Case Study 1B: UPFs

Tyler Arista

2024-09-26

Planning a Model

1. Identify response variable & key predictor of interest
   - Response:
     - FM_change(change in fat mass)
   - Key Predictor:
     - Diet type(ultra_processed vs unprocessed)
2. Are there any confounding variables?
   - Age
     - Could affect both diet type and FM_change
   - Baseline fat mass(baseline_FM)
     - A subject's starting fat mass could influence how much fat is gained or lost depending on their diet
   - Baseline body weight
     - A subject's baseline body weight could influence both the diet's effectiveness & fat mass change
3. Are there any colliders?
   - No clear colliders
4. Are there any moderators of the predictor-response relationship?
   - Sex
     - A subject's sex could moderate the effect of diet type on change in fat mass.
5. Are there any precision covariates?
   - Resting energy expenditure(REE)
     - Only impacts FM_change & not expected to influence diet type
   - Leptin levels
     - Could also only affect FM_change & not influence diet type

6. Are there any mediation chains?
   - Diet type -> REE -> FM_change
     - Diet type might affect resting energy which impacts fat mass change
   - Diet type -> leptin -> FM_change
     - Changes in letpin levels due to the diet could influence fat mass change

Decision on final model

Going through our model planning checklist, I have included the following variables in my model:

- Diet type(key predictor)
- Age(confounder)
  - Helps control for differences in metabolism & fat storage between different age groups
- Baseline Fat Mass(confounder)
  - Controls for different starting points in fat mass, making sure that diet type is assessed independently of initial fat mass
- Sex(moderator)
  - Sees whether the diet type affects males & females differently

Fit your (new) Model

```r
model <- lm(FM_change ~ diet + age + baseline_FM + sex, data = upf_by_diet_data)

upf_by_diet_data <- upf_by_diet_data |>
  mutate(preds = predict(model),
         resids = resid(model))

summary(model)
```

```

Call:
lm(formula = FM_change ~ diet + age + baseline_FM + sex, data = upf_by_diet_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.63426 -0.20787  0.00577  0.23090  0.95951 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)   
(Intercept)          0.122021   0.287854   0.424  0.67332   
dietUltra-processed  0.474647   0.144739   3.279  0.00182 **
dietUnprocessed     -0.291608   0.144739  -2.015  0.04893 * 
age                 -0.011462   0.010836  -1.058  0.29485   
baseline_FM          0.002728   0.006183   0.441  0.66084   
```

```
sexMale              0.337081   0.133848   2.518  0.01478 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4577 on 54 degrees of freedom
Multiple R-squared:  0.4026,    Adjusted R-squared:  0.3473
F-statistic: 7.279 on 5 and 54 DF,  p-value: 2.839e-05
```

$$\widehat{FM}_{\text{change}} = 0.122 + 0.475 \cdot \text{diet}_{\text{Ultra-Processed}} - 0.292 \cdot \text{diet}_{\text{Unprocessed}} - 0.011 \cdot \text{age} + 0.002 \cdot \text{baseline\_FM} + 0.337 \cdot \text{sex}_{\text{male}} + \epsilon$$

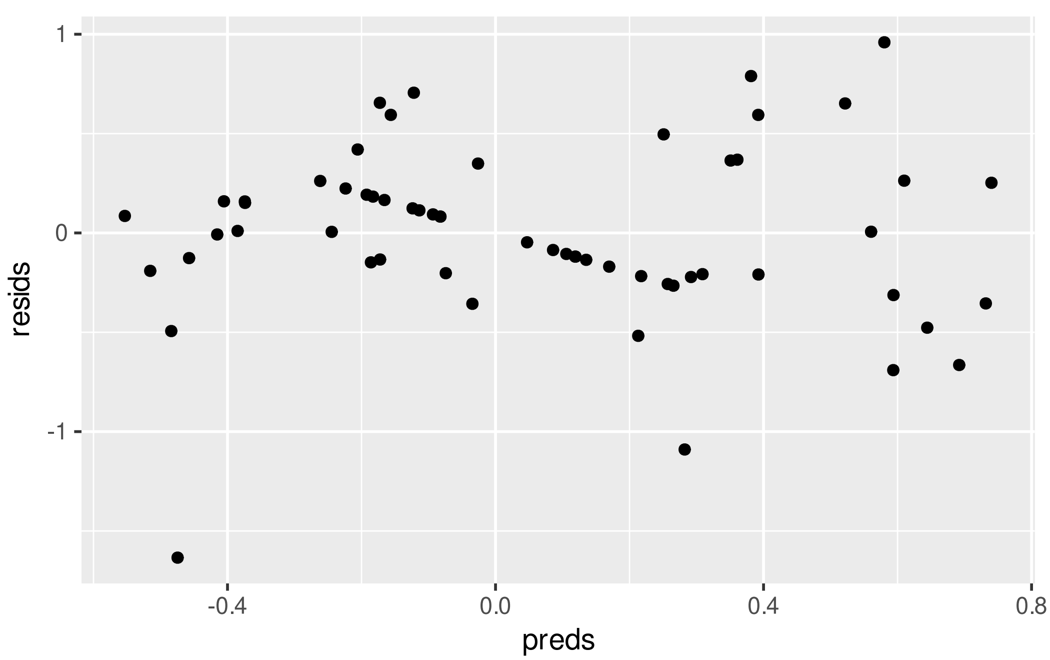
where: - diet(_{} = 1** if the diet is ultra-processed, and **0** otherwise. - diet(_{} = 1** if the diet is unprocessed, and **0** otherwise. - sex(_{} = 1** if the participant is male, and **0** if female.

$$\epsilon \sim N(0, 0.458)$$

Model Assessment

Residuals vs Fitted Plot

```
gf_point(resids ~ preds, data = upf_by_diet_data)
```



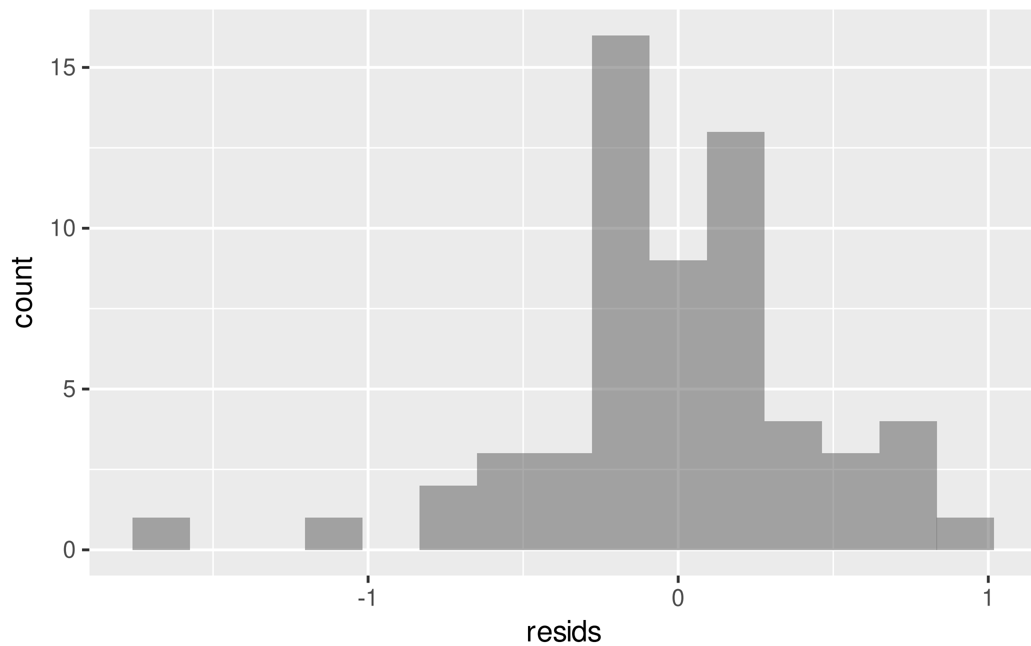
- Which condition(s) it helps you check
  - This helps us check lack of non-linearity
- Whether you think the condition(s) are met or not

- Yes, the conditions seem to be met

- What specific evidence you saw in the plot that allowed you to make your decision about whether the condition was met
  - The residuals are scattered randomly around the zero line and there is no clear pattern in the spread of the residuals, which means that the variance is constant

Histogram of the Residuals

```
gf_histogram(~resids, data = upf_by_diet_data,
             bins = 15)
```

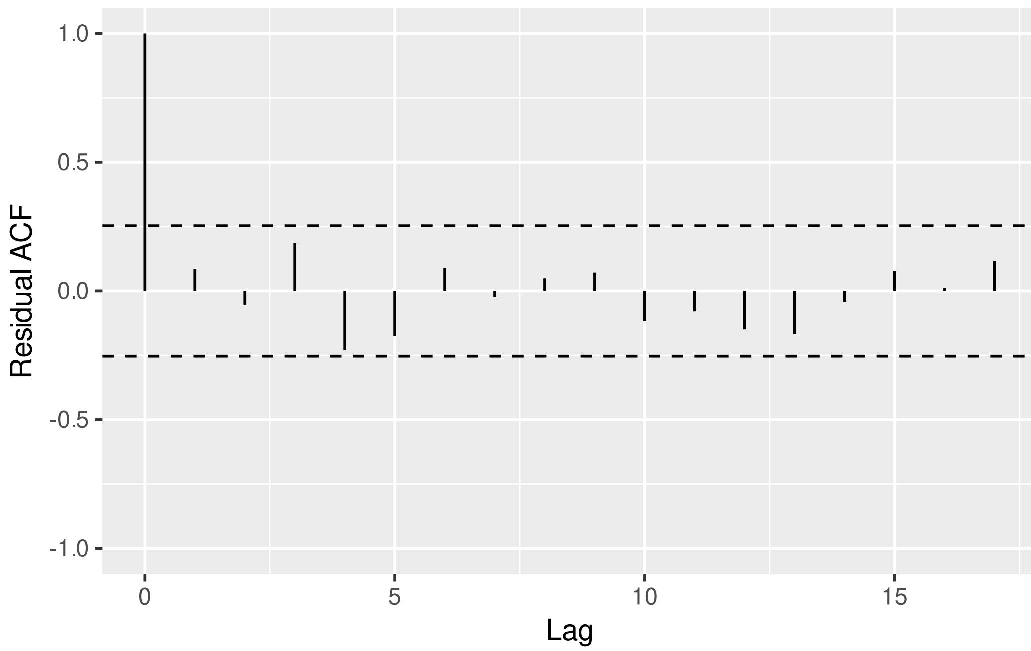


- Which condition(s) it helps you check
  - This plot helps check if the residuals are normal
- Whether you think the condition(s) are met or not
  - Yes, the condition seems to be met
- What specific evidence you saw in the plot that allowed you to make your decision about whether the condition was met
  - The histogram shows a somewhat symmetric distribution of residuals, with no extreme skewness or major outliers.

ACF Plot

```
```

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



- Which condition(s) it helps you check
  - This plot helps check independence of residuals
- Whether you think the condition(s) are met or not
  - No, the conditions don't seem to be met
- What specific evidence you saw in the plot that allowed you to make your decision about whether the condition was met
  - The ACF plot shows that not all of the autocorrelation values fall within the confidence intervals

Prediction Plot

###Hypothetical Data

```
expanded_data <- expand.grid(
  diet = factor(c("Ultra-processed", "Unprocessed")),
  age = mean(upf_by_diet_data$age, na.rm = TRUE),
  baseline_FM = mean(upf_by_diet_data$baseline_FM, na.rm = TRUE),
  sex = factor(c("Male", "Female"))
)
```

Make Predictions

```
preds <- predict(model,
                 newdata = expanded_data,
                 se.fit = TRUE)

glimpse(preds)
```

```
List of 4
 $ fit           : Named num [1:4] 0.643 -0.123 0.306 -0.46
  ..- attr(*, "names")= chr [1:4] "1" "2" "3" "4"
 $ se.fit        : Named num [1:4] 0.122 0.122 0.122 0.122
  ..- attr(*, "names")= chr [1:4] "1" "2" "3" "4"
 $ df            : int 54
 $ residual.scale: num 0.458
```

Convert to CI

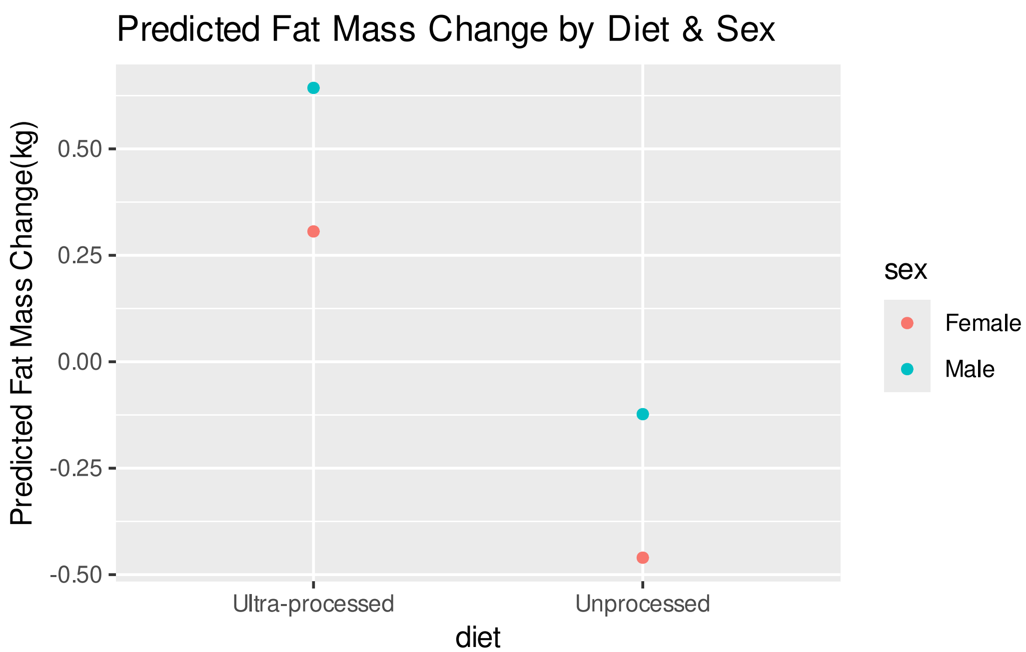
```
expanded_data <- expanded_data |>
  mutate(pred = preds$fit,
         pred.se = preds$se.fit,
         CI_lower = pred - 1.96 * pred.se,
         CI_upper = pred + 1.96 * pred.se)

glimpse(expanded_data)
```

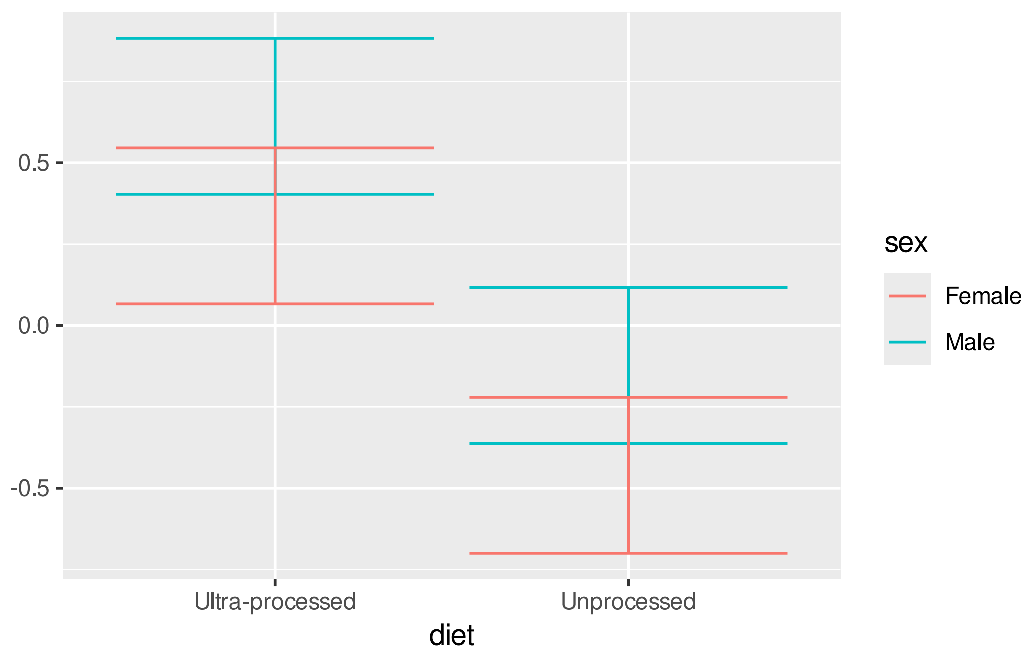
```
Rows: 4
Columns: 8
$ diet        <fct> Ultra-processed, Unprocessed, Ultra-processed, Unprocessed
$ age         <dbl> 31.2, 31.2, 31.2, 31.2
$ baseline_FM <dbl> 24.58544, 24.58544, 24.58544, 24.58544
$ sex         <fct> Male, Male, Female, Female
$ pred        <dbl> 0.6431880, -0.1230674, 0.3061066, -0.4601488
$ pred.se     <dbl> 0.1222845, 0.1222845, 0.1222845, 0.1222845
$ CI_lower    <dbl> 0.40351025, -0.36274510, 0.06642886, -0.69982650
$ CI_upper    <dbl> 0.8828657, 0.1166103, 0.5457843, -0.2204711
```

Prediction Plot

```
gf_point(pred ~ diet, color = ~sex, data = expanded_data) |>
  gf_labs(y = "Predicted Fat Mass Change(kg)", title = 'Predicted Fat Mass Change
by Diet & Sex')
```




```r
gf_errorbar(CI_lower + CI_upper ~ diet, color = ~sex, data = expanded_data)
```



Based on this prediction plot, it seems like that individuals on the ultra-processed diet tend to have a higher predicted fat mass change compared to those on an unprocessed diet, with some variation between males & females. But the overlapping confidence intervals suggest uncertainty in these predictions, and we can't draw firm conclusions from thsi plot alone.