Dummy-Variable Regression

for categorical predictors

👨‍🏫 Statistical Reasoning Lecture #3
Alexander Savi, 2025

📚 Mehmetoglu & Mittner Ch. (6), 9, 10.3, .5, .6

Split Grid by Nicola Rennie (aRt package) 🎨

Ms. Signal 🧑‍🤝‍🧑 Mr. Noise

📷 Jorn van Eck (UvA)

📢 Announcements

🔁 Recap

Last week

Multiple regression

- ❏ Partial slope coefficient (“having controlled for”)
- ❏ Adjusted R^2
- ❏ Standardization (for comparing slopes)

Moderation / interaction analysis

- ❏ Simple slopes analysis (slope of X_1 depends on X_2)
- ❏ Centering (for meaningful slope coefficients)

Last year

Four [levels of measurement]

- ❏ Nominal scale: 🍏🍓🍌
- ❏ Ordinal scale: ☹️😐😀
- ❏ Interval scale: 📅🌡️ (°C)
- ❏ Ratio scale: 🕔📏⚖️🌡️ (K)

Two [types of variables]

- ❏ Categorical
  - ❏ Dichotomous: ♂♀
  - ❏ Polytomous: 🍏🍓🍌🍑
- ❏ Numerical
  - ❏ Discrete: number of …
  - ❏ Continuous: 🌡️

🔮 Today

Topics

1 | Statistical reasoning with GLM
2 | Multiple linear regression
3 | Dummy-variable regression
    3.1 | Dummy-variable regression
    3.2 | Moderation/interaction analysis
4 | Logistic regression
5 | Multilevel and longitudinal analysis
6 | Statistics superpowers
7 | Bayesian statistics

Learning goals

Estimate the relationships between more than two categorical variables.

Determine whether the relationship between a categorical and a continuous variable depends on a third categorical variable.

Dummy-variable regression

with one categorical predictor

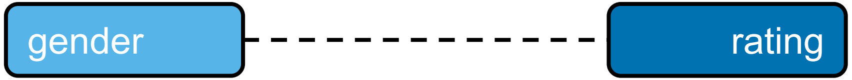
Split Grid by Nicola Rennie (aRt package) 🎨

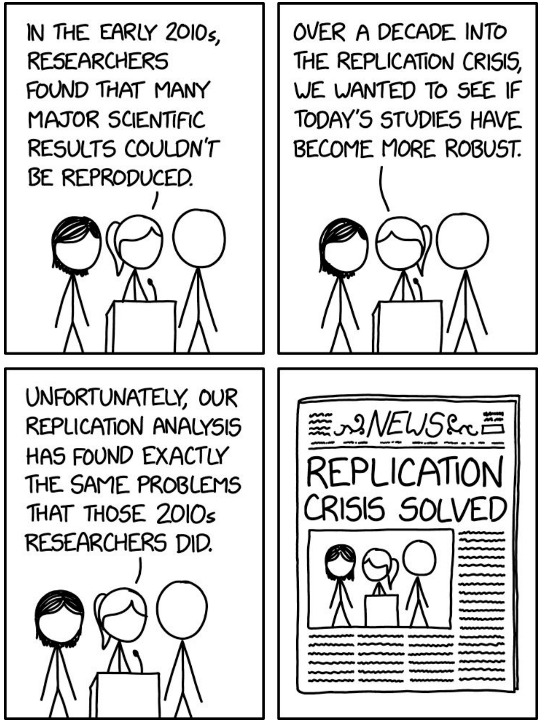
Student evaluations | Gender

De student als consument maakt vrouwelijke docenten extra kwetsbaar

Nieuws | door Frans van Heest

13 september 2023 | Vrouwelijke docenten worden aantoonbaar gediscrimineerd door studentenevaluaties, maar toch blijft het instrument voor veel universiteiten belangrijk om medewerkers te beoordelen. Cursusevaluaties moedigen echter middelmatig onderwijs aan en zijn extra nadelig voor vrouwen.





🔒ScienceGuide (Sep. 13, 2023)

🎨Randall Munroe (wtf)

Categorical independent variables

I gave each category a number (male = 0, female = 1), and look, no errors!

```
Call:
lm(formula = mod, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.83433 -0.36357  0.06567  0.40718  0.90718 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.23433    0.03298 128.384  < 2e-16 ***
gender      -0.14151    0.05082  -2.784  0.00558 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
mod <- score ~ gender
dat <- dat |> mutate(gender = ifelse(gender ==
"male", 0, 1))
summary(lm(mod, data = evals))
```

I guess you’re right, but will it always?

```
	Two Sample t-test

data:  score by gender
t = -2.7844, df = 461, p-value = 0.005583
alternative hypothesis: true difference in means between group female and group male is not equal to 0
95 percent confidence interval:
 -0.2413779 -0.0416378
sample estimates:
mean in group female   mean in group male 
            4.092821             4.234328 
```

```
t.test(mod, data = evals, var.equal = TRUE)
```

🍼 Dummies for dummies | Dummy-coding

Dichotomous

♂= 0
♀= 1

Why does it work?

Polytomous

🍏= 0
🍓= 1
🍌= 2
🍑= 3

Why does it not work?

Dummy-coding

Original data

Y	X
5	🍏
6.2	🍓
2	🍌
4.7	🍑

Dummy-coded data

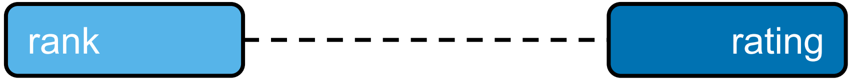
Y	🍏	🍓	🍌	🍑
5	1	0	0	0
6.2	0	1	0	0
2	0	0	1	0
4.7	0	0	0	1

🪄 Factor

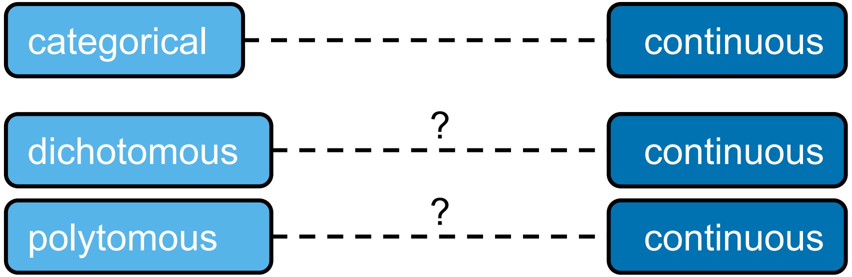
```
class(data$categorical)
data$categorical <- factor(data$categorical)
```

Student evaluations | Academic rank

🧑‍🏫 Teacher
🧑‍🎓 Tenure track
🧑‍🔬 Tenured



Model thinking



$$\text{score} = \beta_0 + \beta_1(\text{rank}_{\text{tenure track}}) + \beta_2(\text{rank}_{\text{tenured}}) + \epsilon$$

```
mod <- score ~ rank
```

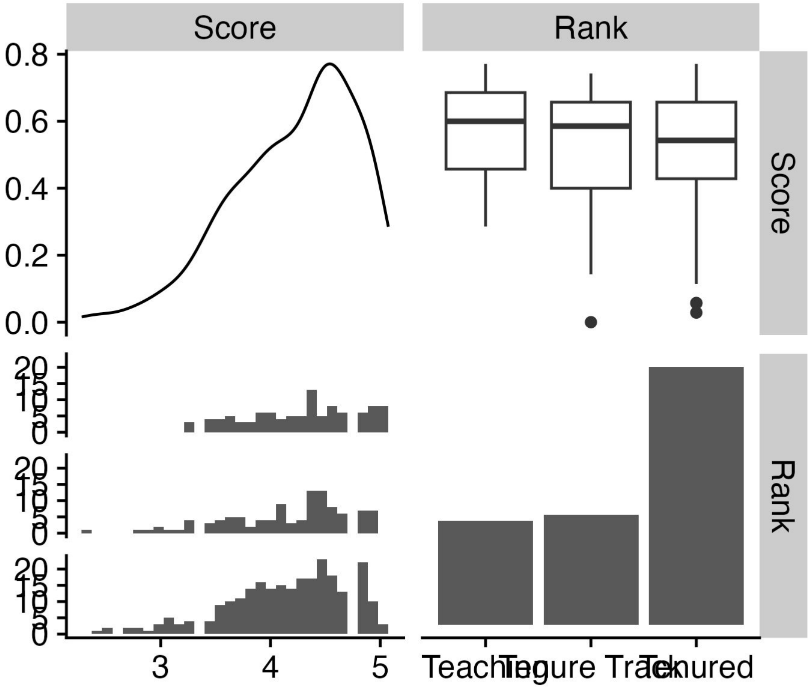
Independent

Dependent	Categorical	Continuous
Categorical		
Continuous	Dummy-variable regression	Simple regression Multiple regression

📖 9.1, 9.2, 9.4

Data | Transformation & visualization

```
data(evals)
class(evals$rank)
levels(evals$rank)
summary(evals$rank)
GGally::ggpairs(evals, columns = c("score",
"rank"))
```



📖 5.2, 5.3 (density, boxplot, histogram, bar chart)

Results | With & without intercept

```
mod <- score ~ rank
fit <- lm(mod, data = evals)
summary(fit)
```

```
Call:
lm(formula = mod, data = evals)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8546 -0.3391  0.1157  0.4305  0.8609 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.28431    0.05365  79.853   <2e-16 ***
ranktenure track -0.12968    0.07482  -1.733   0.0837 .  
ranktenured      -0.14518    0.06355  -2.284   0.0228 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5419 on 460 degrees of freedom
Multiple R-squared:  0.01163,Adjusted R-squared:  0.007332 
F-statistic: 2.706 on 2 and 460 DF,  p-value: 0.06786
```

$$\widehat{\text{score}} = 4.28 - 0.13(\text{rank}_{\text{tenure track}}) - 0.15(\text{rank}_{\text{tenured}})$$

```
mod <- score ~ 0 + rank
fit <- lm(mod, data = evals)
summary(fit)
```

```
Call:
lm(formula = mod, data = evals)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8546 -0.3391  0.1157  0.4305  0.8609 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
rankteaching      4.28431    0.05365   79.85   <2e-16 ***
ranktenure track  4.15463    0.05214   79.68   <2e-16 ***
ranktenured       4.13913    0.03407  121.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5419 on 460 degrees of freedom
Multiple R-squared:  0.9835, Adjusted R-squared:  0.9834 
F-statistic:  9163 on 3 and 460 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{score}} = 4.28(\text{rank}_{\text{teaching}}) + 4.15(\text{rank}_{\text{tenure track}}) + 4.14(\text{rank}_{\text{tenured}})$$

🍼 For dummies dummies | Changing the reference group

```
evals$rank <- relevel(evals$rank, ref = "tenure track")
levels(evals$rank)
```

```
Call:
lm(formula = mod, data = evals)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8546 -0.3391  0.1157  0.4305  0.8609 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.15463    0.05214  79.680   <2e-16 ***
rankteaching  0.12968    0.07482   1.733   0.0837 .  
ranktenured  -0.01550    0.06228  -0.249   0.8036    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5419 on 460 degrees of freedom
Multiple R-squared:  0.01163,Adjusted R-squared:  0.007332 
F-statistic: 2.706 on 2 and 460 DF,  p-value: 0.06786
```

🪄 Linear combinations (see next slide)

- ❏ $\beta_{\text{Teaching}} - \beta_{\text{Tenure Track}}$
- ❏ $\beta_{\text{Tenured}} - \beta_{\text{Tenure Track}}$
- ❏ $\beta_{\text{Tenured}} - \beta_{\text{Teaching}}$

⚠️ Family-wise error rate = $1 - (1 - \alpha)^m$

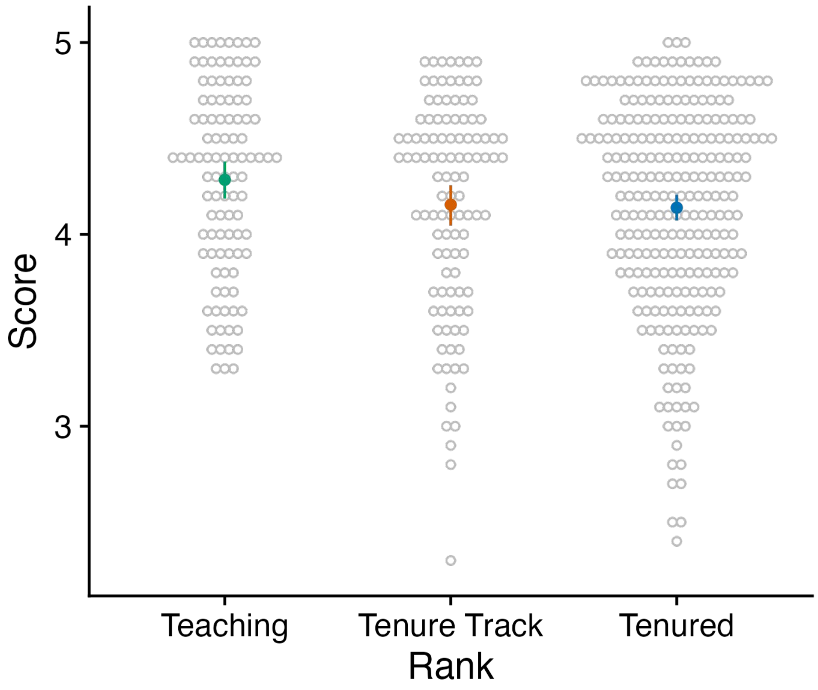
- ❏ m = 3
- ❏ α = .05
- ❏ FWER ≈ .14

Bonferroni adjustment = α / m ≈ .02, but power!

Results | Pairwise multiple comparison adjustment

```
tukey <- glht(fit, linfct = mcp(rank = "Tukey"),
vcov = sandwich)
summary(tukey)
```

```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = mod, data = evals)

Linear Hypotheses:
                             Estimate Std. Error t value Pr(>|t|)
teaching - tenure track == 0  0.12968    0.07279   1.782   0.1753
tenured - tenure track == 0  -0.01550    0.06388  -0.243   0.9678
tenured - teaching == 0      -0.14518    0.06002  -2.419   0.0417 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```



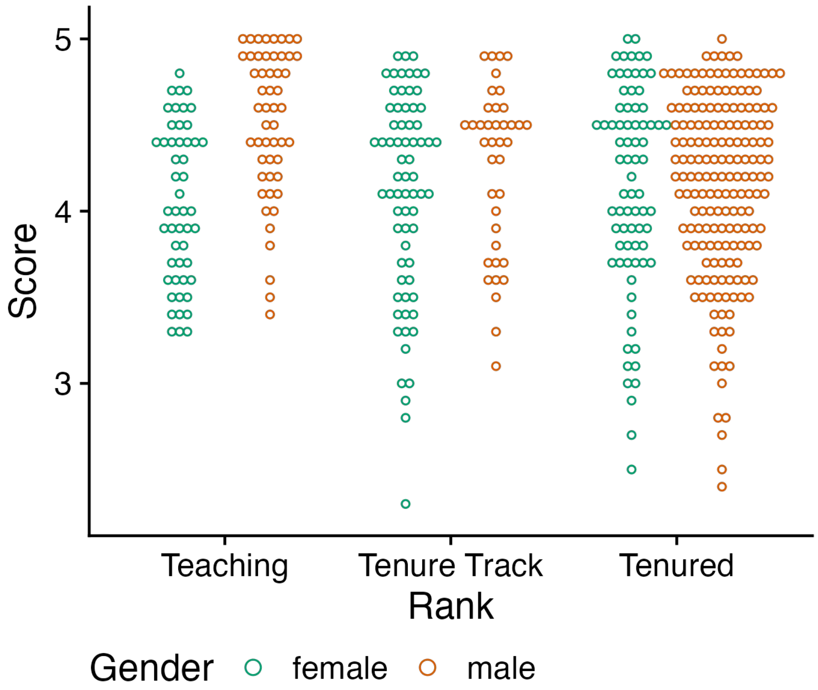
Dummy-Variable Regression

with more than one categorical predictor

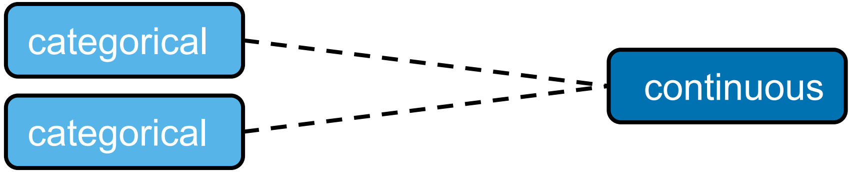
Split Grid by Nicola Rennie (aRt package) 🎨

Student evaluations | Gender + rank

- Male > Female
- Teaching > Tenured



Model thinking



What's the meaning of *score*, when…

- ❑ $\beta_1 = \beta_2 = \beta_3 = 0$
- ❑ $\beta_1 = 1$
- ❑ $\beta_1 = \beta_2 = 1$
- ❑ $\beta_3 = 1$

$$\text{score} = \beta_0 + \beta_1(\text{gender}_{\text{male}}) + \beta_2(\text{rank}_{\text{tenure track}}) + \beta_3(\text{rank}_{\text{tenured}}) + \epsilon$$

```
mod <- score ~ gender + rank
```

📖 9.6

Results

```
fit <- lm(mod, data = evals)
summary(fit)
```

```
Call:
lm(formula = mod, data = evals)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7941 -0.3418  0.1011  0.4105  0.9781 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4.19887    0.05954  70.520  < 2e-16 ***
gendermale         0.16760    0.05272   3.179  0.00158 ** 
ranktenure track -0.10476    0.07450  -1.406  0.16033    
ranktenured      -0.17699    0.06373  -2.777  0.00570 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5366 on 459 degrees of freedom
Multiple R-squared:  0.03292,	Adjusted R-squared:  0.0266 
F-statistic: 5.208 on 3 and 459 DF,  p-value: 0.001519
```

$$\widehat{\text{score}} = 4.2 + 0.17(\text{gender}_{\text{male}}) - 0.1(\text{rank}_{\text{tenure track}}) - 0.18(\text{rank}_{\text{tenured}})$$

- ❏ What does the *intercept* estimate represent?
- ❏ What does the *ranktenured* estimate represent?

The mean score of females in a tenure track is .10 points lower than the mean score of female teachers, having controlled for males and tenured females.

- ❏ Is there an effect *rank*, having controlled for *gender*?

Results | *F*-test

```
car::linearHypothesis(
  model = fit,
  hypothesis.matrix = c("ranktenure track = 0",
                        "ranktenured = 0"))
```

```
Linear hypothesis test:
ranktenure track = 0
ranktenured = 0

Model 1: restricted model
Model 2: score ~ gender + rank

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    461 134.39
2    459 132.16  2    2.2382 3.8869 0.02119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Rank has a significant relation with *score*, having controlled for *gender* (α = .5).

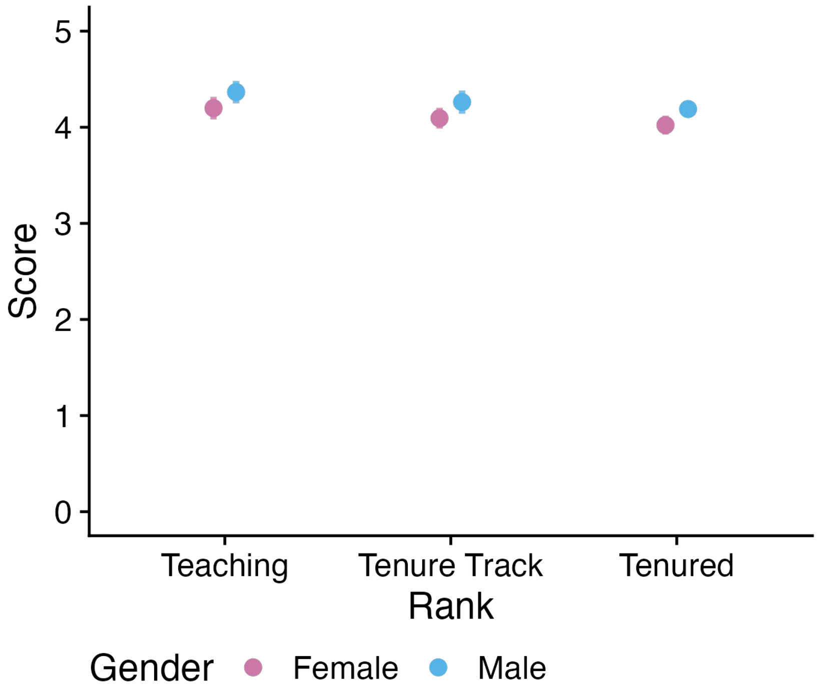
```
car::linearHypothesis(
  model = fit,
  hypothesis.matrix = c("gendermale = 0"))
```

```
Linear hypothesis test:
gendermale = 0

Model 1: restricted model
Model 2: score ~ gender + rank

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    460 135.06
2    459 132.16  1    2.9093 10.104 0.001579 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Gender has a significant relation with *score*, having controlled for *rank* (α = .5).

Results | Visualization

```
?interactions::cat_plot
cat_plot(model = fit,
         pred = "rank",
         modx = "gender",
         geom = "point",
         interval = TRUE) +
  ylim(0, 5)
```



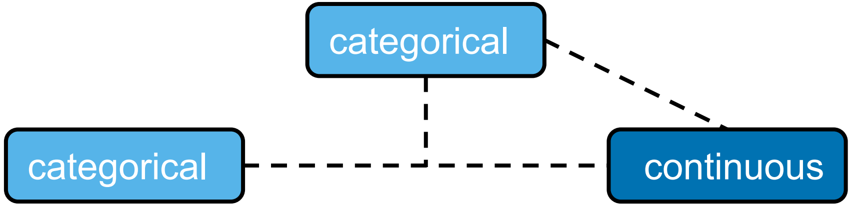
- ❏ Relationship between score and gender is restricted to be the same across ranks (and vice versa). I.e., we used an additive model.

Dummy-variable regression

One categorical predictor and a categorical moderator.

Split Grid by Nicola Rennie (aRt package) 🎨

Student evaluations | Gender × rank



$$\text{score} = \beta_0 + \beta_1(\text{gender}_{\text{male}}) + \beta_2(\text{rank}_{\text{tenure track}}) + \beta_3(\text{rank}_{\text{tenured}}) + \beta_4(\text{gender}_{\text{male}} \times \text{rank}_{\text{tenure track}}) + \beta_5(\text{gender}_{\text{male}} \times \text{rank}_{\text{tenured}}) + \epsilon$$

```
mod <- score ~ gender * rank
```

📖 10.5

Results

```
fit <- lm(mod, data = evals)
summary(fit)
```

```
Call:
lm(formula = mod, data = evals)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.79710 -0.34520  0.07885  0.37885  0.87500 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  4.03800    0.07498  53.855  < 2e-16 ***
gendermale                   0.48315    0.10501   4.601 5.46e-06 ***
ranktenure track             0.05910    0.09847   0.600 0.548660    
ranktenured                  0.08700    0.09654   0.901 0.367982    
gendermale:ranktenure track -0.32385    0.14936  -2.168 0.030660 *  
gendermale:ranktenured      -0.46296    0.12773  -3.625 0.000322 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5302 on 457 degrees of freedom
Multiple R-squared:  0.05996,	Adjusted R-squared:  0.04967 
F-statistic:  5.83 on 5 and 457 DF,  p-value: 3.11e-05
```

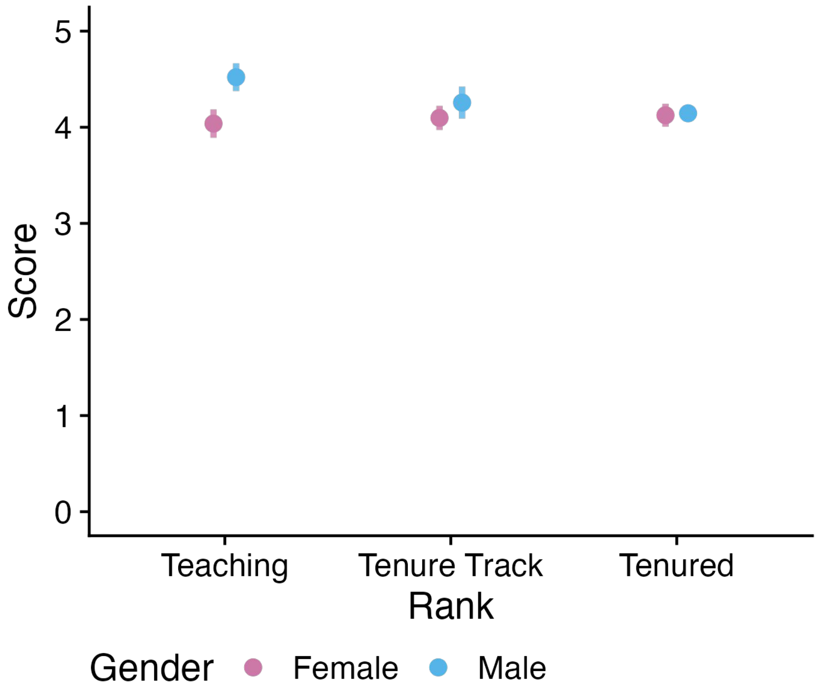
$$\widehat{\text{score}} = 4.04 + 0.48(\text{gender}_{\text{male}}) + 0.06(\text{rank}_{\text{tenure track}}) + 0.09(\text{rank}_{\text{tenured}})$$
$$0.32(\text{gender}_{\text{male}} \times \text{rank}_{\text{tenure track}}) - 0.46(\text{gender}_{\text{male}} \times \text{rank}_{\text{tenured}})$$

```
sim_slopes(model = fit, pred = "gender",
           modx = "rank", johnson_neyman = FALSE)
```

```
SIMPLE SLOPES ANALYSIS

Slope of gender when rank = teaching:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.48   0.11     4.60   0.00

Slope of gender when rank = tenure track:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.16   0.11     1.50   0.13

Slope of gender when rank = tenured:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.02   0.07     0.28   0.78
```

Results | Visualization

```
?interactions::cat_plot
cat_plot(model = fit,
         pred = "rank",
         modx = "gender",
         geom = "point",
         interval = TRUE) +
  ylim(0, 5)
```



- ❑ Relationship between score and gender may vary across ranks (and vice versa). I.e., we used a nonadditive model.
- ❑ Slope is no longer significant for each rank.

💩 R is the shit

```
library("equatiomatic")
extract_eq(fit, intercept = "beta", wrap = TRUE,
use_coefs = TRUE)
```

```
$$
\begin{aligned}
\operatorname{\widehat{score}} &= 4.2 + 0.17(\operatorname{gender}_{\operatorname{male}}) - 0.1(\operatorname{rank}_{\operatorname{tenure\ track}}) - 0.18(\operatorname{rank}_{\operatorname{tenured}})
\end{aligned}
$$
```

Copy-paste and download as image:

$$
\begin{aligned}
\widehat{\text{score}} &= 4.04 + 0.48(\text{gender}_{\text{male}}) + 0.06(\text{rank}_{\text{tenure track}}) + 0.09(\text{rank}_{\text{tenured}}) \\
&\quad 0.32(\text{gender}_{\text{male}} \times \text{rank}_{\text{tenure track}}) - 0.46(\text{gender}_{\text{male}} \times \text{rank}_{\text{tenured}})
\end{aligned}
$$

📦 equatiomatic package

Cooling Down

Split Grid by Nicola Rennie (aRt package) 🎨

💯 Exam*

(*course manual is leading)

Statistical Reasoning

📖 Chapter 7, 8, 9, 10, 11, 12, 15 (pdf available)

👨‍🏫 Lectures (pdf handouts available)

🧑‍💻 Weekly assignments (not available)

✍️ Colophon

Slides

alexandersavi.nl/teaching/

License

⚠️ Don’t look here!

Show that an ANOVA and linear regression analysis return the same results.

Share your attempt (and tell whether you needed hints)!

Hints (select and copy/paste the invisible text below to reveal it)

0.

1.

2.

3.