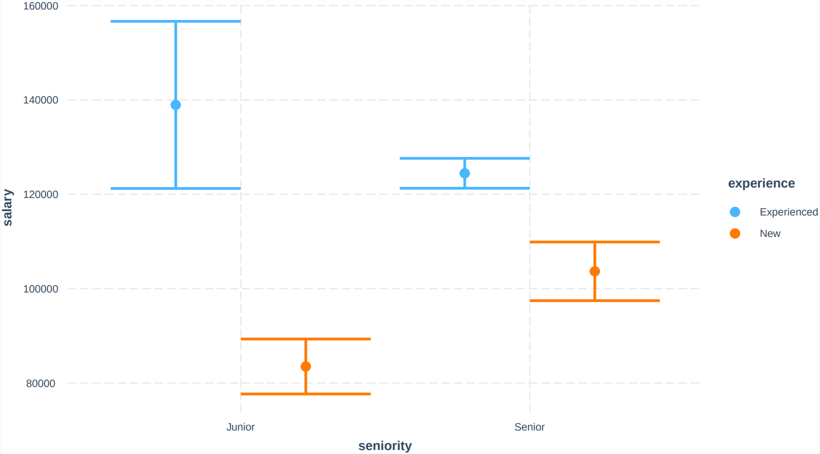


Interactions between two categorical predictors

```
# Fake variables...
Salaries$seniority <- ifelse(Salaries$yrs.service > 5,
                             "Senior",
                             "Junior")
Salaries$experience <- ifelse(Salaries$yrs.since.phd > 15,
                              "Experienced",
                              "New")
# Model with no interactions
fit <- lm(salary ~ seniority + experience, data = Salaries)

# Model with interactions
fit_i <- lm(salary ~ seniority * experience, data = Salaries)

#summary(fit) # run this code
#summary(fit_i) # run this code
```

Interactions between two categorical predictors

```
interactions::cat_plot(fit_i,
                       pred = "seniority",
                       modx = "experience"
                       )
```



Interactions in multiple linear regression: example

Wagner, Compas, and Howell (1988) hypothesized that individuals who experience more stress, as assessed by a measure of daily hassles, will exhibit higher levels of symptoms than those who experience little stress. That is what, in analysis of variance terms, would be the main effect of hassles. However, they also expected that if a person had a high level of social support to help deal with his or her stress, symptoms would increase only slowly with increases in hassles. For those who had relatively little social support, symptoms were expected to rise more quickly as hassles increased.

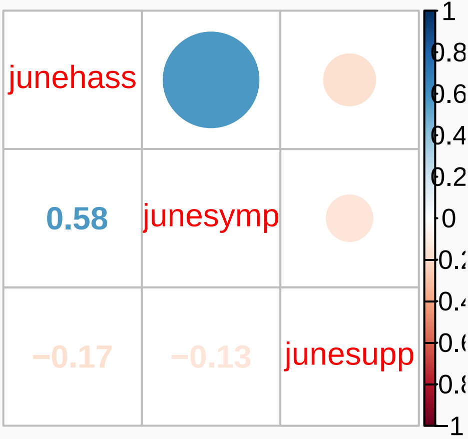
```
library(dplyr); hassles <- read.csv("hassles.csv")
hassles %>%
  select(id, junehass, junesymp, junesupp) %>%
  slice_head(n = 5)
```

```
  id junehass junesymp junesupp
1  5      176       73       10
2 26      379       88       50
3 58      126      118       45
4 41      193       79       40
5 63      229      127       40
```

Interactions in multiple linear regression: example

First we can look at the correlation matrix between all three variables.

```
corrplot::corrplot.mixed(
  cor(hassles %>% select(junehass, junesymp, junesupp)))
```



Interactions in multiple linear regression: example

If we just multiply Hassles and Support together, there will be two problems with what results.

In the first place, either Hassles or Support or both will be highly correlated with their product, which will make for multicollinearity in the data.

Second, any effect of Hassles or Support in the regression analysis will be evaluated at a value of 0 for the other variable. In other words the test on Hassles will be a test on whether Hassles is related to Symptoms if a participant had exactly no social support. Similarly the test on Support would be evaluated for those participants who have exactly no hassles.

To circumvent these two problems we are going to **center** our data

```
hassles$c_hassles <- hassles$junehass - mean(hassles$junehass)
hassles$c_supp <- hassles$junesupp - mean(hassles$junesupp)
```

Interactions in multiple linear regression: example

We can compare two models – with interaction term and without it.

```
fit_no_interact <- lm(junesymp ~ c_hassles + c_supp,
                      data = hassles)
fit_interact <- lm(junesymp ~ c_hassles * c_supp,
                   data = hassles)
anova(fit_no_interact, fit_interact)
```

```
Analysis of Variance Table

Model 1: junesymp ~ c_hassles + c_supp
Model 2: junesymp ~ c_hassles * c_supp
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     53 16151
2     52 14840  1    1311.3 4.5948 0.03677 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions in multiple linear regression: example

```r
summary(fit_interact)
```

```

Call:
lm(formula = junesymp ~ c_hassles * c_supp, data = hassles)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.734 -13.379  -0.444   8.689  35.523 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      89.584940   2.291503  39.094  < 2e-16 ***
c_hassles         0.085942   0.019213   4.473 4.22e-05 ***
c_supp            0.146358   0.305244   0.479   0.6336    
c_hassles:c_supp -0.005065   0.002363  -2.144   0.0368 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.89 on 52 degrees of freedom
Multiple R-squared:  0.3885,	Adjusted R-squared:  0.3532 
F-statistic: 11.01 on 3 and 52 DF,  p-value: 1.046e-05
```

Interactions in multiple linear regression: example

```
interactions::interact_plot(fit_interact,
              c_hassles, # predictor variable
              c_supp) # moderator variable
```