Case Study:Correlation between race ethnicity and novelty

Danyili Hong

This is because the file couldn't render

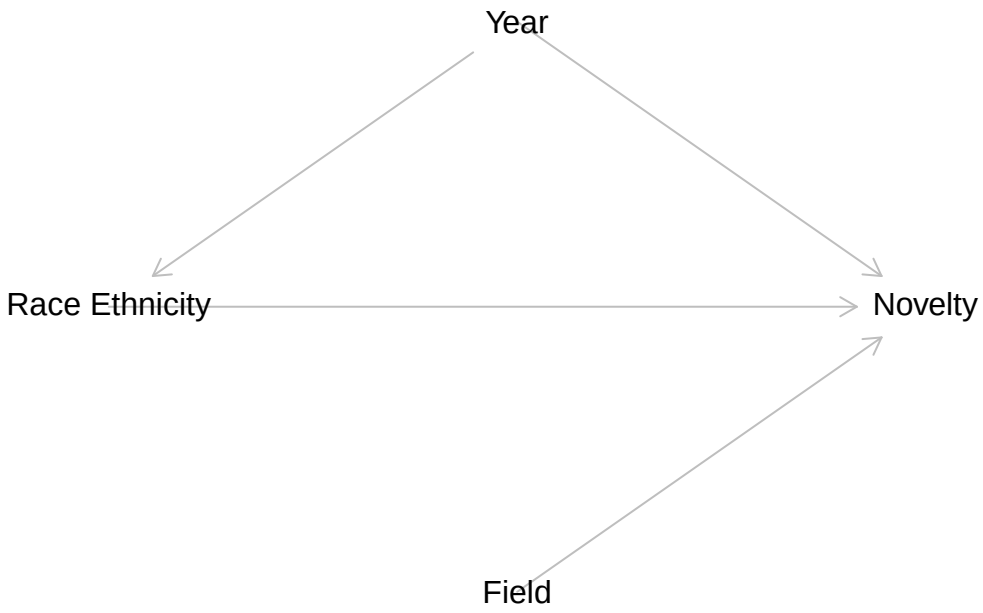
```
# Set the CRAN mirror
options(repos = "https://cran.rstudio.com/")


# Replace 'CRAN_mirror_URL' with the URL of the CRAN mirror you want to use.
# For example:
# options(repos = "https://cran.rstudio.com/")
```

Planning a model:

How would race ethnicity and field of people from different years impact the level of novelty? I will investigate whether race ethnicity, field, and year are associated with novelty, while accounting for effects of moderator sex.

```
# Create a DAG
dag <- dagitty('
  dag {
    "Race Ethnicity" [pos="1,2"]
    "Year" [pos="2,1"]
    "Field" [pos="2,3"]
    "Novelty" [pos="3,2"]

    "Race Ethnicity" -> "Novelty" [pos="1,2"]
    "Year" -> "Novelty" [pos="2,1"]
    "Field" -> "Novelty" [pos="2,3"]
```

```
    "Year" -> "Race Ethnicity"
  }
')

# Plot the DAG
plot(dag)
```



Causal Diagram: My main predictor is race ethnicity with my response variable novelty. I have year as my confounder, which influence race ethnicity and novelty level. I have field as my mediator influencing novelty, at the same time it could be precision covariate.

And since there are 4195, it fulfills n/15 rule.

```
innovation <- read_csv('https://sldr.netlify.app/data/phd_innovation.csv', show_col_types
nrow(innovation)
```

```
[1] 4195
```

```
nmodel <- lm(novelty ~ race_ethnicity + field + year, data = innovation)
summary(nmodel)
```

```
Call:
lm(formula = novelty ~ race_ethnicity + field + year, data = innovation)

Residuals:
   Min     1Q Median     3Q    Max
-5.656 -3.522 -1.545  1.876 33.635

Coefficients:
                                           Estimate Std. Error t value
(Intercept)                               -30.53964   27.54555  -1.109
race_ethnicityAsian                         0.70139    1.40565   0.499
race_ethnicityBlack or African American     0.56545    1.41834   0.399
race_ethnicityEthnicity not reported        1.03825    1.48281   0.700
race_ethnicityHispanic or Latino            0.61522    1.42053   0.433
race_ethnicityMore than one race            0.99800    1.50053   0.665
race_ethnicityOther race or race not reported  0.50596    1.54801   0.327
race_ethnicityWhite                         0.18061    1.38506   0.130
fieldEngineering                           -0.08386    0.31417  -0.267
fieldHumanities and arts                    0.09449    0.30943   0.305
fieldLife sciences                         -0.03971    0.31287  -0.127
fieldMathematics and computer sciences      0.56017    0.31306   1.789
fieldOther                                  0.44429    0.31044   1.431
fieldPhysical and earth sciences           -0.51178    0.31651  -1.617
fieldPsychology and social sciences         0.45058    0.31835   1.415
year                                        0.01920    0.01371   1.400
                                          Pr(>|t|)
(Intercept)                                 0.2676
race_ethnicityAsian                         0.6178
race_ethnicityBlack or African American     0.6902
race_ethnicityEthnicity not reported        0.4838
race_ethnicityHispanic or Latino            0.6650
race_ethnicityMore than one race            0.5060
race_ethnicityOther race or race not reported  0.7438
race_ethnicityWhite                         0.8963
fieldEngineering                            0.7895
fieldHumanities and arts                    0.7601
fieldLife sciences                          0.8990
fieldMathematics and computer sciences      0.0736 .
fieldOther                                  0.1525
fieldPhysical and earth sciences            0.1060
fieldPsychology and social sciences         0.1570
year                                        0.1615
```

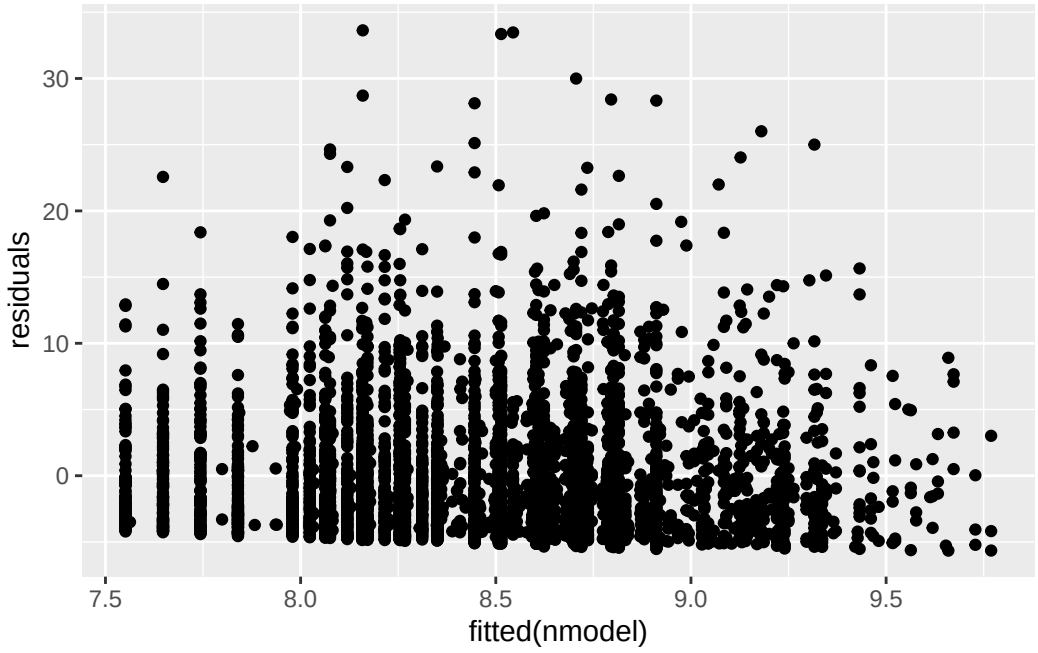
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.978 on 4179 degrees of freedom
Multiple R-squared:  0.007298,  Adjusted R-squared:  0.003735
F-statistic: 2.048 on 15 and 4179 DF,  p-value: 0.009758
```

Fit your Model

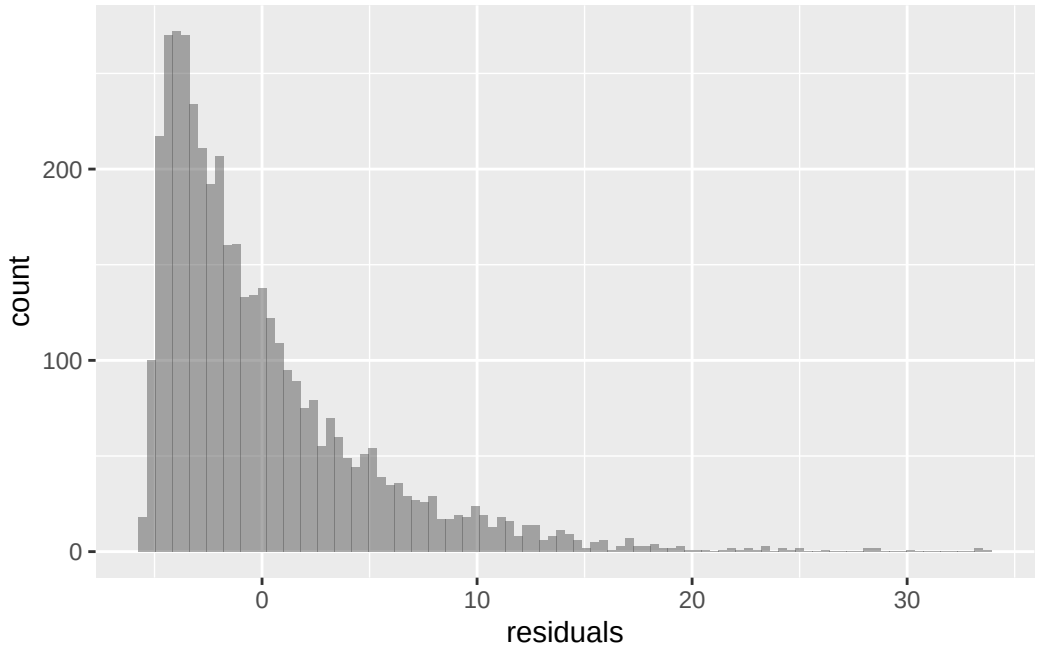
$$y = -30.54 + 0.7Race_{Asian} + 0.57Race_{AfricanAmerican} + 1.04Race_{NotReported}$$
$$+0.62Race_{Latino} + 1Race_{Morethanone} + 0.51Race_{Other} + 0.18Race_{White}$$
$$-0.08Field_{Engineering} + 0.09Field_{HumanityandArts} - 0.04Field_{LifeScience}$$
$$+0.56Field_{MathandCS} + 0.44Field_{Other} - 0.51Field_{PhysicalandEarthScience}$$
$$+0.45Field_{PhychologyandSocialScience} + 0.02year + \epsilon,$$
$$\epsilon \sim N(0, 4.978)$$

#Model Assessment

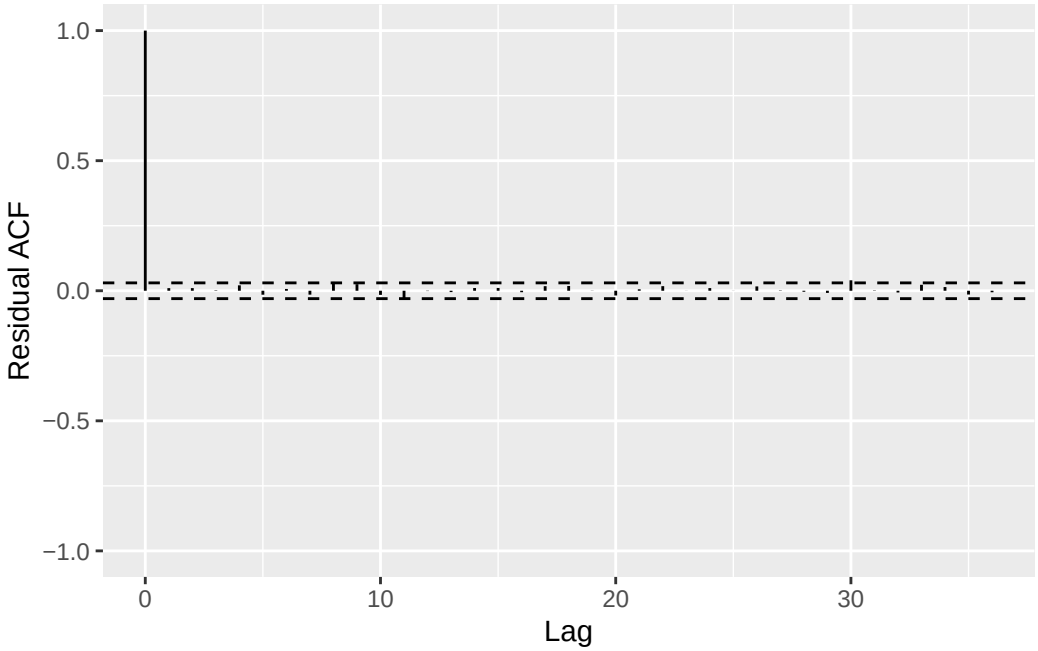
```
innovation$residuals <- residuals(nmodel)
gf_point(residuals ~ fitted(nmodel), data = innovation)
```



```
gf_histogram(~residuals,data = innovation, bins=100)
```




```
s245::gf_acf(~nmodel) |> gf_lims(y = c(-1,1))
```



Residuals and fitted model: This plot shows a random scatter of points around the horizontal line at zero, which indicates the linearity condition is met.

Histogram of residuals: This plot appears to be roughly bell-shaped and symmetric. This plot is right-skewed, and some outliers are visible. This suggests that the residual normality condition is not met.

ACF plot of residuals: This plot indicates that autocorrelation values for all lags are within the confidence bands, only one of them spikes out of the bound. This suggests that the independence of residuals is maintained.

Since not all of the conditions are met, I conclude that the model is not appropriate for drawing valid conclusions.

Prediction Plot

```
fake_data_categorical <- expand.grid(
  race_ethnicity = c("Asian", "Black or African American",
                     "Ethnicity not reported", "Hispanic or Latino",
                     "More than one race",
                     "Other race or race not reported", "White"),
                      field = "Engineering",
```

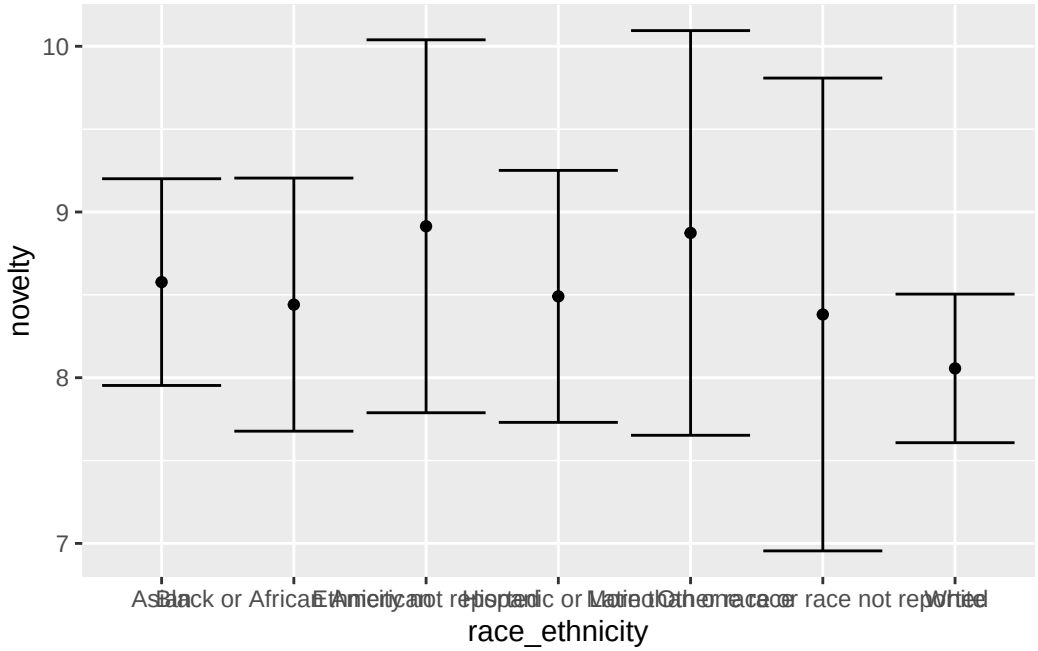
```
                      year = 2005)

preds <- predict(nmodel, newdata = fake_data_categorical, se.fit = TRUE)
glimpse(preds)
```

```
List of 4
 $ fit           : Named num [1:7] 8.58 8.44 8.91 8.49 8.87 ...
  ..- attr(*, "names")= chr [1:7] "1" "2" "3" "4" ...
 $ se.fit        : Named num [1:7] 0.318 0.39 0.574 0.388 0.623 ...
  ..- attr(*, "names")= chr [1:7] "1" "2" "3" "4" ...
 $ df            : int 4179
 $ residual.scale: num 4.98
```

```
fake_data_categorical <- fake_data_categorical |>
  mutate(pred = preds$fit, pred.se = preds$se.fit)
fake_data_categorical <- fake_data_categorical |>
  mutate(CI_lower = pred - 1.96*pred.se, CI_upper = pred + 1.96*pred.se)
glimpse(fake_data_categorical)
```

```
Rows: 7
Columns: 7
$ race_ethnicity <fct> Asian, Black or African American, Ethnicity not reporte~
$ field          <fct> Engineering, Engineering, Engineering, Engineering, Eng~
$ year           <dbl> 2005, 2005, 2005, 2005, 2005, 2005, 2005
$ pred           <dbl> 8.576926, 8.440986, 8.913791, 8.490757, 8.873536, 8.381~
$ pred.se        <dbl> 0.3182187, 0.3896102, 0.5741098, 0.3878813, 0.6229591, ~
$ CI_lower       <dbl> 7.953217, 7.677350, 7.788535, 7.730510, 7.652537, 6.954~
$ CI_upper       <dbl> 9.200635, 9.204622, 10.039046, 9.251005, 10.094536, 9.8~
```

```
gf_point(pred ~ race_ethnicity, data = fake_data_categorical) |>
  gf_labs(y='novelty') |>
  gf_errorbar(CI_lower + CI_upper ~ race_ethnicity)
```



The figure above shows model predictions illustrating how novelty is associated with race ethnicity. To make these predictions, field and year were held constant, field as Engineering and year as 2005.

Model Selection

```
car::Anova(nmodel)
```

```
Anova Table (Type II tests)

Response: novelty
               Sum Sq   Df F value   Pr(>F)
race_ethnicity    229    7  1.3223 0.235108
field             459    7  2.6464 0.009927 **
year               49    1  1.9606 0.161526
Residuals      103576 4179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions and Interpretation

Model assessment: all of the them met the condition so this is a appropriate model but the p-value suggests that it failed to reject the null hypothesis.

Prediction plot: the plot shows between different race groups they have different novelty score range and median so it shows the association between race ethnicity and novelty score.

null hypothesis: people in different race ethnicity and field from different years has impact on novelty level.

Model selection: the p-value from the ANOVA was 0.2351 which provides no evidence against the null hypothesis. So, according to this result, we are not quite confident that when race ethnicity changes, novelty does not change.