Basic Literacy in Statistics Score Analysis

Priscilla Chen

Plan

Question: Is there an association between Calvin students' BLIS scores and the Timepoint?

Predictor: Timepoint

Response: BLIS scores

Confounding Variables: None.

Colliders: None.

Precision Covariates: courses and Duration. Courses may affect only the response variable, BLIS scores, because the higher level courses may teach more STAT knowledge and leads to higher BLIS scores. But Courses don't affect the predictor, Timepoint. Duration (time the student took to complete the BLIS) may affect the response variable BLIS scores, the more time the student spend on the test may indicate he/she was more careful and detailed, which may leads to higher scores, but Duration does not affect the timepoint. The dataset is big enough to satisfy the n/15 rule, so these two precision covariates should be included. The n/15 rule states that the number of parameters want to estimate need to be smaller to the sample size over 15. Because there are 293 rows of data, and four parameters, the n/15 rule is satisfied.

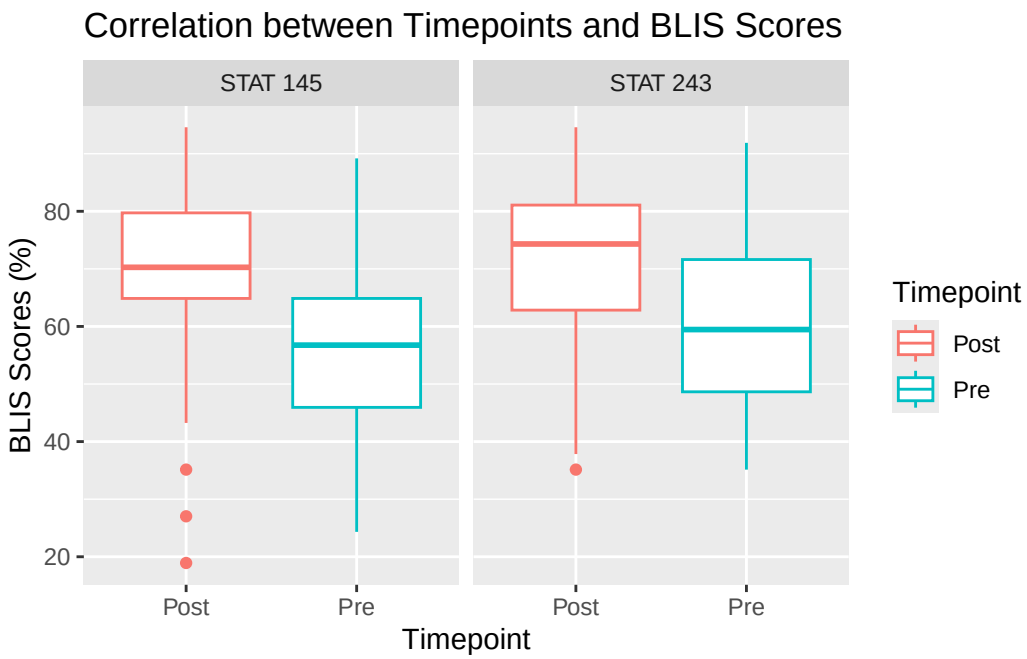
Mediation Chains: None.

Since the model planning and data loading are complete, we can explore the data using graphic.

Graphic

```
gf_boxplot(Percent~Timepoint |Course, color = ~Timepoint,
           title ="Correlation between Timepoints and BLIS Scores",
           data = file) |>
gf_labs(
    x = "Timepoint",
    y = "BLIS Scores (%)"
  )
```

```
Warning: Removed 51 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Explanation of Graphic: This boxplot graph shows a correlation between Timepoint and Blis Scores facet by courses, color shows different timepoint. The graph suggests that scores before courses (Timepoint = Pre) have lower mean than scores after courses (Timepoint = Post), and course STAT 243 seems to have a better BLIS score before and after course than STAT 145 course.

Fit

After exploring the data using graphic, we can fit a model using the parameters in our model planning.

```r
# create Model
mod <- lm(Percent ~ Timepoint
          + Duration
          + Course,
          data = cleaned_file)
summary(mod)
```

```
Call:
lm(formula = Percent ~ Timepoint + Duration + Course, data = cleaned_file)

Residuals:
    Min      1Q  Median      3Q     Max
-45.175  -8.712   1.448   9.027  34.761

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     63.09518    2.57537  24.499  < 2e-16 ***
TimepointPre   -13.89299    1.85624  -7.484  1.4e-12 ***
Duration         0.19708    0.06233   3.162  0.00177 **
CourseSTAT 243   3.28041    1.87966   1.745  0.08224 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.1 on 237 degrees of freedom
Multiple R-squared:  0.2313,    Adjusted R-squared:  0.2215
F-statistic: 23.77 on 3 and 237 DF,  p-value: 1.75e-13
```

Equation:

$$Percent = 63.095 - 13.89I_{pre} + 0.20Duration + 3.28I_{stat243} + \epsilon,$$

where: I_{pre}= 1 if timepoint = Pre; 0 if timepoint = Post;

I_{stat243} = 1 if Course = Stat243; 0 if Course = Stat 145

$$\epsilon \sim N(0, 14.1)$$

R_squared

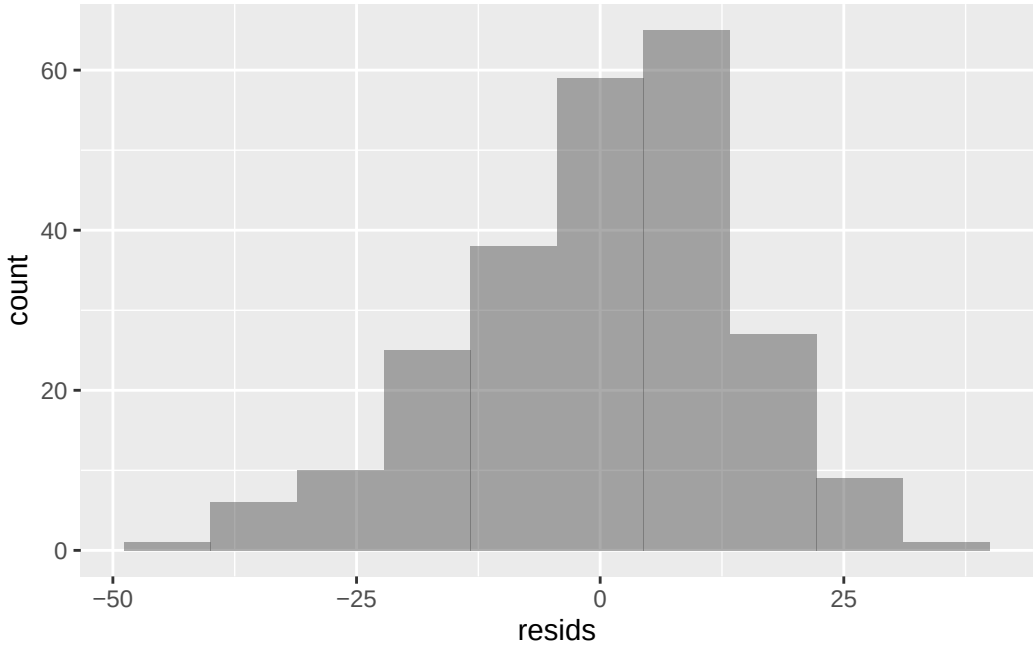
$R^2 = 0.2215$ (adjusted R-Square). The porpotion of the varience in response variable, BLIS score, that is explained by the predictors are 0.2215.

Assessment

In order to operate linear regression, we need to assess if the conditions are met.

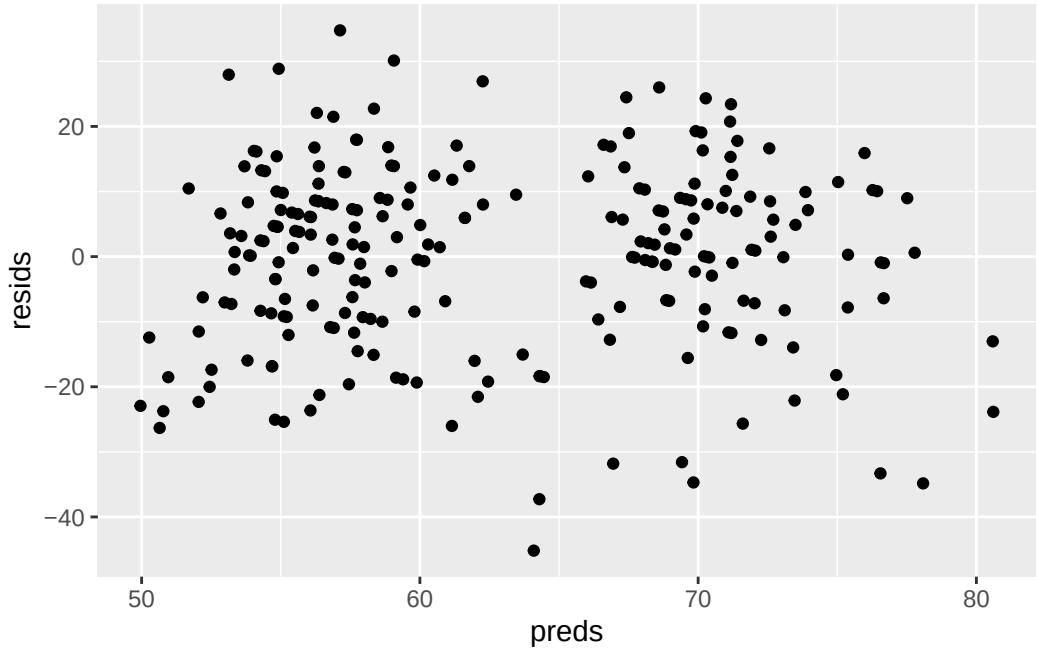
```
cleaned_file <- cleaned_file |>
  mutate(preds = predict(mod),
         resids = resid(mod))
```

```
# Residual Normal: Histogram
gf_histogram(~resids, data = cleaned_file, bins = 10)
```



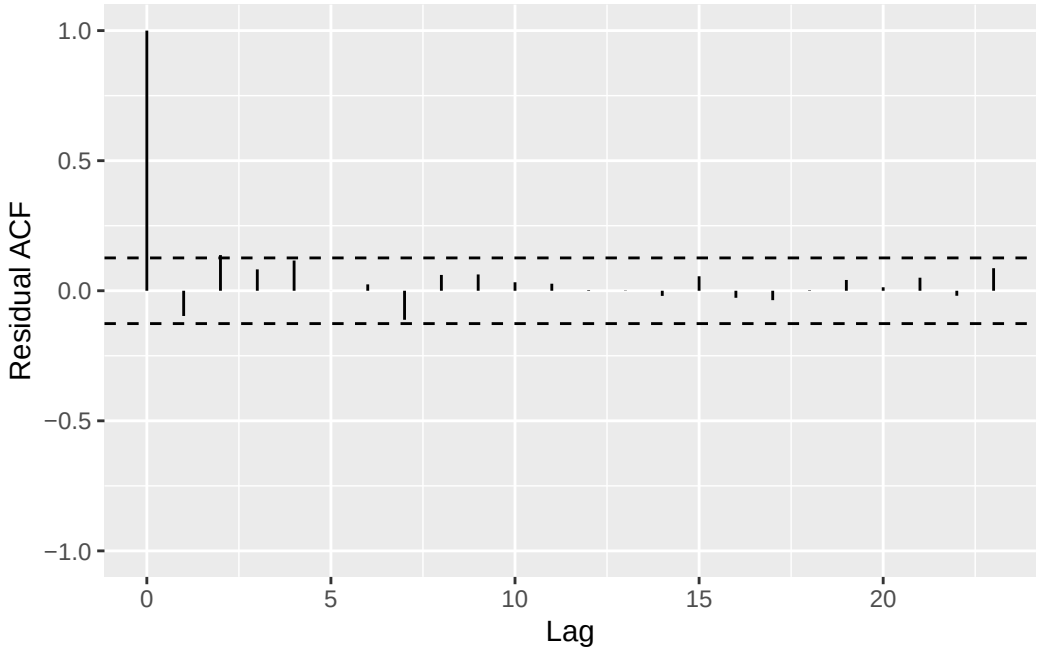
Histogram helps us check if our dataset is normally distributed, which is a requirement for fitting linear model. I think this condition is met because the residuals look unimodal, symmetrical, and loosely fit like a normal bell curve with the middle has higher counts than the left and right.

```
# Residuals VS Fitted Plot
gf_point(resids~preds, data = cleaned_file)
```



Residuals VS Fitted Plot helps us check if the dataset satisfies the lack of non-linearity and constant Residual Variance. I think both of these conditions are met because there are many points scattered randomly to indicate this graph as no trend and the points fit loosely into a rectangle between resids (-40,30), despite some outlier points that exceeds the rectangle.

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



The ACF plot checks the independence of residuals. I think this condition is met, because all lines except Lag(0) are within or slightly touching the confidence bounds.

Overall, I think all conditionals are met, so the model passes assessment and can be used to draw conclusions.

Interpretation

Prediction plot

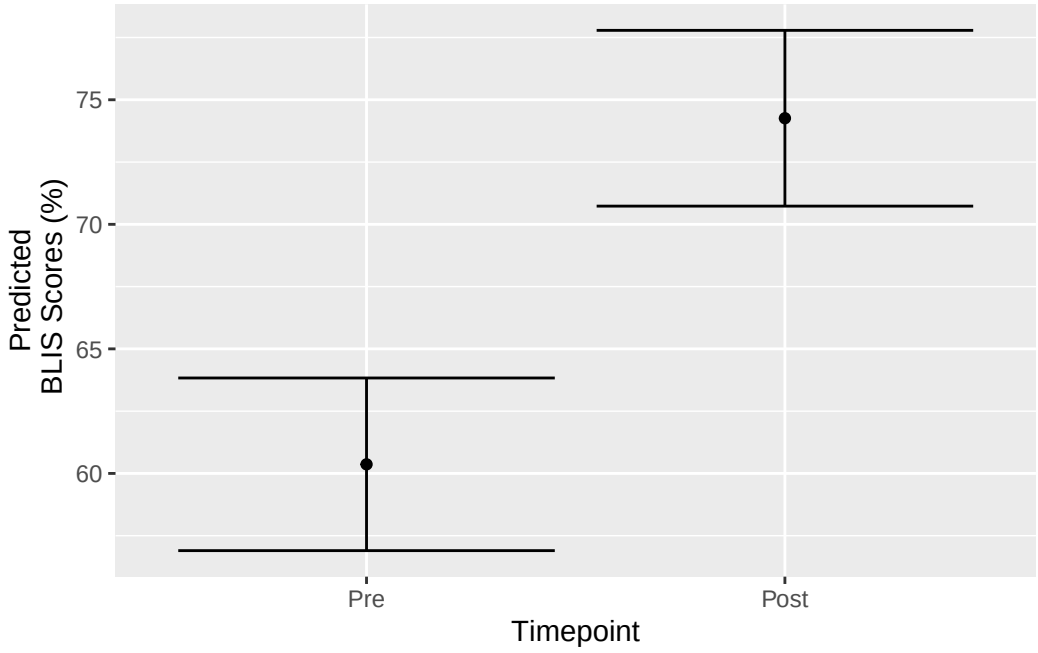
Since the model met all assessment conditions, we can proceed to making predictions:

```
# Create hypothetical data in which all others are kept constant.
fake_data <-
  expand.grid(Duration = 40,
              Course = "STAT 243",
              Timepoint = c('Pre','Post'))
preds <- predict(mod,
                 newdata = fake_data,
                 se.fit = TRUE)
```

```
fake_data <- fake_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(fake_data)
```

```
Rows: 2
Columns: 7
$ Duration  <dbl> 40, 40
$ Course    <fct> STAT 243, STAT 243
$ Timepoint <fct> Pre, Post
$ pred      <dbl> 60.36596, 74.25894
$ pred.se   <dbl> 1.768142, 1.799738
$ CI_lower  <dbl> 56.90040, 70.73146
$ CI_upper  <dbl> 63.83152, 77.78643
```

```
gf_point(pred ~ Timepoint,
 data = fake_data) |>
 gf_labs(y='Predicted\n BLIS Scores (%)') |>
 gf_errorbar(CI_lower + CI_upper ~ Timepoint)
```



Explanation for the prediction plot: This is a prediction plot of BLIS scores(%) from categorical predictor Timepoint.There are noticeable difference in BLIS scores between group before course (Timepoint = Pre) and group after course (Timepoint = post), while keeping the other variables constant. The variables are kept constant as the following: Duration = 40 and Course = "STAT 243".

Any relevant model selection result

```
# get Confident Intervals
confint(mod)
```

```
                     2.5 %      97.5 %
(Intercept)     58.02163544  68.1687214
TimepointPre   -17.54983090 -10.2361425
Duration         0.07429648   0.3198712
CourseSTAT 243  -0.42256070   6.9833843
```

```
# Hypothesis Test
car::Anova(mod)
```

```
Anova Table (Type II tests)

Response: Percent
          Sum Sq  Df F value    Pr(>F)
Timepoint  11133   1 56.0172 1.402e-12 ***
Duration    1987   1  9.9986  0.001771 **
Course       605   1  3.0458  0.082243 .
Residuals  47100 237
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

There is very strong evidence to suggest that BLIS score is associated with timepoint. The model passes all three model assessment test, therefore can be used in linear regression. The Anova inference suggests that there is very strong evidence (p_value = 1.402e-12) that BLIS score is associated with Timepoint, decreases 13.89 if Timepoint is Pre course (95% CI: -17.55 to -10.24) when other predictors are kept constant at Duration = 40 and Course = "STAT 243".