Test 1 Revisions: Assessing the Relationship Between Glycemic Status & Triglyceride Levels

Tyler Arista

2024-10-26

Revisions

I focused on improving clarity and flow throughout the report. My primary audience is college students who have some modeling knowledge but are not familiar with the STAT 245 course content. Key changes include:

- Adding linking text between sections to create a better flow & provide context
- Simplifying technical language where possible, but retaining accuracy
- Revising explanations of statistical concepts, making them accessible to a broader audience

Technical Revisions

- I made corrections to the following items from the Test 1 Rubric:
  - **Model Plan: Checklist: Checklist (–4 points)**
  - **Model Assessment, Residuals vs fitted (–2 points)**
  - **Model Assessment, Other graph (–2 points)**
  - **Prediction Plot: Graph relevant to question of interest shown (–3 points)**

Introduction

This analysis uses data from the Tehran Glucose Study, which explores the relationship between blood sugar and other health markers over time. One key marker of interest is triglyceride levels, a known indicator of cardiovascular risk. In this study, we aim to see if triglyceride levels are associated with different glycemic statuses (Normal, Prediabetes, Diabetes).

Model Plan

Research Question

Is there an association between a person's blood triglyceride level and their glycemic category (whether they have Diabetes)?

Response Variable: Triglyceride Levels

Triglyceride levels in the blood (measured in mmol/L) are chosen as the response variable because they serve as a key indicator of cardiovascular health, which may vary with different levels of blood sugar control. This makes triglyceride levels a meaningful measure for assessing the metabolic impact of glycemic management.

Predictor Variable: Glycemic Category

The predictor variable is **glycemic category**, which has three levels: **NFG/NGT** (Normal Fasting Glucose, Normal Glucose Tolerance), **Prediabetes**, and **Diabetes mellitus**. Glycemic category represents varying levels of blood sugar control, and it is hypothesized that poorer control (moving from NFG/NGT to Diabetes) may contribute to higher triglyceride levels, reflecting metabolic changes.

Confounders, Precision Covariates & Colliders

To ensure a comprehensive and unbiased assessment, the model considers potential confounders, precision covariates, and excludes colliders, based on the checklist:

Confounding Variables

These variables influence both the predictor (glycemic category) and the response (triglyceride levels), potentially creating misleading associations if not controlled for: - **Age:** Older individuals are more likely to have both higher triglyceride levels and diabetes. Including age as a confounder helps isolate the direct effect of glycemic category on triglyceride levels. - **BMI:** Higher BMI is associated with increased triglyceride levels and a greater risk of diabetes, making it a key confounder. Controlling for BMI helps reduce bias in estimating the effect of glycemic category.

Precision Covariate

- **Smoking Status:** Smoking affects triglyceride levels but does not directly influence glycemic category. It is included as a precision covariate to reduce residual variance and improve model accuracy.

Colliders

These variables are influenced by both glycemic category and triglyceride levels, so they are excluded to avoid bias: - **Low Physical Activity** and **Education:** Both are influenced by diabetes risk and triglyceride levels, making them colliders. To prevent biased results, these variables are not included in the model.

Model Checklist Review

Following the model planning checklist:

- The response and predictor variables are clearly identified
- Confounding variables (age and BMI) are included to control for their influence
- The precision covariate (smoking status) is included to improve model accuracy
- Colliders (low physical activity and education) are excluded to avoid bias.

Rationale

The relationship between glycemic category and triglyceride levels is of interest because it can indicate how metabolic health changes as blood sugar control worsens. By including age, BMI, and smoking status, the model aims to provide a more accurate and unbiased estimate of the effect of glycemic category on triglyceride levels.

n/15 Rule

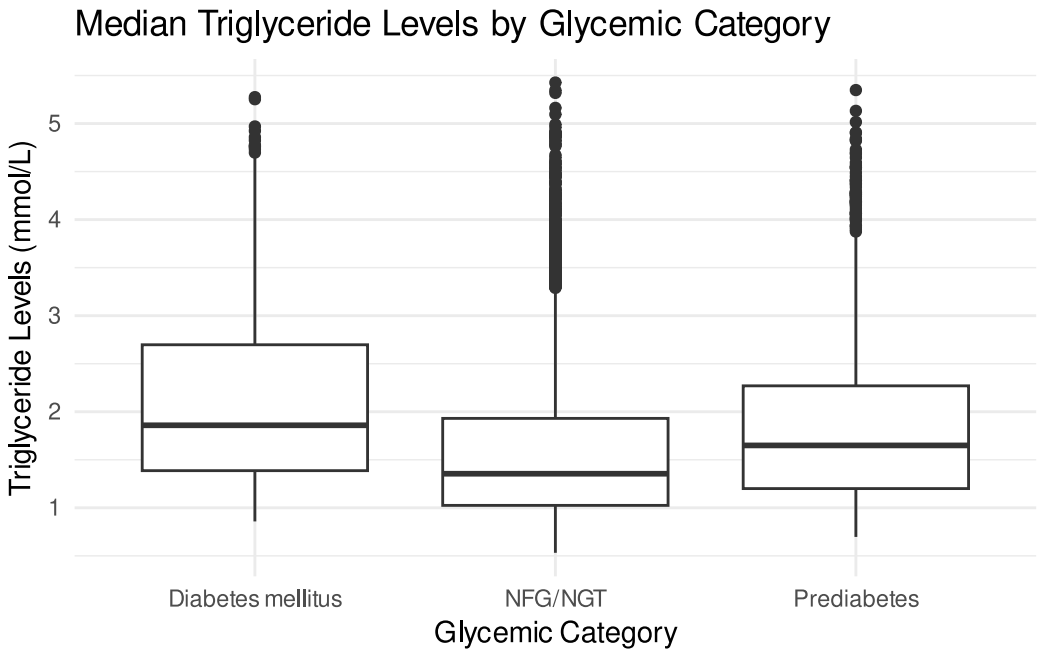
The n/15 rule is used to determine the maximum number of parameters that can be included in the model to avoid overfitting. The dataset has 7,718 observations, allowing for a maximum of approximately 514 parameters (7,718 / 15).

For this analysis: - **Glycemic category:** 2 parameters (one for each level beyond NFG/NGT). - **Age:** 1 parameter. - **BMI:** 1 parameter. - **Smoking status:** 2 parameters (one each for 'Past' and 'Current', with 'Never' as the reference). - **Intercept:** 1 parameter. - **Residual standard deviation:** 1 parameter.

This results in a total of **8 parameters**, which is well within the n/15 limit of 514. Therefore, the model adheres to the n/15 rule.

Data Exploration

```
gf_boxplot(triglyceride ~ glycemic_category,
           data = tgs_data,
           title = "Median Triglyceride Levels by Glycemic Category",
           xlab = "Glycemic Category",
           ylab = "Triglyceride Levels (mmol/L)") +
  gf_theme(theme_minimal())
```



Explanation of Graphic

- The box plot displays the **distribution of triglyceride levels** across three glycemic categories: **NFG/NGT**, **Prediabetes**, and **Diabetes mellitus**. The categories are ordered by median triglyceride levels to clearly highlight differences.
- **Median triglyceride levels** are highest among individuals with **Diabetes mellitus** (around **2.1 mmol/L**), followed by those with **Prediabetes** (around **1.9 mmol/L**), and lowest among those with **NFG/NGT** (around **1.6 mmol/L**).
- The **variability within each category** is represented by the interquartile range (IQR). The **Diabetes** and **Prediabetes** groups show wider IQRs, indicating greater variation in triglyceride levels compared to the **NFG/NGT** group.
- The title, "Median Triglyceride Levels by Glycemic Category," provides context, emphasizing that the plot focuses on the median differences among the glycemic categories.

Model Fitting

```
model <- lm(triglyceride ~ glycemic_category + age + BMI + smoking, data =
tgs_data)

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

summary(model)
```

```
Call:
lm(formula = triglyceride ~ glycemic_category + age + BMI + smoking,
    data = tgs_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3560 -0.6328 -0.2538  0.3672  3.8315

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.235e+00  6.969e-02  32.069  < 2e-16 ***
glycemic_categoryNFG/NGT    -5.173e-01  2.871e-02 -18.020  < 2e-16 ***
glycemic_categoryPrediabetes -2.418e-01  3.024e-02  -7.995 1.49e-15 ***
age                         -8.341e-05  7.986e-04  -0.104  0.91682
BMI                         -9.326e-04  1.232e-03  -0.757  0.44899
smokingNever                -9.399e-02  2.975e-02  -3.160  0.00158 **
smokingPast                  1.624e-02  4.064e-02   0.400  0.68949
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8322 on 7711 degrees of freedom
Multiple R-squared:  0.05514,   Adjusted R-squared:  0.0544
F-statistic:    75 on 6 and 7711 DF,  p-value: < 2.2e-16
```

$$3 \cdot \text{glycemic\_category}_{\text{NFG/NGT}} - 0.2418 \cdot \text{glycemic\_category}_{\text{prediabetes}} - 0.00008341 \cdot age - 0.0009326 \cdot BMI - 0.09399 \cdot \text{smoking}_{\text{Neve}}$$

where:

- $\text{glycemic\_category}_{\text{NFG/NGT}}$ = 1 if the glycemic category is NFG/NGT, and 0 otherwise.
- $\text{glycemic\_category}_{\text{prediabetes}}$ = 1 if the glycemic category is prediabetic, and 0 otherwise.
- $\text{smoking}_{\text{never}}$ = 1 if the smoking status is 'Never', and 0 otherwise.
- $\text{smoking}_{\text{past}}$ = 1 if the smoking status is 'Past', and 0 otherwise

$$\epsilon \sim N(0, 0.8322)$$

Report the (adjusted) R^2 value of the model

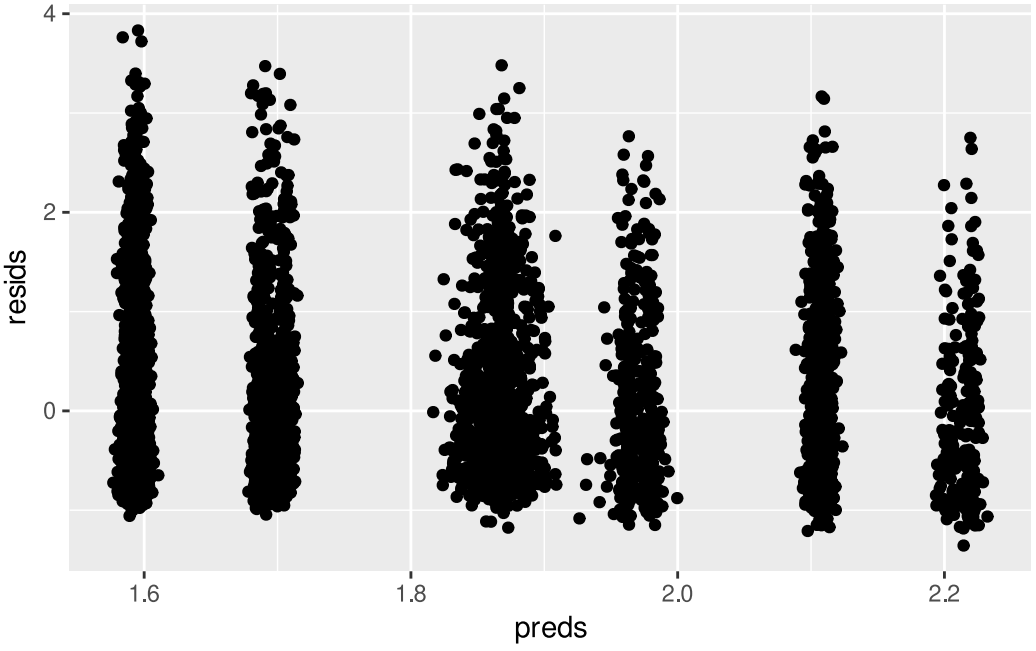
- Adjusted R-Squared: 0.0544
  - The adjusted R-squared value of 0.0544 indicates that about 5.44% of the variation in triglyceride levels is explained by the model, which includes glycemic category, age, BMI, and smoking status as predictors. While this suggests that glycemic category, along with other variables, contributes to predicting triglyceride levels, the majority of the variation remains unexplained, implying that additional factors are likely at play. It's important to understand that adjusted R-squared measures how well the model accounts for variation in the response variable, not the strength or detectability of individual associations. To assess the strength

of evidence for the observed relationships, we focus on p-values from the model summary, which indicate the likelihood that the detected associations could occur by chance.

Model Assessment

Residuals vs Fitted Plot

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

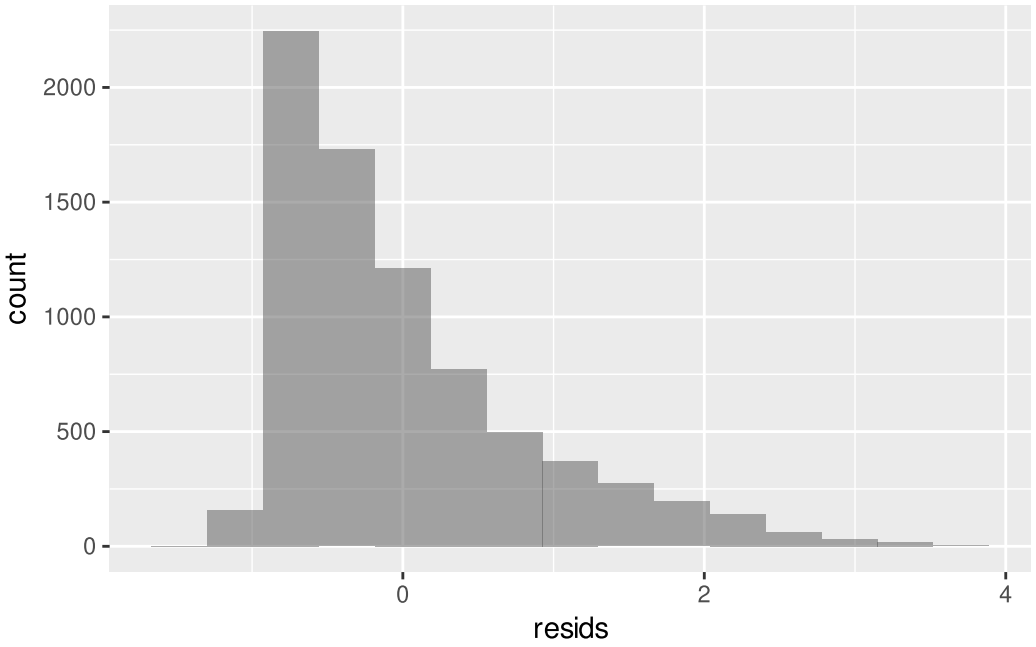


- Which condition(s) it helps you check
  - This plot checks for linearity and constant variance (homoscedasticity) in the model.
- Whether you think the condition(s) are met or not
  - **Linearity:** The condition appears to be met.
  - **Constant Variance:** There is evidence of heteroscedasticity (unequal variance), suggesting that the constant variance condition is not fully met.
- What specific evidence you saw in the plot that allowed you to make your decision about whether the condition was met
  - The plot shows that residuals are generally centered around zero across different levels of predicted values, indicating that the relationship between glycemic category and triglyceride levels is approximately linear.
  - However, the spread of residuals varies across predicted values. Specifically, there is greater spread at predicted values around **1.8** and **2.0**, which suggests potential heteroscedasticity.

This indicates that the variability of residuals is not constant, with some groups having more dispersed residuals than others.

Histogram of the Residuals

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



- Which condition(s) it helps you check
  - This plot checks the normality of residuals, which is a key assumption for the validity of the linear regression model.
- Whether you think the condition(s) are met or not
  - The condition of normality is not 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 right-skewed distribution of residuals, indicating a significant deviation from normality. This skewness is substantial enough to violate the normality assumption, which means that the model fails the assessment and cannot be expected to provide meaningful or valid results. As a result, any conclusions drawn from this model should be considered invalid, as the underlying assumption of normally distributed residuals is critical for accurate inference.

Interpretation

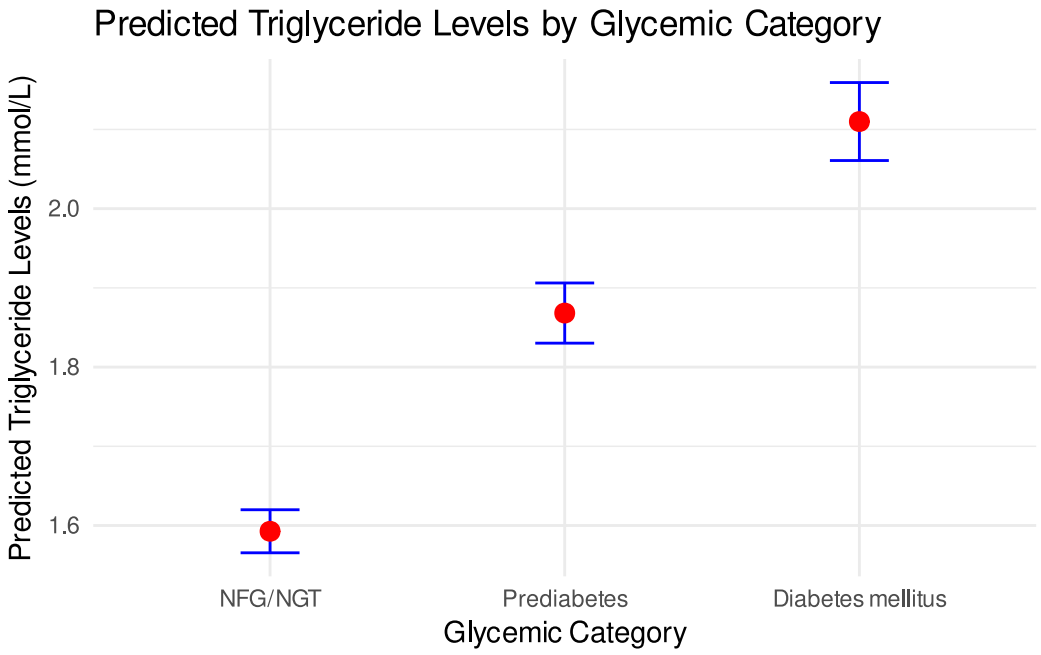
Prediction Plot

```
# Generate a hypothetical dataset for glycemic category
pred_data <- data.frame(
  glycemic_category = factor(c("NFG/NGT", "Prediabetes", "Diabetes mellitus"),
                             levels = c("NFG/NGT", "Prediabetes", "Diabetes
mellitus")),
  age = mean(tgs_data$age, na.rm = TRUE),   # setting age to mean
  BMI = mean(tgs_data$BMI, na.rm = TRUE),   # setting BMI to mean
  smoking = factor("Never", levels = c("Never", "Past", "Current"))  # setting
smoking to "Never"
)

# Make predictions with confidence intervals
preds <- predict(model, newdata = pred_data, interval = "confidence")

# Add predictions and confidence intervals to the hypothetical dataset
pred_data <- pred_data |>
  mutate(predicted_triglyceride = preds[, "fit"],
         lower_ci = preds[, "lwr"],
         upper_ci = preds[, "upr"])

# Plot the prediction with confidence intervals
gf_errorbar(lower_ci + upper_ci ~ glycemic_category, data = pred_data,
            width = 0.2, color = "blue") %>%
  gf_point(predicted_triglyceride ~ glycemic_category, data = pred_data,
           color = "red", size = 3) %>%
  gf_labs(title = "Predicted Triglyceride Levels by Glycemic Category",
          x = "Glycemic Category",
          y = "Predicted Triglyceride Levels (mmol/L)") +
  gf_theme(theme_minimal())
```



Prediction Plot Explanation

The prediction plot displays the expected triglyceride levels across different glycemic categories, holding age, BMI, and smoking status constant at their average or most common values. The red points represent the predicted triglyceride levels, with error bars showing the 95% confidence intervals. As shown in the plot, individuals with **Diabetes mellitus** have the highest predicted triglyceride levels, approximately **2.1 mmol/L**, followed by those with **Prediabetes** at about **1.9 mmol/L**. The lowest predicted levels are seen in individuals with **NFG/NGT**, around **1.6 mmol/L**.

The confidence intervals indicate the uncertainty in the predictions, and their lack of overlap suggests distinct differences in triglyceride levels across the glycemic categories. The intervals are narrowest for the **NFG/NGT** group and widest for the **Diabetes mellitus** group, reflecting varying levels of precision in the model's predictions. However, it is essential to note that since the model fails the assumptions of normality and constant variance, these predictions cannot be considered reliable or conclusive. Thus, while the plot suggests a relationship between glycemic category and triglyceride levels, the validity of these findings is compromised.

ANOVA

```
anova(model)
```

```
Analysis of Variance Table

Response: triglyceride
                   Df Sum Sq Mean Sq  F value    Pr(>F)
```

```
glycemic_category    2  297.1 148.566 214.5036 < 2.2e-16 ***
age                  1    0.0   0.003   0.0038    0.9506
BMI                  1    0.9   0.879   1.2691    0.2600
smoking              2   13.6   6.824   9.8522  5.33e-05 ***
Residuals         7711 5340.7   0.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Explanation

The ANOVA table evaluates the contribution of each predictor in explaining triglyceride levels:

- **Degrees of Freedom (Df)**: Glycemic category has 2 degrees of freedom, as it consists of three groups (NFG/NGT, prediabetes, and diabetes), while age and BMI each have 1 degree of freedom. The smoking variable has 2 degrees of freedom due to its three levels (Never, Past, Current).
- **F-value**: The F-value of 214.5036 for glycemic category indicates that it explains a significant amount of variation in triglyceride levels compared to residual variation, suggesting that glycemic category is an important predictor.
- **p-value**: The very small p-value (< 2.2e-16) for glycemic category suggests strong statistical evidence of an association between glycemic category and triglyceride levels. In contrast, age and BMI have p-values greater than 0.05, indicating no statistically significant association with triglyceride levels in this model. The smoking variable, however, has a p-value of 5.33e-05, indicating a statistically significant association with triglyceride levels.

Conclusion

The analysis indicates that glycemic category is associated with triglyceride levels, as seen in both the model's coefficients and the prediction plot. The prediction plot shows that individuals with **Diabetes mellitus** have higher predicted triglyceride levels than those with **Prediabetes** or **NFG/NGT**, suggesting a positive relationship between worsening glycemic control and triglyceride levels. However, the model explains only a small portion of the variation in triglyceride levels, with an adjusted R^2 of **5.23%**, indicating that many other factors likely contribute to the variation in triglyceride levels.

In terms of statistical evidence, the p-values for some coefficients suggest that there is a detectable relationship between glycemic category and triglyceride levels, particularly for individuals with **Diabetes mellitus**. However, it is essential to note that the model fails key assessment checks, including normality and constant variance of residuals. This failure undermines the reliability of the results and suggests that any conclusions drawn from this model are not valid. Therefore, while the results suggest a potential association, they cannot be considered conclusive.

To better understand the relationship between glycemic status and triglyceride levels, future models should include additional predictors, such as **age**, **BMI**, and **smoking status**, and address the assumption violations observed here. Incorporating more predictors and refining model assump-

tions would likely provide a clearer and more reliable understanding of how glycemic control affects triglyceride levels and cardiovascular risk.