Working with Data about an Individual Type 1 Diabetic - Failing to Fit Linear Models and Lessons Learned

Jeremy Chu

21/04/2021

Abstract

Through looking at type 1 diabetes data collected on an individual through a duration of 9 months, this study attempts to examine the factors attributed to the overnight fluctuations in an insulin-dependent diabetic's blood sugar levels, while illustrating the difficulties encountered in fitting a regression model to the available data. Through a variation of linear and logistic regressions, the study found that the individual blood sugar levels consistently rise overnight regardless of carb and insulin intake, potentially owing to either uncontrolled diabetes management or the dawn phenomenon. While the study aims to provide results that are more applicable towards a personal objective, there is hope that the success and failures of the regression models used here will help inform further type 1 diabetes research and offer a comprehensive individual dataset for additional research on how blood sugar levels fluctuate at night, and how food and insulin intake plays a part.

Contents

1 Introduction

Diabetes research is primarily centered around two fields: Studying the factors leading to diabetes, and finding critical factors that impact the lives of diabetic people. When approaching the latter topic, available research has been sparse and scattered, with some examining the relationships between meal frequency and meal timings with blood sugar levels (Jette Bertelsen and Hermansen 1993), and others comparing electrocardiograms (ECG/EKG) with continuous glucose monitors (CGM) to discover some previously unknown correlations (Fabien Dubosson 2018). Data on the lives of diabetics is sparse, and even when institutions generously make their data public, the time period of data collection is often quite short.[1] Borrowing the foundational research questions laid upon by previous diabetes researchers, this project looks at a single case for an extended period of time. Rather than comparing the diabetic experience of multiple individuals for a few weeks, I instead opted to examine 1 individual for 9 months. Please note that the individual is a type 1 diabetic, and therefore the data and results should only be taken in relevance to insulin-dependent diabetics, and should not be associated with the experiences of type 2 diabetics.

The individual in question is my fiancée, which makes the topic of consent much simpler than otherwise would be. The aim of the project is two-fold. First and foremost, this is a project designed to derive value for my fiancée, hence the questions will be seeking answers practical in

[1]The GitHub repository containing a collection of diabetes studies and open datasets can be found here: https://github.com/irinagain/Awesome-CGM/

nature. Secondly, I hope to be able to contribute to other available research; to see how a more extended observation period compares to a shorter but wider scope. The project aims to address three key questions:

1. Does the individual's blood sugar level follow a pattern of time of day
2. Does eating before sleep influence whether she wakes up with a high/in range/low blood sugar
3. Does taking insulin right before bed influence whether she wakes up with a high/in range/low blood sugar

The questions are structured around time rather than insulin and carb intake throughout the day due to the limitations of the data. As will be further detailed in the data source section, the data used for analysis is aggregated by hour, where insulin intake numbers increase when blood sugar level rises. This results in unfortunate regression lines where it gives off the illusion that the more insulin the individual takes, the higher her blood sugar. The project will therefore first go over and illustrate why the data is unfeasible to perform predictions on daytime blood sugar levels before analyzing the overnight changes of blood sugar in the subject.

One of the primary concerns for diabetic people is waking up in the middle of the night with a low blood sugar due to having too much insulin in them before bed, or waking up and realizing that they spent the night with a high blood sugar, resulting in having to start off the day groggy and tired due to something called the dawn phenomenon (Francesca Porcellati and Fanelli 2013).[2] Rather than focusing my efforts on predicting carb to insulin ratios, where the now present CGM and insulin pump combinations are already able to assist with their in-builit algorithms, this project aims to dissect what happens when the subject sleeps, and hopes to offer better insight into what to do to wake up in range. An understanding of the individual's nightly shifts of blood sugar levels would ideally connect my research to a wider discourse and tie back into studies dedicated to researching ways to manage the dawn phenomenon (Francesca Porcellati and Fanelli 2013; Teri B. O'Neal 2020). Ultimately, the goal is to personally derive value from this study, while at the same time offering the research community a unique set of findings alongside a long and extensive dataset to build upon. All files associated with this project can be found on GitHub.[3]

The paper will be structured as followed: the Data Source section will provide a brief rundown on data collection along with data transformation results. Sections Diabetes Terminology and About the Individual (Ethics) will further supplement the necessary background information about the project before going forward in terms of medical terminology and the level of consent for the data. The Looking at the Individual's Performance section will give an overview on the individual's blood sugar level across the time span of the project, illustrating that the project is working with an uncontrolled diabetic's data. Following that, the Models and Results section will first detail why the data is unable to support predicting blood sugar levels from carb and insulin intake, before moving on to logistic and linear regression models performed on data that specifically targets when the individual is asleep. This section will also explain why the models ultimately are not a good fit. Due to the nature of only working with one person's data, the Discussion section will explain why causality cannot be determined from this dataset along with other limitations in terms of conclusiveness and representability of the project. Despite this, I have opted to include a Moving Forward section at the end, discussing the opportunities in which other researchers could take this study and its data forward, especially in investigating managing the dawn phenomnon in people with uncontrolled diabetes.

[2]The dawn phenomenon implies the increase in blood sugar levels overnight, it is still under research

[3]GitHub repository for study: https://github.com/JeremyJChu/diabetes

Table 1: Original Insulin Data

Field Name	Description
Time	Time data was logged
Basal Amount	Amount of basal taken at the time of data log
Bolus Type	Whether bolus was fast-acting or normal
Bolus Volume	Amount of bolus taken at the time of data log
Immediate Volume	A setting on the pump where insulin can be quickly administered with a button press
Extended Volume	Type of bolus taken where users can take some insulin up front and have the rest admimnistered later
Duration	No data
Carbs	Carbs taken at the time of data log
Total Daily Dose	Summation of total daily bolus taken
Total Daily Basal	Summation of total daily basal
Serial Number	Serial number of device, removed from dataset prior cleaning

[a] 1 of 2 original datasets

2 Methods

All of the analysis was done using R (R Core Team 2020). Reproducible file paths utilized the here package (Müller 2020), and data cleaning/transformation was done with tidyverse (Wickham et al. 2019). All graphs were plot and arranged using a combination of ggplot2, gridExtra, and gghighlight (Wickham 2016; Auguie 2017; Yutani 2020). Tables were created using knitr::kable (Xie 2014) and kableExtra (Zhu 2020). The creation of the final product was aided with ggpubr, knitr, bookdown and magick (Kassambara 2020; Xie 2014, 2016; Ooms 2021). Regression diagnosis was performed using the performance package (Lüdecke et al. 2021).

3 Data Source

The data for this project comes from my fiancee's (name omitted for privacy's sake) Dexcom Continuous Glucose Monitor (CGM). A CGM is a device that tracks glucose levels 24/7, providing constant updates to the individual's phone and insulin pump, allowing for better diabetes management (Dexcom 2021). It also has an additional function of saving the data and allowing it to be extracted for further analysis. By connecting the CGM and insulin pump to a computer, individuals are able to download data on their glucose level readings, when they took insulin, when they had carbs etc. For the purposes of this project, I simply took that data, scrubbed it to ensure privacy, and cleaned it for use. The original datasets contained the fields shown in tables 1 and 2, and the cleaned version can be seen in tables 3 and 4. A comprehensive datasource datasheet can be found in the appendix.

Table 2: Original CGM Data

Field Name	Description
Time	Time data was logged
mmol/L	Blood sugar level at the time of data log
...3	No data
Serial Number	Serial number of device, removed from dataset prior cleaning

[a] 2 of 2 original datasets

Table 3: Cleaned CGM Data

Field Name	**Description**
Year	Year data was logged
Month	Month data was logged
Day	Day of the month data was logged (1-31)
Hour	Hour of day data was logged (0-23)
Blood Sugar	Average hourly blood sugar level at the hour of data log
Basal Amount	Total hourly basal taken at the hour of data log
Bolus Volume	Total hourly bolus amount taken at the hour of data log
Carbs	Total hourly carbs taken at the hour of data log
Time of Day	Morning, afternoon, or evening based on the hour
Time of Day Coded	Time of day coded into 1, 2, 3
Range	What the range of blood sugar is for the hour (low, in range, high)
Range Coded	Range coded into 1,2,3
Insulin Food	Whether bolus taken was with food or not (1: No carbs, Yes insulin, 2: Yes carbs, Yes insulin, 3: Yes carbs, No insulin)
Sleep	Whether the individual is asleep

[a] 1 of 2 cleaned and coded datasets

Table 4: Cleaned CGM Data

Field Name	**Description**
Year	Year data was logged
Month	Month data was logged
Day	Day of the month data was logged (1-31)
Bed Food	Did the individual consume food during the hours of 0-7
Bed Carbs	How many carbs did the individual consume between the hours of 0-7
Wakeup Range	What is the range of the individual at 7am (0: in range, 1: high, 2: low)
Night Insulin	How much insulin (bolus) did the individual take between the hours of 0-7
Any Insulin	Did the individual take any insulin at night (0: No, 1: Yes)

[a] 2 of 2 cleaned and coded datasets

4 Diabetes Terminology

For readers unfamiliar with diabetes, as the project revolves around medical data, some clarification of the fields and terminology used is necessary before proceeding.

Type 1 and Type 2 Diabetes

The two major types of diabetes. Type 1 diabetes' risk factors are associated with genetics and family history. It is an autoimmune disease where the pancreas is unable to produce sufficient insulin, thus resulting in patients being insulin dependent. It is a lifelong insulin dependency that cannot be cured or reversed.

Meanwhile, the risk factors for type 2 diabetes are mainly attributed to lifestyle to age, more controllable factors. It is a situation where the patient's body becomes resistant to insulin. Unlike type 1, type 2 diabetes is potentially reversible through a rigorously controlled diet and exercise. (Diabetes Care Community 2021).

A1c

Throughout this paper the term A1c will be mentioned sporadically. A1c, also referred to as hemoglobinA1c or glycated hemoglobin, is a blood test that measures the 2-3 month average of glucose in blood (Labtests Online 2021). In other words, it is a measure of how managed individual diabetics are. If they are usually in range, their A1c levels will reflect that, if they are uncontrolled, then likewise the A1c will return higher averages.

Blood Sugar/Glucose Levels

Blood sugar or glucose levels is simply the amount of sugar in a person's blood. For the purposes of this project, these values will be represented in mmol/L (millimoles per litre), the UK standard. In the US and continental Europe, the values would instead be represented in mg/dL (milligrams per decilitre). While the calculations differ slightly between the two, they perform the same function in measuring blood glucose concentration (Diabetes UK 2019). Once blood glucose levels fall or rise above a certain range, the individual can be labelled as currently having a low/high blood sugar. Low blood sugars can result in a loss of consciousness with a risk of death whereas high blood sugars for prolonged periods of time will eventually result in a host of health complications.

The general target range usually classifies less than 4 mmol/L as a low blood sugar, and higher than 11 as a high blood sugar. The dataset used in this project will reflect this range.

Continuous Glucose Monitor (CGM) and Insulin Pump

The data comes from the Dexcom G6 Continuous Glucose Monitor System, linked to the Tandem t:slim X2 insulin pump (Dexcom 2021; Tandem 2021). The Dexcom CGM is essentially a transmitter that continuously monitors an individual's blood sugar levels and reports the data onto the device's display screen. When connected to the Tandem insulin pump, the user has the option to choose between Control-IQ Technology and Basal-IQ Technology. In relation to this project, the data is collected while the individual was using Basal-IQ technology. In a nutshell, Basal-IQ connects with the Dexcom CGM to predict blood sugar levels 30 minutes in advance and stops insulin delivery if the user is expected to drop below a certain threshold.[4]

Basal and Bolus

[4]For more information please visit the Tandem website: https://www.tandemdiabetes.com/en-ca/products/t-slim-x2-insulin-pump/basal-iq

During this project, two types of insulin will be referred to: Basal and Bolus. The differentiation is simple. Basal insulin refers to insulin that is released in the background at all times, regardless of carb consumption; bolus insulin refers to insulin released in response to food (Giles 2020). There are also different characteristics of insulin brands that will not be written in detail, simply note that for this individual, rapid-acting insulin is taken. Rapid-acting insulin takes around 15 minutes to reach to blood stream and will continue to work for up to 4 hours after intake.

5 Literature Review

The main bodies of studies and literature that this project is founded upon is as follows:

5.1 Dawn Phenomenon and Overnight Control

Francesca Porcellati and Fanelli (2013) discuss and summarizes over 30 years of dawn phenomenon research, and specifically references the different methods of treatment prescribed to type 1 and type 2 diabetics. The authors were able to quantify to magnitude of the dawn phenomenon, that of a 15-25 mg/dL blood sugar elevation for type 1 diabetics, and discuss the effectiveness of oral treatment and basal insulin in managing the phenomenon, ultimately concluding that it is a subject that remains an area of research interest today despite the amount of research done on the subject. Teri B. O'Neal (2020) cover more of the diagnosis side of the dawn phenomenon, and discusses the development of a formula to "calculate the magnitude of early morning hyperglycemia without CGM." The authors further stress the importance of a more aggressive control over glucose to counteract dawn phenomenon. Tsalikian (2005) meanwhile looks at overnight glycermic control at a different angle, examining whether low blood sugars occur in children at night after afternoon exercise, ultimately finding that exercise in the afternoon does impact nightly blood sugar levels.

5.2 Food and Blood Sugar Levels

In relation to meal frequencies, Jette Bertelsen and Hermansen (1993) looked at the impact of meal frequencies for type 2 diabetics that do not take insulin, ultimately concluding that consuming more smaller meals throughout the day led to less blood sugar fluctuations than if the subjects had eaten two large meals. Nguyen Thanh Ha (2019) look at the association between adherence to dietary recommendation and patients' fasting blood sugar levels. The study found statistical significance in how much carbs a meal has and the number of meals a day with patients' fasting blood sugar levels. Taking it in a different direction, Heather Hall (2018) bring the investigation to healthy individuals. By using a CGM, the study monitors the impact of meals on the blood sugar levels of patients with no diabetes diagnosis, finding that individuals respond differently to different foods, but there are some foods that result in an elevation of blood sugar levels in the majority of adults. It is this study's hope to provide a type 1 diabetic window into the larger discourse, addressing situations where type 1 diabetics need to lose weight to reduce insulin resistance (The next section will elaborate why this is important).

6 About the Individual (Ethics)

A brief explanation on the circumstances surrounding the individual used for this case study. This does not impact the research done in this paper, but is rather used to provide further context for additional research down the line.

The individual is a female aged between 20-30, ethnically South Asian, and a type 1 diabetic. What differentiates her from a typical type 1 diabetic is that she is a type 1 with insulin resistance. While previously uncommon, these cases are becoming more and more prevalent and as such so will the usefulness of this case study. Additionally, the individual was misdiagnosed with type 2 diabetes at age 19 before getting a correct diagnosis of type 1 diabetes with insulin resistance at age 22.

7 Looking at the individual's performance

Before I get any deeper into running models to analyze and predict my fiancee's blood sugar levels and insulin intake, I would like to preface that during the year 2020 (the year the data was collected), she would be what you would call an uncontrolled diabetic.

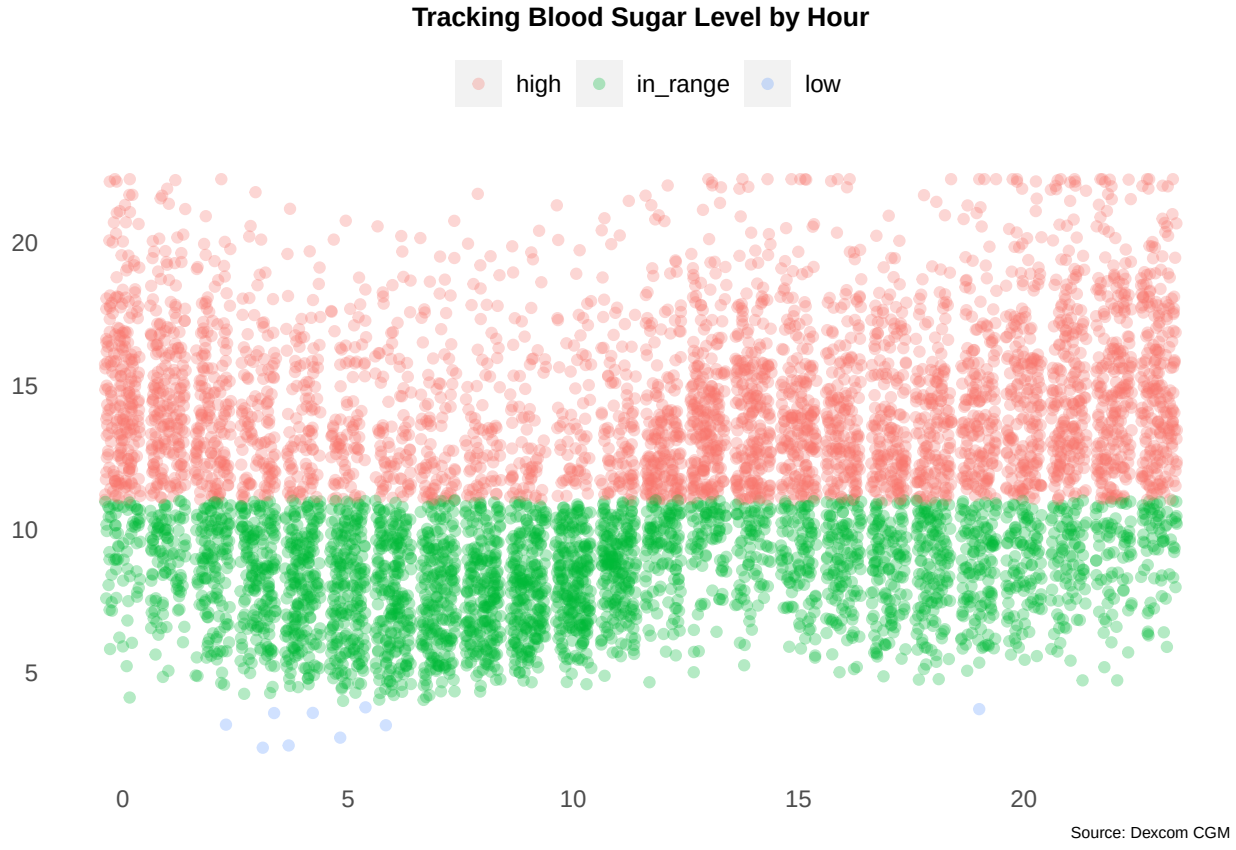


Figure 1: The individual spends significantly more time in a state of high blood sugar

Almost consisently, at all times, the individual has been in a state of high blood sugar rather than in range (See Figure 1). This suggests that the data reflects a situation in which the insulin she takes is insufficient to counteract both her natural blood sugar generation and the amount of carbs she

eats. As the data itself is skewed, note that any prediction done will be reflective of insulin intake for an uncontrolled diabetic. It is not representative of diabetics in control of their blood sugar levels and therefore the insulin intake presented in this project will undoubtedly be significantly higher if comapred to in control individuals. There will be more datapoints in which insulin is taken without carb consumption (See Figure 2) because of either insufficient insulin taken in the previous meal, or incorrect carb-insulin settings in the pump that failed to account for a need for increased insulin intake.

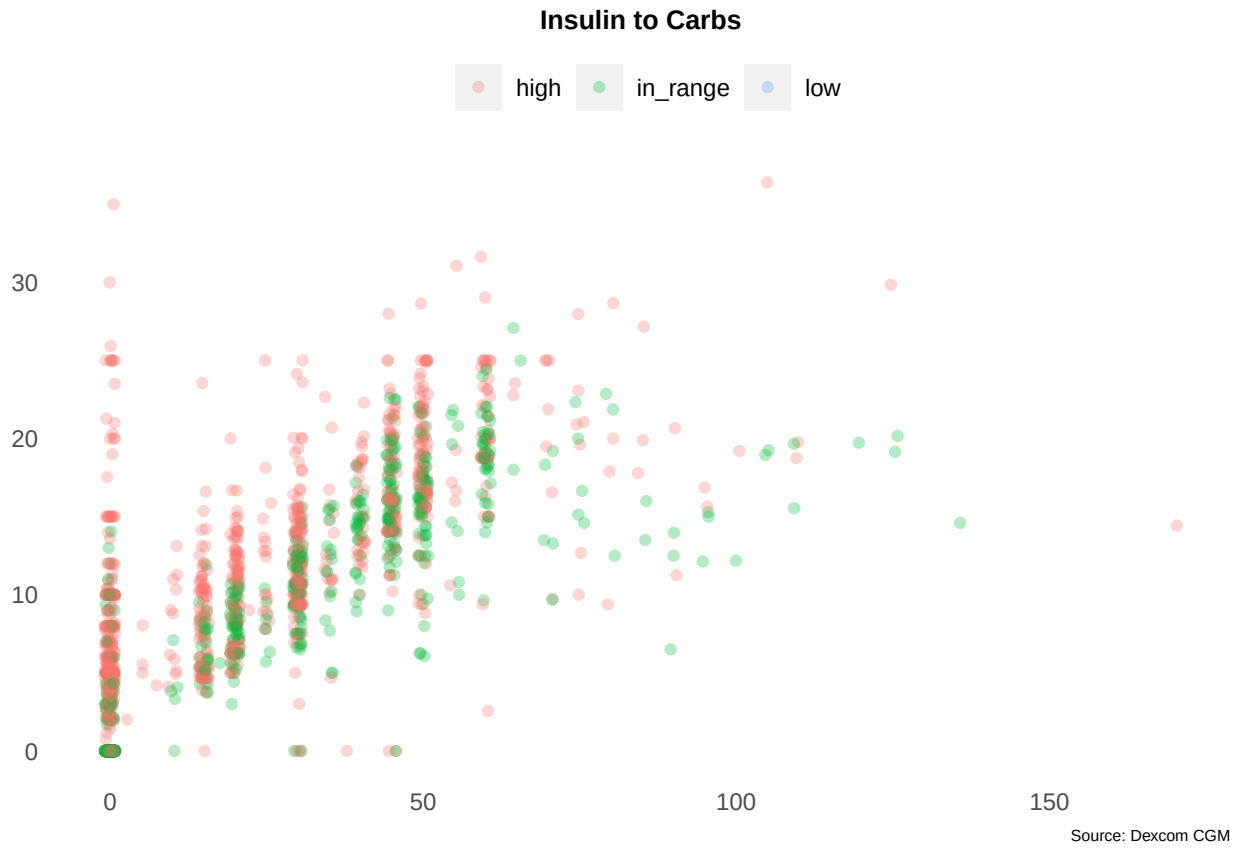


Figure 2: As expected, the individual takes more insulin when eating more carbs, but there is a significant amount of insulin taken when no carbs are eaten

7.1 Eating Habits. Meal Times, Frequencies.

In terms of insulin taking habits (See Figure 3), the individual typically starts the morning off in range. She then noticeably has the least amount of times she takes insulin in the afternoon. When the evening comes around, her blood sugar level is more often high than in range.

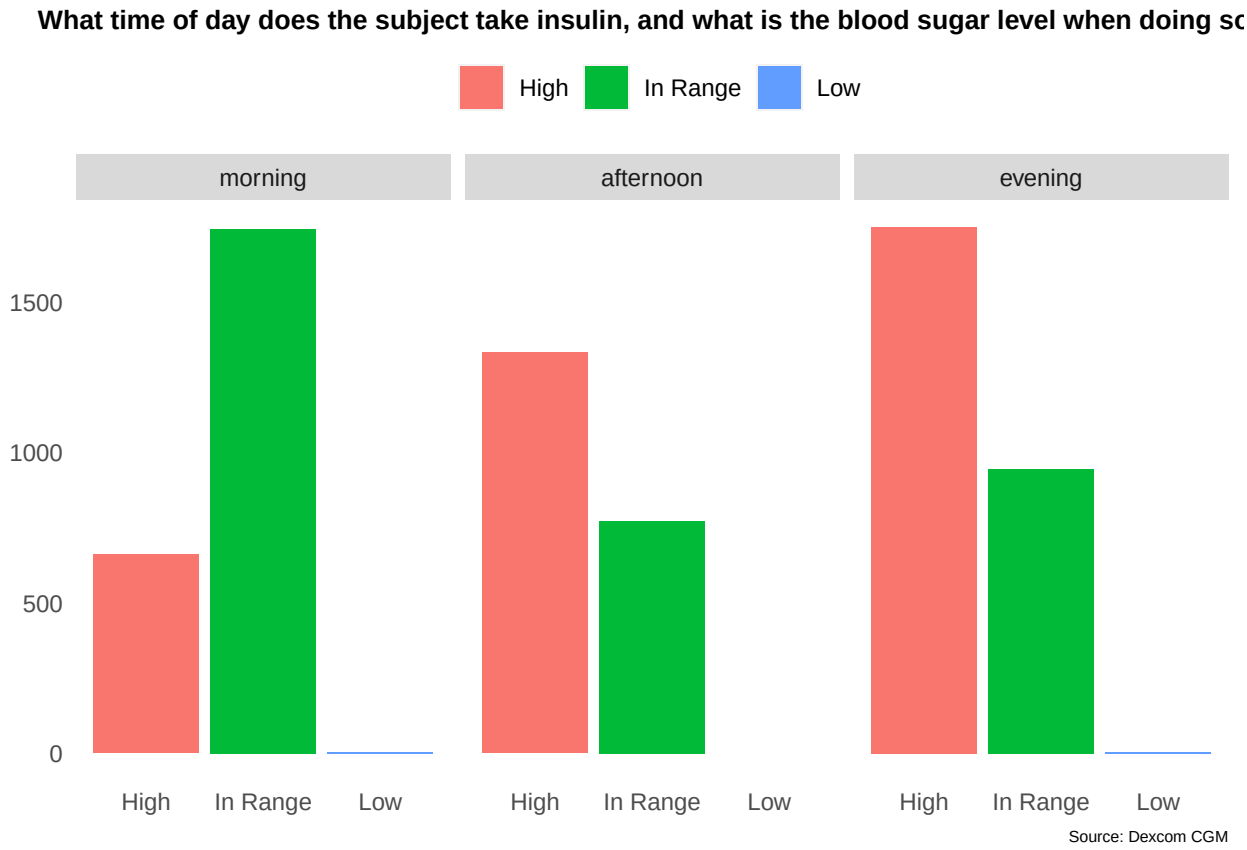


Figure 3: The subject starts the day off in typically in range, then gets more lax before ending up with typically high blood sugar levels in the evening

For meal frequencies, the subject on average eats around 3-4 meals a day (See Figure 4). Meals in this situation are defined as the consumption of carbs, so snacks are counted as 1 meal. A brief glance at WebMD suggests that the Recommended Daily Allowance is 130 grams of carbs per day (D'Arrigo 2015). A more in-depth look from the Centers for Disease Control and Prevention (CDC) reveals that ultimately there is no standard, but rather diabetics should aim to get half their calories from carbs (Centers for Disease Control and Prevention 2019a). Since I know the subject personally, the doctor recommendation provided was 30 grams of carbs for breakfast, 45 grams of carbs each for lunch and dinner, and 30 grams of carbs for snacks throughout the day. This brings up the subject's daily allowance to up to 150 per day. With the exception of December 2019, the subject has mostly fallen within this range (See Table 5). While the type of carbs consumed elicits varied reactions on blood sugar levels, the project will be progressing with the assumption that the subject is not overeating carbs and therefore leading to high spikes in blood sugar levels, but rather the reasoning for high blood sugar levels stem elsewhere.

Table 5: Carb Intake Breakdown

Year	Month	Total Carbs	Average Carbs/Day
2019	December	5498	183
2020	January	4055	135
2020	May	4460	149
2020	June	3195	106
2020	July	4745	158
2020	August	4640	155
2020	September	4475	149
2020	October	3387	113
2020	November	3013	100
2020	December	3820	127

[a] Average carbs rounded to nearest interger

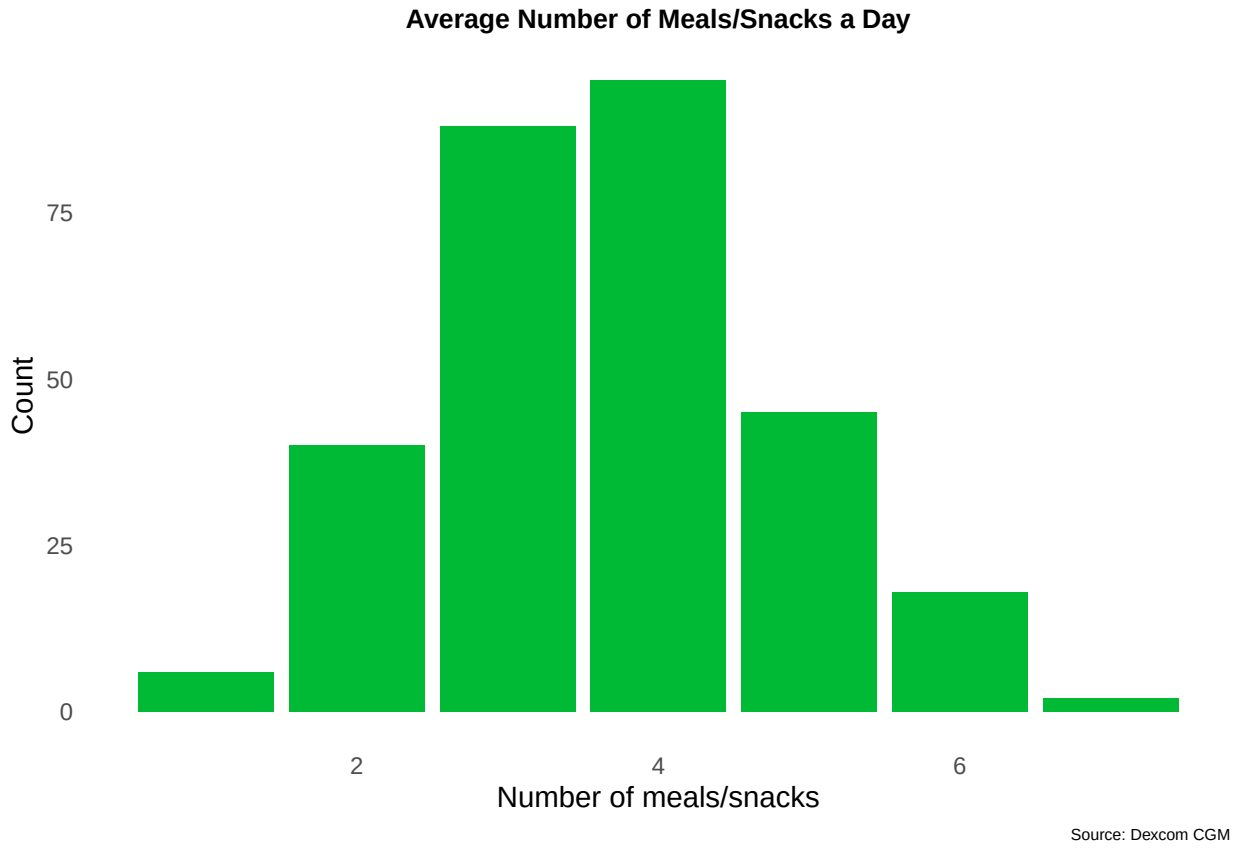


Figure 4: The subject typically eats 3-4 times a day

8 Model and Results

With all the factors outlined above, I decided to run a few statistical models on varying dependent variables. Initially, I wanted to evaluate and predict volume of insulin intake and blood sugar levels. Predicting insulin intake would allow for better adjustment of the tandem pump insulin settings

and carb ratios, whereas predicting blood sugar levels could reveal possible patterns in blood sugar levels in relation to meal frequencies or time of day.

8.1 Predicting Bolus Insulin Based on Carbs, Time of Day, and Basal

To start, I looked at whether carb intake, time of day, and basal amount had a correlation with how much insulin the subject took. Based on previous observations, it would be a natural assumption that insulin intake, that is bolus volume increases with the number of carbs consumed. The question is, would time of day also matter? As we were not shown otherwise during data exploration, my null hypothesis is no, time of day is not a significant factor correlated to how much insulin the subject takes. Meanwhile, the relationship between basal and bolus is as expected. For each unit increase in basal, insulin intake decreases by 1. As the functionality of insulin remains the same regardless of type, it is expected that an increase in one will result in a decrease in another. Results in table 6.

Table 6

	(1)
(Intercept)	3.506 ***
	(0.306)
carbs	0.251 ***
	(0.004)
time_of_day_coded	0.040
	(0.112)
basal_amount	-1.002 ***
	(0.082)
N	2376
R2	0.686
logLik	-6796.572
AIC	13603.144

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

A linear regression was chosen based on the assumption of linearity between carb intake and bolus amount. Accordingly, there should be no surprises in which bolus intake decreases with an increase in carb consumption. Time of day was chosen more in relation to being an uncertain variable. In an ideal situation, time of day should have no linearity with blood sugar levels as a controlled diabetic would have a flat and stable curve, keeping their blood sugar levels within the acceptable range. However, given that the individual is not in control during the scope of this study, the inclusion of time of day becomes a more investigative choice. As for basal amount, there should be linearity

present due to the idea that basal insulin is a constant supply of insulin that is periodically given to bring down "high resting blood glucose levels" (Medical News Today 2019). With higher basal amount given in an hour, theoretically there should be less bolus injected.

To test for validity of a linear regression for these variables, I implemented the performance package to check for linearity and collinearity in particular (Lüdecke et al. 2021). The diagnostics report can be seen in figure 5.

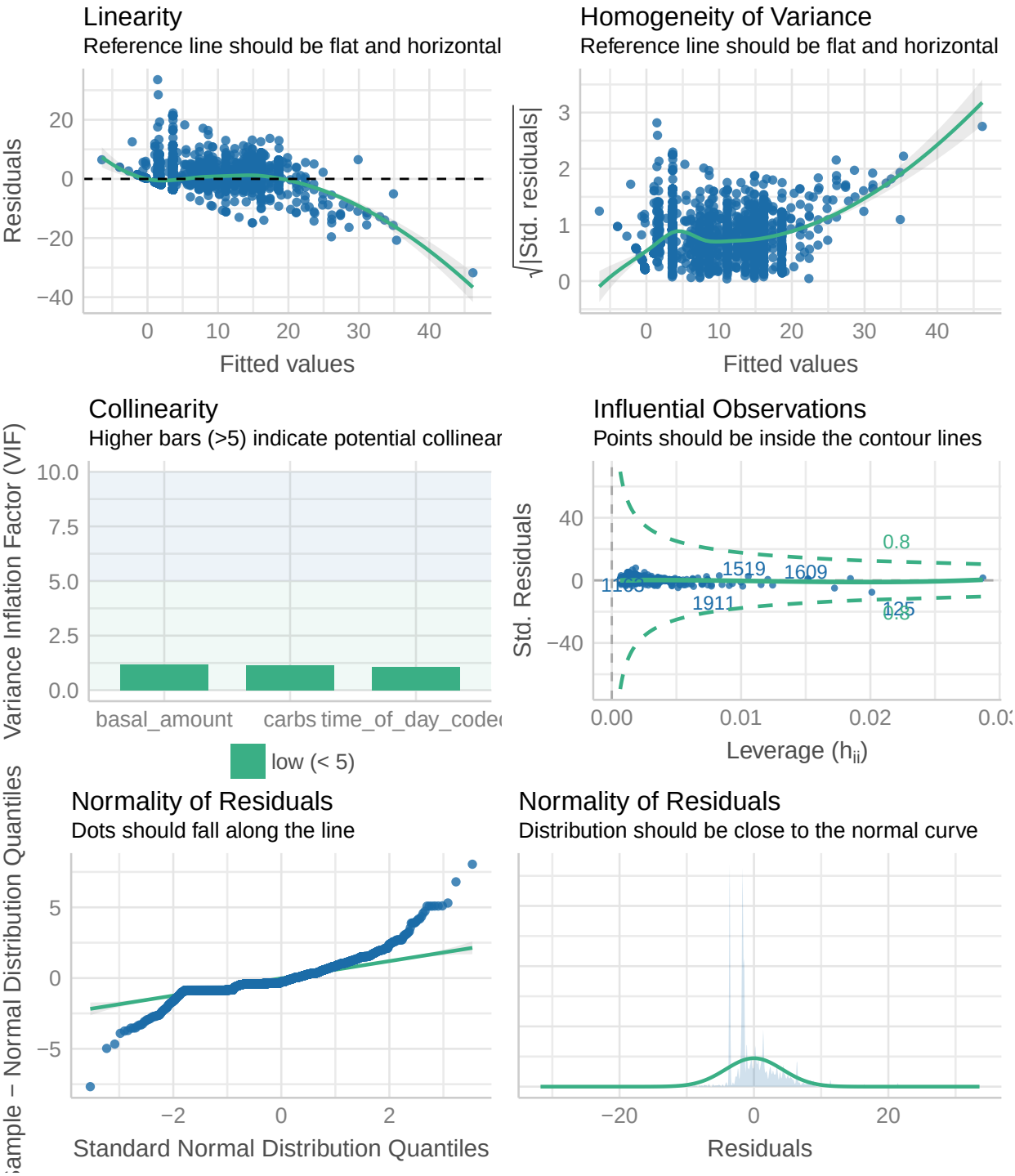


Figure 5: Diagnostics of linear regression

The results were as expected. Carb intake shows a positive, significant correlation with insulin intake (See figure 6). Insulin intake should increase by 0.25 units per 1 extra gram of carbs consumed. Similarly, insulin intake (bolus) decreases by 1 mmol/L for each unit of basal taken. However, the diagnostics are showing significant outliers in the data, something to keep in mind moving forward.

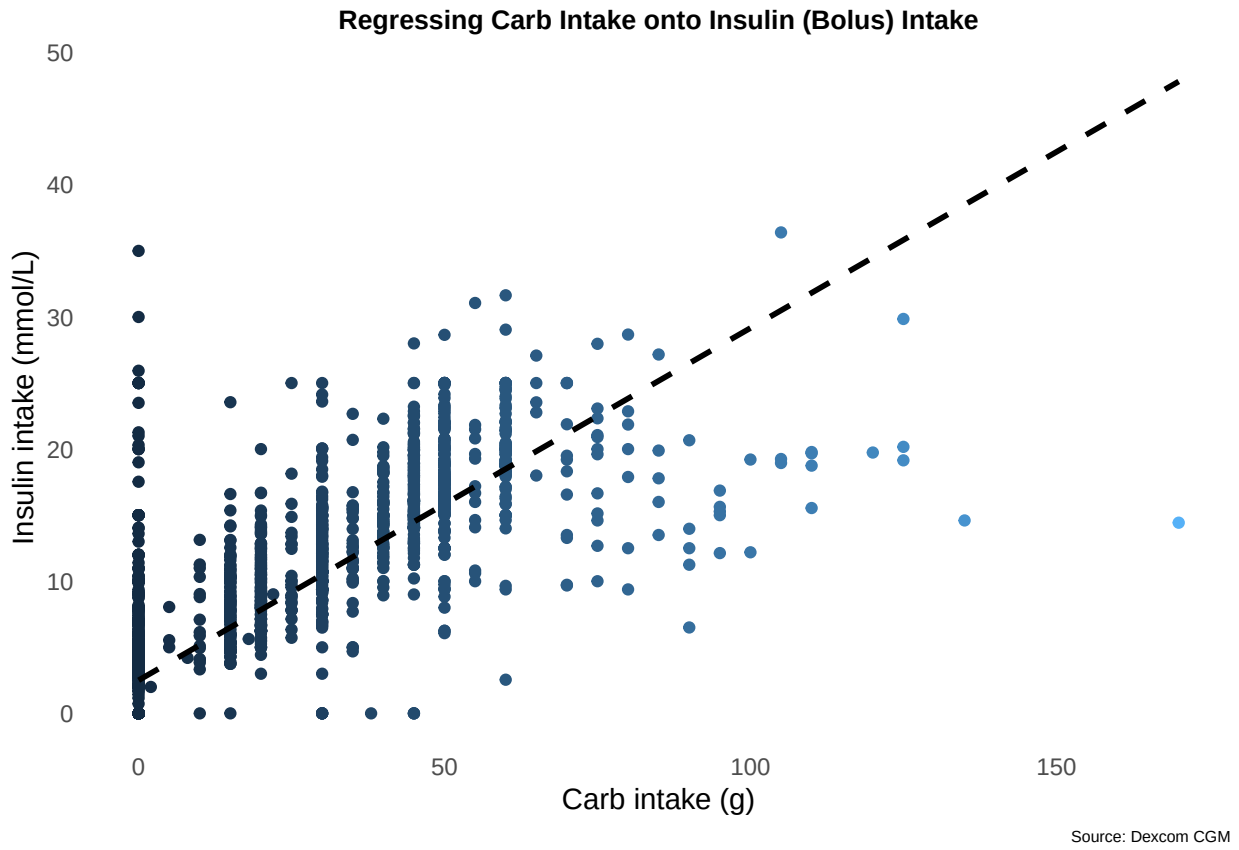


Figure 6: Positive correlation between carb intake and insulin intake

8.2 A Dead End Model - Predicting Blood Sugar Levels

Rather than insulin intake, which is a calculated measure that the insulin pump also calculates, a more valuable question is figuring out factors correlating with blood sugar levels. Table 7 shows the results of a regression predicting blood sugar levels based on carb intake, time of day, bolus amount, and basal amount.

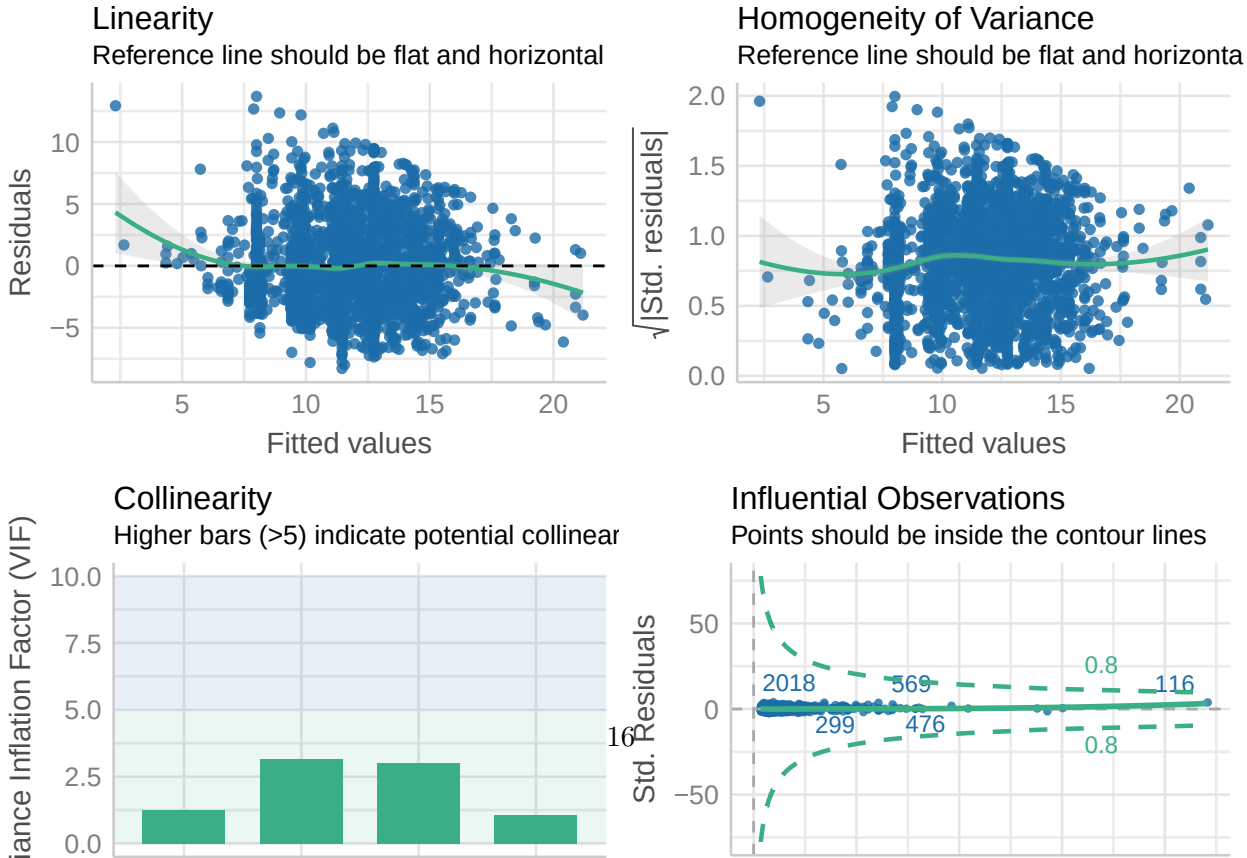
Despite all factors showing statistical significance, the results did not make sense. The data is showing a negative correlation with carbs, suggesting that for every 1g of carb eaten, the individual's blood sugar level would drop by 0.08 mmol/L, infringing on the very foundations of how blood sugar level operates. To make matters worse, the results also demonstrate a positive correlation with insulin intake. For unknown reasons, the individual blood sugar levels are hypothetically rising by 0.3 mmo/L per unit increase in insulin. Taking these results at face value, it would mean that the individual's blood sugar level decreases with carb consumption and increases with insulin intake.

Running a diagnostics on the model shows that none of the assumptions apply. There is absolutely no linearity nor normality of residuals. The variance is not homogenous. There are major errors in using a linear model for data that should have linearity. The question is why does data that should normally show linearity result in such chaos.

Table 7

	(1)
(Intercept)	7.768 ***
	(0.257)
carbs	-0.079 ***
	(0.005)
time_of_day_coded	1.659 ***
	(0.091)
basal_amount	-0.678 ***
	(0.069)
bolus_volume	0.326 ***
	(0.017)
N	2339
R2	0.325
logLik	-6203.912
AIC	12419.824

*** p < 0.001; ** p < 0.01; * p < 0.05.



Resolving this contradiction is simpler than it appears, and will explain why this paper must be cautions in working with the variables of insulin and carb intake. Figure 8 gives a window into the correlation between insulin intake and blood sugar levels. With no obvious relationship, it is clear why the regression results would not be reliable. Similarly, figure (9) details why carbs shows a negative correlation with blood sugar levels in the regression results. Ultimately, the non-linearity of the data triumps over the slight negative correlation that can be observed.

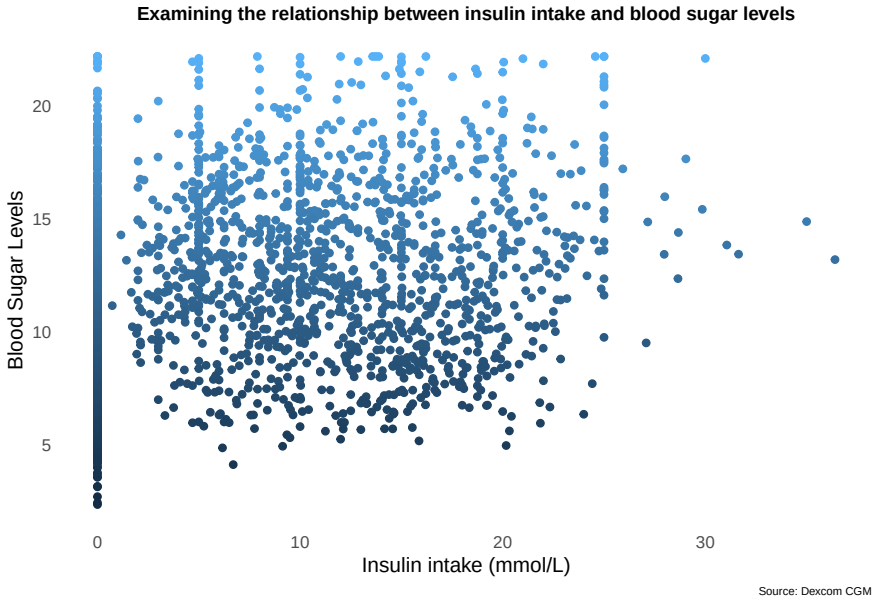


Figure 8: The non-linear relationship between insulin intake and blood sugar levels

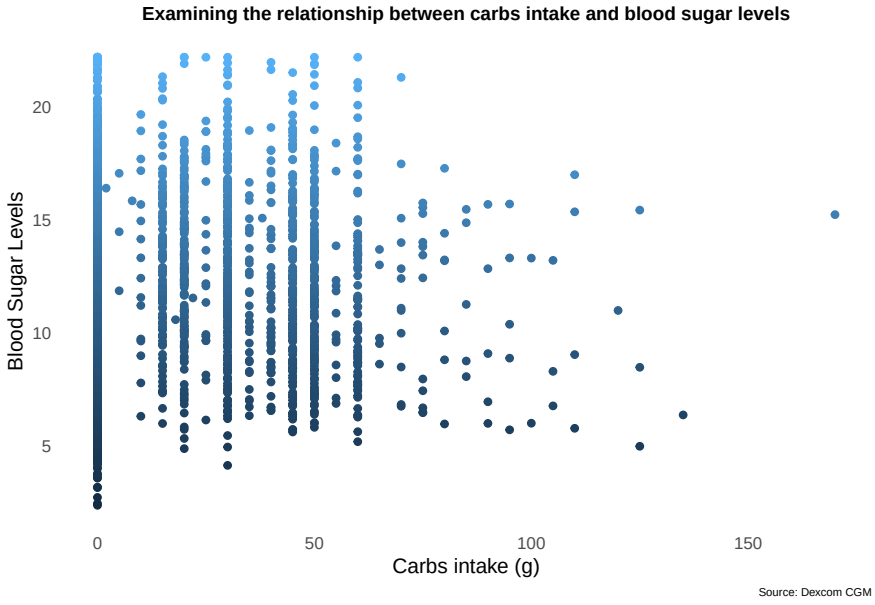


Figure 9: Non-linearity with slight negative correlation between carb intake and blood sugar levels

The problem is figuring out why, and it turns out to be more straightforward than expected. It all

comes back to the limits of the data. The data is collected in a manner where the CGM logs the user's current blood sugar level and saves that data to a time and date. Typically, best practice is to take insulin before eating food. So for example when the individual eats lunch, in this specific case study the probability of her blood sugar level being high is the greatest, she first logs how much carbs she is going to eat, and then administers the corresponding required amount of insulin. This means a generally higher blood sugar level is recorded when insulin is taken. All that data is logged before the meal is actually consumed. That insulin then works in her body for up to 4 hours. To further complicate the data, because insulin and carb intake are minority events throughout the day, the original dataset has countless null values where there is time and blood sugar level data, but not insulin and/or carb intake data. As such the data had to be cleaned in a way that it aggregated by hour. Therefore in the cleaned dataset it is natural for the association between the average hourly blood sugar level and combined hourly insulin intake to be positive, and for carbs to be strangely correlated with blood sugar levels. Ultimately, there are more occurrences, on average, for an uncontrolled diabetic in which insulin is taken in response to high blood sugar levels. In addition, the factors associated with food consumption and insulin intake are simply too numerous for an hourly aggregation to properly encapsulate. It is therefore unfortunate but expected that a simple linear regression would result in predictions that defy reality.

8.3 Bedtime

With that said, this is why it is impossible for this project to proceed in the direction of predicting blood sugar levels in the day based on insulin and carb intake. The data simply cannot support this line of investigation. Therefore, I instead opted to investigate the period of time in which there would be no, or at least be minimal carb and insulin intake - bedtime. I mention minimal because there are outlier situations in which the individual wakes up in the middle of the night with a low blood sugar and would have to subsequently rapidly consume a large amount of carbs. She then would have to take a careful amount of insulin to counteract the sudden intake of carbs while making sure she does not drop to a low again. For the most part, this study will ignore these outlier situations, instead focusing on the typical nights in which she sleeps through the night.

To do so, I turn to the second cleaned dataset (See table 4). The idea is to see what factors influence overnight blood sugar levels and allow for the individual to wake up in range. Setting the null hypothesis as eating at night or taking insulin at night does not influence the individual's blood sugar level, this is under the assumption that the insulin taken counteracts the carbs, I ran a logistic regression with whether the morning blood sugar level was in range (0) or high (1) as the dependent variable. Low was not an option simply because in the 9 months of data available, the individual has not woken up with a low. Note that this does not mean she did not have a low at night, but rather she did not wake up with a low at 7am. Table 8 shows the results.

To begin with, there was no significance in the results, at least in terms of p-value. The independent variables, "ate food at night" and "took insulin at night" were binary coded with 0 and 1. 0 implies a negative result, while 1 implies positive. Both eating food at night and taking insulin at night show higher log odds for waking up with a high than if the individual did not eat food or take insulin at night. But honestly, the data does not leave much for interpretation. Much like with previous attempts, the extent of what can be concluded is that there is a better hypothesis than simply looking at the interaction between eating food and taking insulin at night and the corresponding blood sugar level in the morning the next day.

A diagnostic on the model (see figure 10) indicates the presence of significant outliers. Their

Table 8: Logistic Regression for Whether Food/Insulin Intake Correlates with Waking Up In Range

	Estimate	**Std Error**	**z value**	**p-value**
Intercept	-1.20	0.19	-6.20	0.00
Ate food at night	0.37	0.35	1.10	0.28
Took insulin at night	0.19	0.33	0.58	0.56

Note:
DV: Wake up with blood sugar in range/high
IV: Ate/Did not eat food at night
IV: Took/Did not take insulin at night

presence is not surprising considering that the data of an uncontrolled diabetic is filled with outliers. The act of taking insulin without eating carbs has led to incredibly strange plots as outlined in previous sections. Even then, however, the act of taking correctional insulin exists, in which the individual takes small amounts of insulin to correct for the natural release of blood sugar in the body.

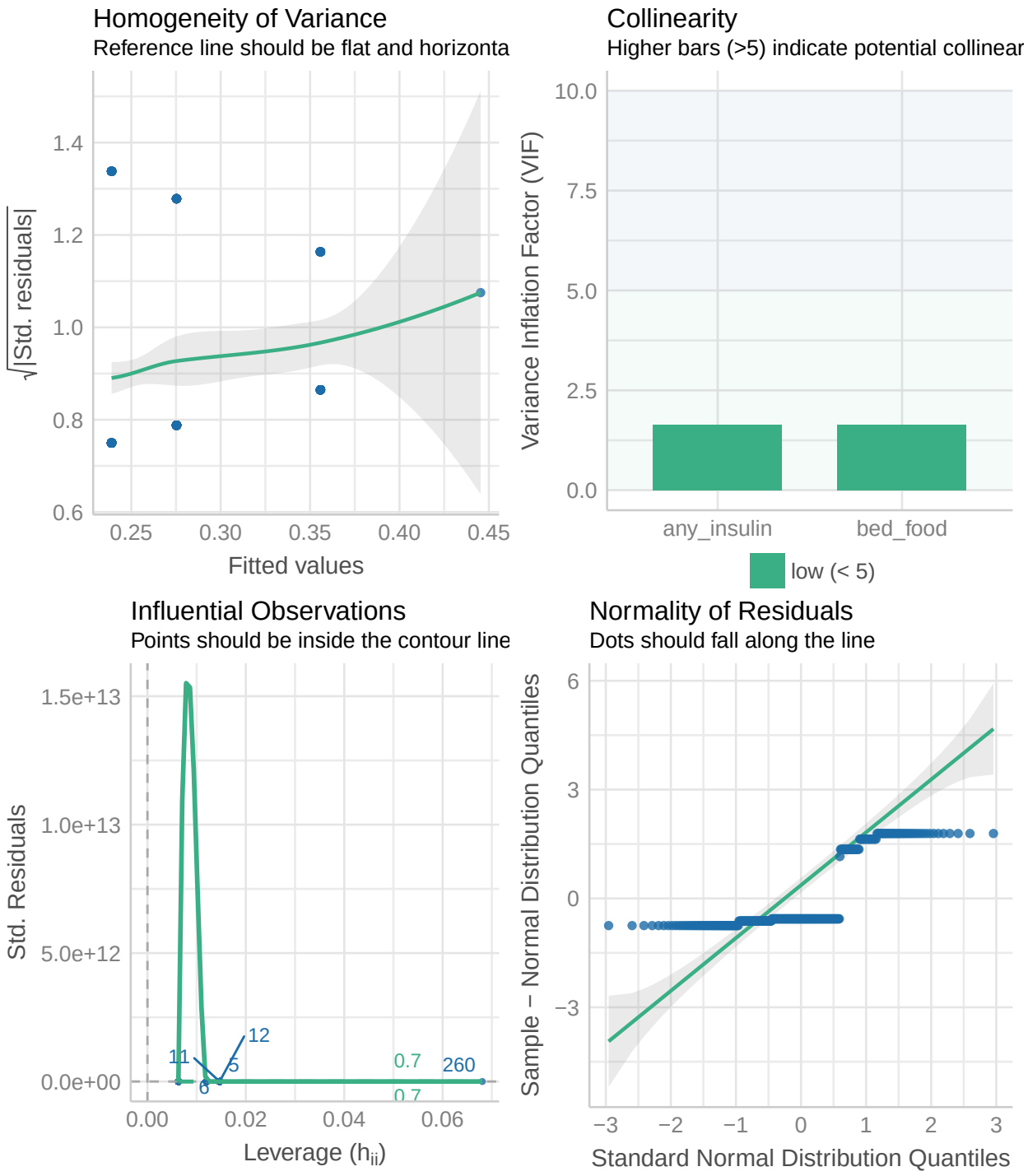


Figure 10: Diagnostics of linear regression

8.4 Splitting into Two Groups

As a final experiment, I split the sleep data into 2 groups. Group 1 accounts for nights where the individual took insulin after eating food, this includes situations in which she partakes in late night snacks or is required to consume additional carbs to prevent predicted low blood sugars. Group 2

covers nights in which the individual took insulin despite a lack of carb consumption, accounting for scenarios in which she has to take correctional insulin before bed due to an already high blood sugar level.

A brief glance at the outcomes of this split is shown in figure 11. For the duration of the study, in scenarios where the individual ate food and took insulin at night, the individual woke up in range 66% of mornings. In comparison, in nights where the individual took insulin without food, the individual woke up in range for 72% of mornings.

After running a multiple regression on both groups, I finally noticed some difference. For both regressions I chose the individual's morning blood sugar level as the dependent variable. The goal was to test for factors leading up to the individual waking up with her blood sugar in range. For independent variables, I firstly chose her blood sugar levels at midnight to see if there was any correlation between both time periods regardless of other factors. Additionally, I chose the amount of insulin the individual took during midnight to 7am as the second dependent variable. The assumption is that for the group where the individual took insulin without eating food, the more insulin the individual took, the lower her blood sugar levels at 7am would be. For the group in which she both ate food and took insulin, the relationship would be less clear as a lot more confounding factors would be present. Simply the act of taking insufficient insulin for her nightly carb consumption would affect the results.

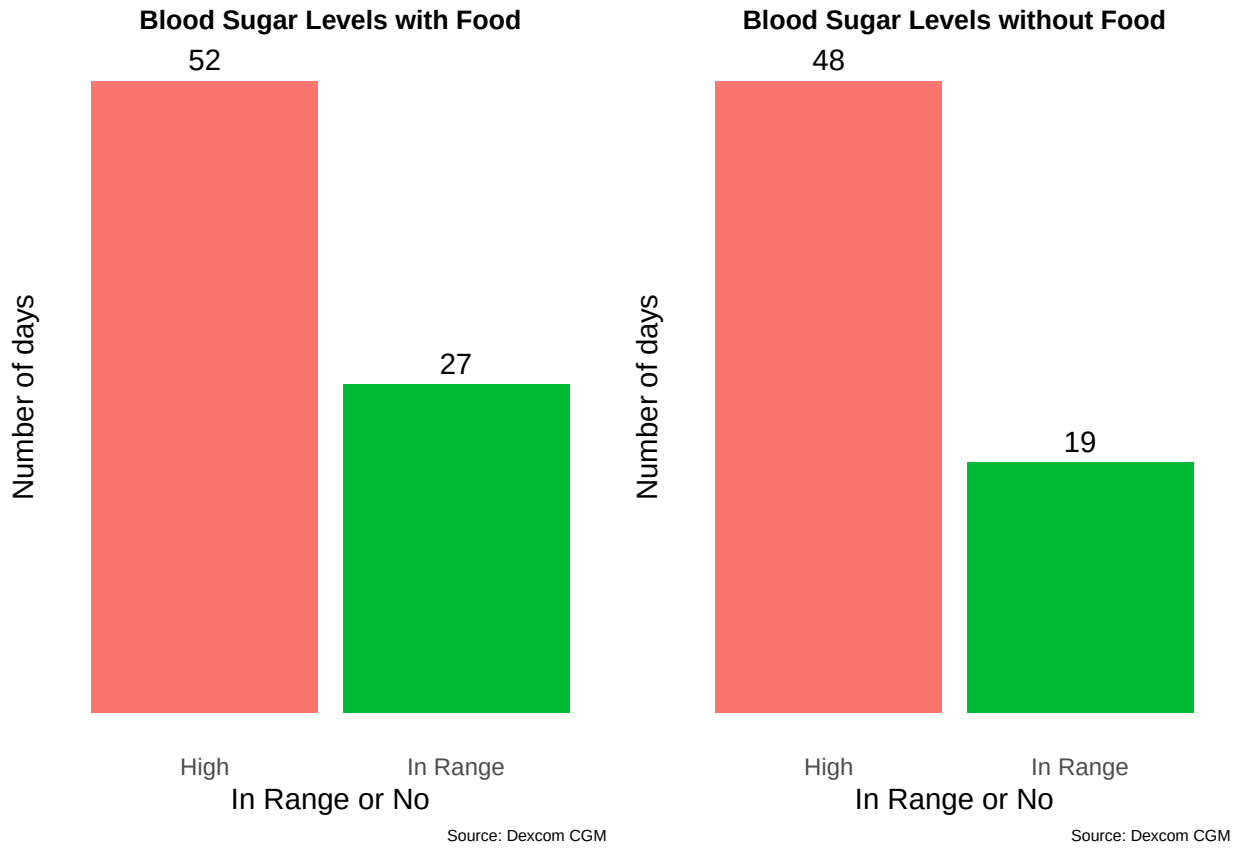


Figure 11: The individual is more likely to wake up in range during nights she did not eat

8.4.1 Group 1: Taking Insulin with Food at Night

Tables 9 show the results. In this case the null hypothesis can be posited as insulin and food intake in the same night has no correlation with the blood sugar levels in the morning. In addition, it is also hypothesized that blood sugar levels before sleep have no correlation with blood sugar levels in the morning. Results show that no significant values can be seen. There is slight positive correlation between both indpendent variables with the dependent variable, with morning blood sugar levels increasing by 0.065 with each 1 unit increase in insulin level before sleep; and morning blood sugar levels increase by 0.052 for each unit of insulin taken at night. This, however, cannot be considered conclusive, not simply because of the high p-value, but also the idea that these variables cannot tell the whole story. Diagnosis results (See Figure 12) show that for the most part, the model relatively passes the linearity, collinearity, and normality checks.

Table 9: Linear Regression on Blood Sugar Levels in the Morning on Nights with Food and Insulin Intake

	Estimate	**Std Error**	**t value**	**p-value**
Intercept	7.800	1.700	4.60	1.7e-05
Insulin Level at Night	0.065	0.140	0.46	6.4e-01
Insulin Took at Night	0.052	0.085	0.61	5.4e-01

Note:
DV: Blood sugar levels when awake
IV: Insulin level at night
IV: Amount of insulin took at night

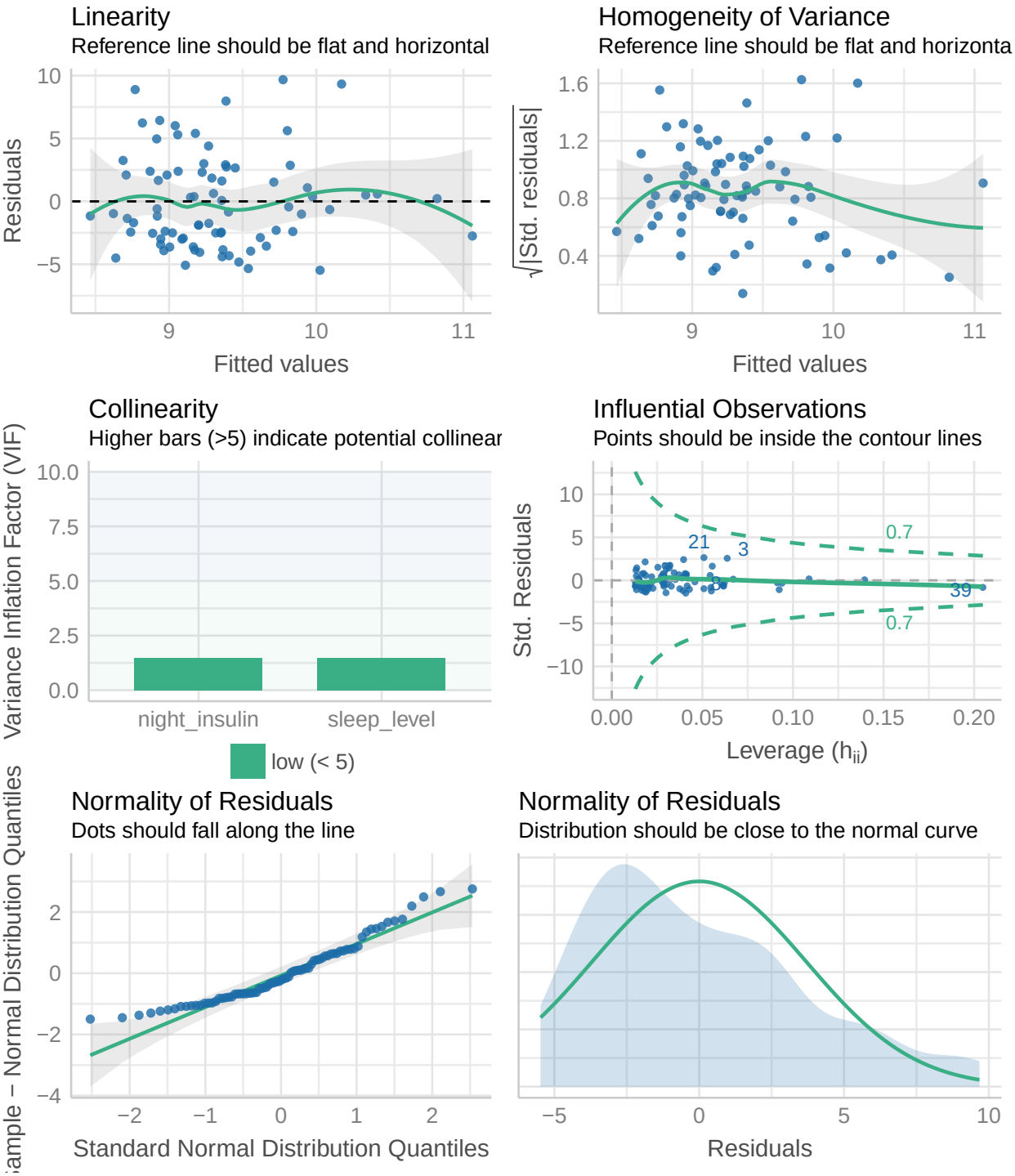


Figure 12: Diagnostics of linear regression on with food group

8.4.2 Group 2: Taking Insulin without Food at Night

For this group, the null hypothesis is mostly similar to the previous group. The difference is that in this case we are looking at nights in which the individual has taken insulin without food consumption. It is assumed that there is no correlation between blood sugar levels in the morning

with both blood sugar levels and amount of insulin taken at night. The results are shown in table 10. In regards to blood sugar levels the night before, there is no statistical significance in the variable. There is a slight negative correlation in that a unit increase in nighttime blood sugar level results in a 0.2 decrease in morning blood sugar level but the results are inconclusive. However, we see a slight statistical significance in taking insulin at night. The results suggest that for each unit increase in insulin taken at night, blood sugar levels in the morning increase by 0.3. Noting this statistically significant correlation, yet with results contrary to the norm, we once again have a contradiction. Of course, as we cannot be certain as to which assumption this p-value implies is incorrect, there is not much we can do in terms of conclusivity. With how small the smaple size is, the data could change very easily.

With that said, there are still some new insights to be gleamed. As mentioned, this particular group references nights in which the individual's blood sugar level is particularly uncontrolled. Insulin taken without food consumption naturally implies an already high blood sugar level. It is therefore not a stretch to correlate insulin intake with morning blood sugar levels as insulin was taken in a conscious effort to adjust the values. Figure 13 shows the slight positive correlation between the two variables. The model diagnosis (See Figure 14) shows that linearity and homogeneity leaves more to be desired, suggesting that this ultimately is not the best fit for the data. A model card can be found in the appendix.

Table 10: Linear Regression on Blood Sugar Levels in the Morning on Nights with Insulin Intake Only

	Estimate	**Std Error**	**t value**	**p-value**
Intercept	9.80	1.70	5.6	4.0e-07
Insulin Level at Night	-0.16	0.12	-1.4	1.8e-01
Insulin Took at Night	0.34	0.16	2.2	3.1e-02

Note:
DV: Blood sugar levels when awake
IV: Insulin level at night
IV: Amount of insulin took at night

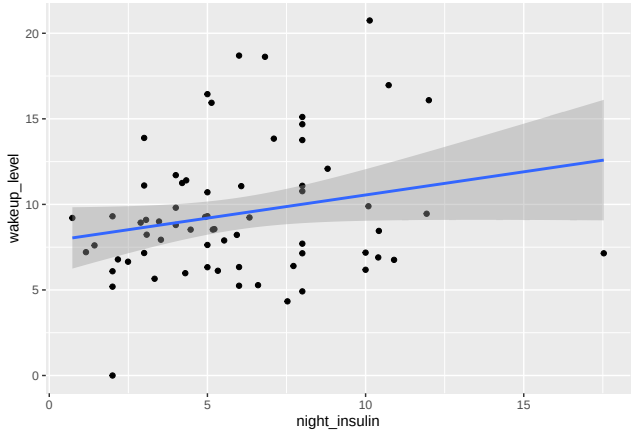


Figure 13: Plotting night insulin amount with morning blood sugar levels

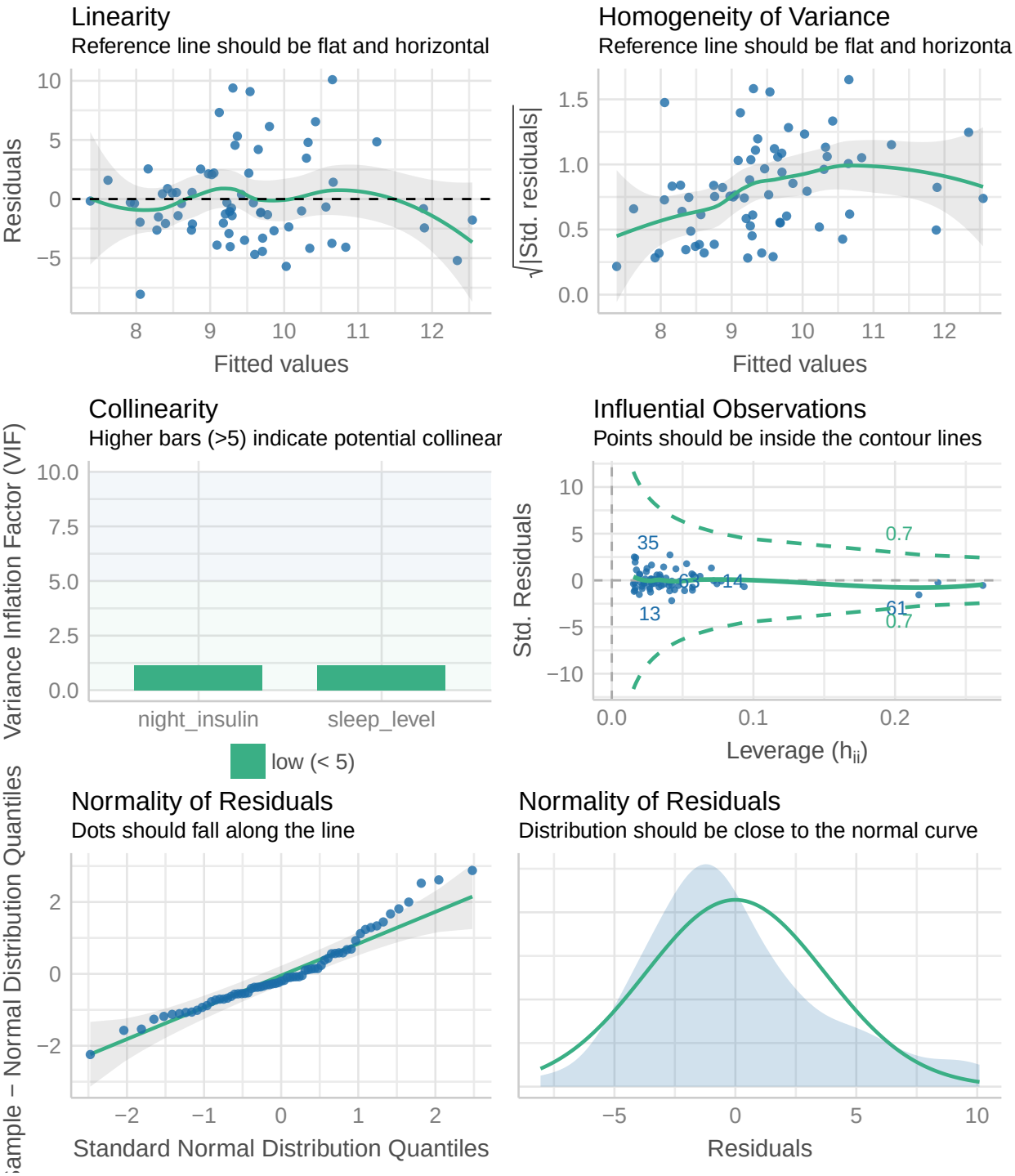


Figure 14: Diagnostics of linear regression on with food group

9 Discussion

9.1 Looking Back at Results - Finding Opportunities in Difficult Data

With so many regression models done, and none of them really fitting the data, what can be gleamed from the results? Despite all the flaws in the data, there is one linear variable that has shown high statistical significance and good model fit. That variable is time of day. If you refer back to table 7, you can see that time_of_day_coded has a positive correlation with blood sugar levels. This means for for every unit increase in time of day, the individual's blood sugar level increases by 1.659. In other words, the individual's blood sugar levels in general rises in the afternoon and then further rises again at night.

Noting that all data points to the individual being an uncontrolled diabetic with the tendency for her blood sugar levels to trend high, several assumptions can be made about the data:

1. There will be a significant amount of outliers where insulin is taken without any carb consumption (See Figure 15).
2. There will be occurrences in which insulin taken is insufficient in counteracting the amount of carbs consumed, leading to a continued high state of blood sugar levels
3. On average, the individual will be taking a lot more insulin than a controlled diabetic, if all other factors are constant

As such, the more reliable data points are those at night, when there would be no active attempts at taking insulin or eating carbs. However, that also means that we eliminate nearly all of our available factors when the time period we examine is literally a period of inaction. Therefore, the only direction is a before and after comparison - blood sugar levels before sleep, and blood sugar levels in the morning. A detailed breakdown of table 8 has already been covered above. The problem, as I soon realized, was that the uncontrolled diabetic outliers were simply too much interference on the scenarios in which the individual was in control of her blood sugar.

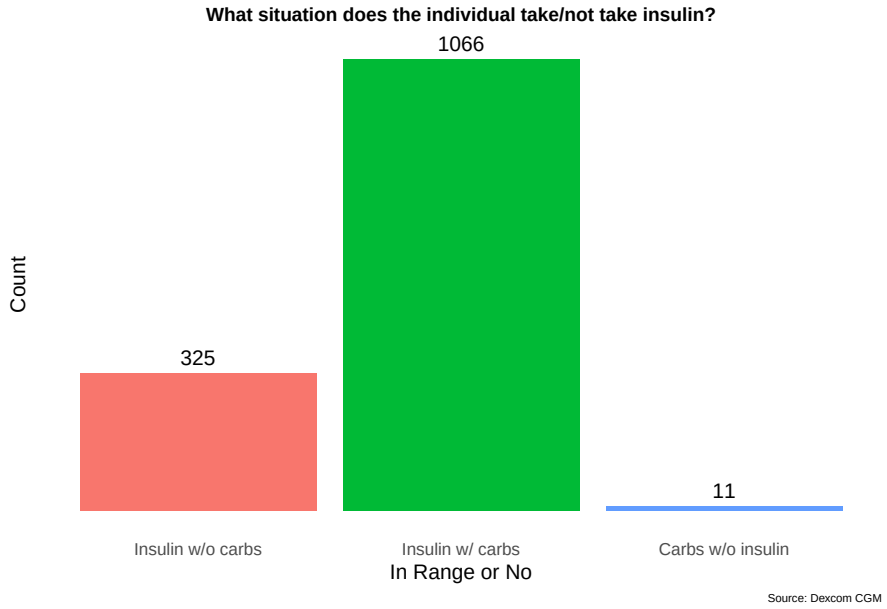


Figure 15: 23% of times insulin was taken without carbs

Hence I decided to split the data into 2 groups: the first group consists of data points in which the individual took insulin after eating carbs between the hours of midnight to 7am; the second group consists of data points in which the individual took insulin between the hours of midnight to 7am without eating food, hoping to cover outlier and correctional insulin cases. Ultimately, there were no significant variables in group 1 (results shown in table 9), whereas insulin intake showed slight significance in group 2 (results shown in table 10). The non-significance in blood sugar levels at night in correlation to blood sugar levels in the morning is unsurprising. There should not be any. There are way too many factors to be able to predict morning blood sugar levels from blood sugar levels 7 hours before. The confounding nature of factors influencing blood sugar levels mean that for group 1, when insulin is taken in response to food, that the model simply does not fit all the variations from simply the influence different types of carbs would have on blood sugar levels. Only when we strip the data points of carb consumption, as in the case of group 2, can we see some slight significant correlation. Blood sugar levels in the morning increased when insulin was taken at night without carbs. With no carbs in the system, blood sugar levels should drop in relation to insulin intake. Since we are not seeing the obvious result, and knowing that the individual spends more time with high blood sugar levels, this implies that either the insulin profile is insufficient in counteracting the natural sugars the blood releases, or that the insulin taken was insufficient in dropping the individual's already high blood sugar levels. A glance at the distribution of before sleep blood sugar levels in group 2 shows that most days the individual goes to sleep with a high blood sugar (>= 11) (See Figure 16). However, if the former 2 hypotheses are false, then this is a clear indicator of the presence of the dawn phenomenon, in which no matter what the individual's blood sugar level will rise throughout the night into the morning. Unfortunately, the limitations of data available is preventing any concrete conclusions.

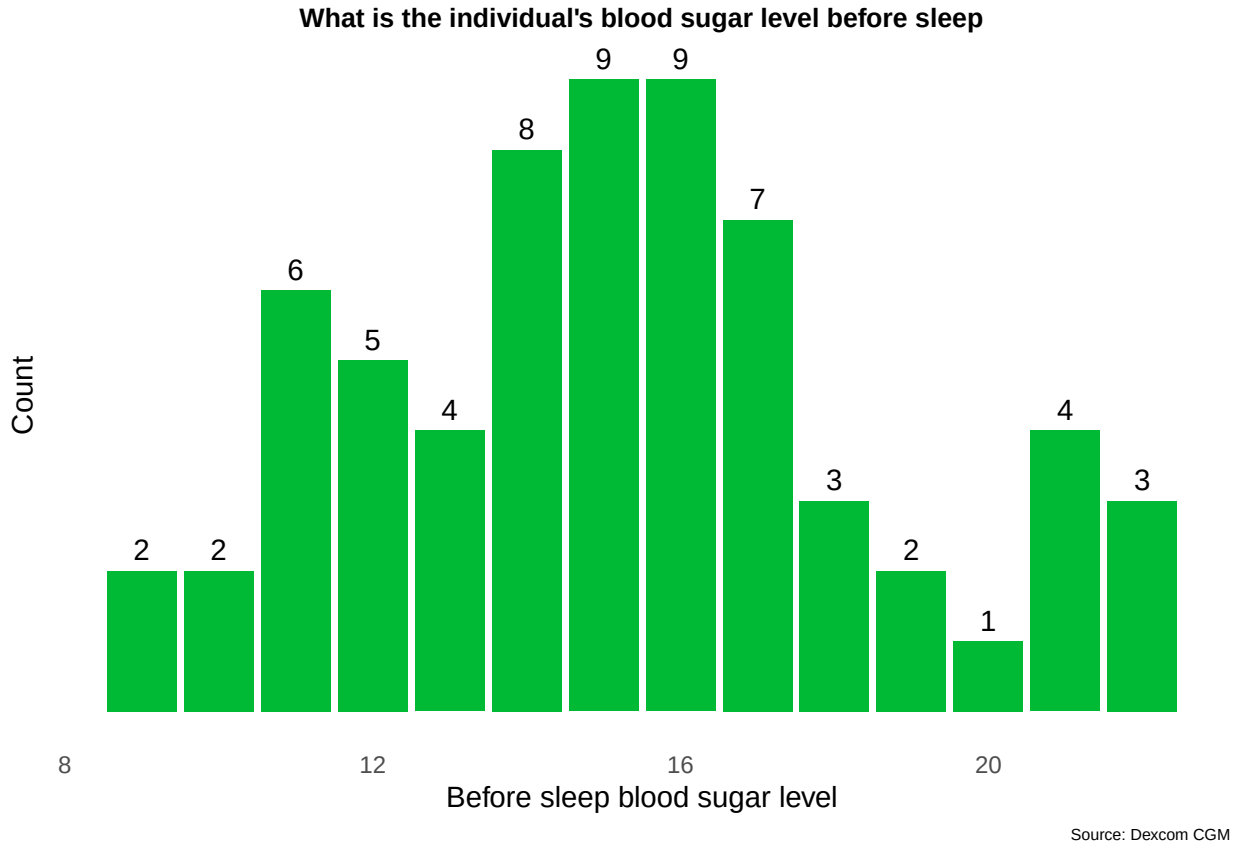


Figure 16: Only 4 out of 67 nights were in range

9.2 Causality, or the lack thereof

At this point, it is unfortunate but I must declare that causality is beyond the scope of this project. Fundamentally, this is a dataset of one person, albeit with a long period of time. This takes away the option for a control-intervention style of approach. As for splitting the data, forgetting the significant reduction of sample size (group 1 had 79 rows, group 2 had 67), the problem is that any split has to make too many assumptions for the result to be credible. We would have to assume that all carbs interact with the individual the same way. We would also have to assume that insulin affects the body at the same rate at all times of day. Considering that the pump changes sites every 3 days, with differing insulin effectiveness each time, the data I have simply cannot support any claims of causality.

9.3 Limitations

9.3.1 Conclusiveness

Ultimately, the flaws of this study comes down to the reality that the factors affecting diabetes are as uncertain as they are numerous. DiaTribe, the online publication of the diatribe foundation, listed 42 factors affecting diabetes in 2018 (Brown 2018). From exercise to dietary habits, no diabetic has the exact same diabetes experience. Moreover, due to privacy limitations, much of the personal information such as exercise habits and weight fluctuations have been withheld from this

study. In addition, as data was only collected through the Dexcom CGM, the only food related data available is carb numbers. With sugars, starches, and fiber each impacting blood sugar levels differently, it is impossible to make conclusions about the individual's dietary habit or even properly predict blood sugar levels from food intake (Centers for Disease Control and Prevention 2019b). As such, this study does not attempt to do so and instead opts to select periods with and without food intake as the comparison.

9.3.2 Representation

At the end of the day, this study only covers data on 1 patient. When looking at clinical trials, often we see the push for randomized controlled trials with a representative sample. It is not erroneous to claim that RCTs are the "current gold-standard primary study design for the determination of the efficacy and safety of medical interventions" (Kennedy-Martin et al. 2015). A single study simple cannot lead to any meaningful large scale conclusions. This study does not claim to do so, and its goal should not be taken as such. Much like there is merit in a well-designed RCT to show the average treatment effects, you would also need outlier data to see how treatments influence edge cases. In diabetes research while much focus has been designed around controlled diabetics with well-monitored diets, less has been tested on uncontrolled diabetes, especially if they are type 1 with insulin resistance. It is this study's hope to provide one piece of a larger investigation in the lives of uncontrolled diabetics. Managing diabetes is a lifestyle change that honestly some people simply cannot afford, whether it be due to time or money. As such, I believe having more data on uncontrolled diabetics and building research around helping them control their diabetes and isolating key variables in influencing their blood sugar levels is key to help mitigate the costs in managing diabetes.

9.4 Lessons Learned

9.4.1 The Project

Working with diabetes data was more complicated than anticipated. The regression results showed how important and difficult the data collection and proper data cleaning methods are in impacting proper predictions. Newer methods of aggregating the data beyond by hour must be brainstormed to take the project further. This is in part because of how the CGM reads data and the reality that insulin is taken when blood sugar is high, and stays in the blood stream in effect for up to 4 hours. Without significantly more data on how insulin interacts with the individual's blood sugar levels, it is difficult to predict blood sugar levels from the current dataset. That said, the alternative that this project tried is also lacking. Focusing on when the individual is asleep runs into the uncertainty known as the dawn phenomenon. In an attempt to predict blood sugar levels in the morning based on nightly blood sugar levels, insulin intake, and carb consumption, the ultimate conclusion that can be derived from all the models is that the individual's blood sugar level somehow rises at night no matter if she has eaten carbs and taken insulin or not. This opens up opportunitites for further research on the dawn phenomenon and allows for better insights in how to manage said phenomenon for uncontrolled diabetics.

9.4.2 The World

Indeed, what the data and this project ultimately reveals is the complexity behind diabetes management. Predicting blood sugar levels and utilizing algorithms are only a part of the solution. Keep in mind that the individual, throughout the entire 9 months of this study, was using a combination of the Dexcom CGM and the tandem pump, with Basal-IQ technology. This means that the device itself already predicts the individual's blood sugar levels, suspends insulin delivery when necessary, and calculates the amount of insulin required based on carb intake, a calculation that is in part alogrithmically based and in another determined through discussions with health professionals. Even with all this the difficulties and the multitude of moving factors when managing diabetes is made apparent from how non-linear the data points intersect.

9.5 So what was the point of this study?

The point of this study therefore is to be a drop in the ocean. A tiny scrap of data in the wide study of diabetes. Certainly there is no detailed carb breakdown, nor does the individual even have consistent sleeping habits. There are even situations in which the individual has forgotten to take insulin. But that is what makes this data real, and true to life. Diabetes is a lifelong illness, one that is exceedingly taxing physically and mentally to control, and one that gets easier the more resources an individual has. What this study has failed to demonstrate, that there lacks a steady correlation between blood sugar levels before and after sleep can hopefully be the foundation of some larger research. The increase in blood sugar levels as the day progresses is also another avenue in which more research can be done.

Personally however, I was satisfied from what I learned worked and did not work from the models ran. The failures of the data reflected the reality that my fiancee needed to recalibrate her pump's basal and carb ratio, as it became obvious that even with insulin intake at night her blood sugar still rose. The positive correlation of blood sugar levels with the time of day showed room for improvement in controlling her diabetes and calls for increased vigilance as the day progresses. This is a rare occurrence in which I have explanations for every outlier and contradiction that gets thrown at me through the data, and ultimately the inability to find a well fitting model and the presence of significant outliers is telling.

9.6 Moving Forward

This study is not over. With the glaring flaws of the dataset there are opportunities both for me and for anyone who wants to expand on this research. In terms of data collection, a new set of data must be gathered. While the flaws of CGM generated data, that is in which we will always need to aggregate by time and insulin will always be taken when blood sugar levels are higher mean that predictions of blood sugar levels after certain amount of hours is exceedingly more difficult than imagined, with enough data it is possible to split off the data so that we are seeing segments of a day in which no other meals are had. That however, would require significantly more samples.

In addition, this study shows the painful importance of addressing as many influential factors as possible in the data collection process. Proper dietary tracking, in which not only the amount of carbs are logged, but also the distribution of sugars, starches, and fibre is essential in getting a proper model. Likewise, tracking exercise each day is essential for better analyses. With the

inclusion of exercise for example, it would become possible to see if exercise as a treatment improves glycemic control.

For the public, this dataset offers a snapshot of the life of an uncontrolled diabetic. It serves as an additional dataset to add onto treatments designed to focus on the average. What would happen when applied to data in which the blood sugar level only rises throughout the day? Is the insulin intake of this particular dataset above average? If so, is it possible to make assumptions of the individual's insulin resistance? These are questions that cannot be answered within this study due to a lack of knowledge about the field of diabetes research at large, but will nonetheless add onto those who can make use of this data. My sincere hope is that the failures of this study can help assist the development of better research for improving the lives of uncontrolled diabetics. As more data gets collected, the project's GitHub repository will be updated accodingly. For suggestions on improving data collection processes or data requests, please contact Jeremy Chu at jeremychuj@gmail.com.

10 References

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Brown, Adam. 2018. "42 Factors That Affect Blood Glucose?! A Surprising Update." https://diatribe.org/42factors.

Centers for Disease Control and Prevention. 2019a. "Carb Counting." https://www.cdc.gov/diabetes/managing/eat-well/diabetes-and-carbohydrates.html#:~:text=On%20average%2C%20people%20with%20diabetes,225%20carb%20grams%20a%20day.

———. 2019b. "Carb Counting." https://www.cdc.gov/diabetes/managing/eat-well/diabetes-and-carbohydrates.html.

D'Arrigo, Terri. 2015. "How to Count Carbs." https://www.webmd.com/diabetes/features/diabetes-counting-carbs#:~:text=The%20Recommended%20Daily%20Allowance%20(RDA,grams%20per%20meal%20for%20women.

Dexcom. 2021. "What Is Continuous Glucose Monitoring (Cgm)?" https://www.dexcom.com/en-CA/what-cgm.

Diabetes Care Community. 2021. "Differences Between Type 1 Diabetes and Type 2 Diabetes." https://www.diabetescarecommunity.ca/diabetes-overview/differences-between-type-1-diabetes-and-type-2-diabetes/?gclid=Cj0KCQjwmIuDBhDXARIsAFITC_7-_kvWgylcWlAApAQ1Tk6L2Ck6QPHR1K0rwcB.

Diabetes UK. 2019. "Blood Sugar Converter." https://www.diabetes.co.uk/blood-sugar-converter.html#:~:text=mmol%2FL%20gives%20the%20molarity,this%20case%20milligrams%20per%20decilitre.

Fabien Dubosson, Stefano Bromuri, Jean-Eudes Ranvier. 2018. "The Open D1namo Dataset: A Multi-Modal Dataset for Research on Non-Invasive Type 1 Diabetes Management." *Informatics in Medicine Unlocked* 13. https://doi.org/https://doi.org/10.1016/j.imu.2018.09.003.

Francesca Porcellati, Geremia B. Bolli, Paola Lucidi, and Carmine G. Fanelli. 2013. "Thirty Years of Research on the Dawn Phenomenon: Lessons to Optimize Blood Glucose Control in Diabetes." https://care.diabetesjournals.org/content/36/12/3860.

Giles, Gary. 2020. "Overview of the Types of Insulin." https://www.verywellhealth.com/basal-and-bolus-insulin-3289548#:~:text=Basal%20insulin%20(sometimes%20called%20background, blood%20glucose%20that%20immediately%20follows.

Heather Hall, Alessandra Breschi, Dalia Perelman. 2018. "Glucotypes Reveal New Patterns of Glucose Dysregulation." *PLOS Biology*. https://doi.org/https://doi.org/10.1371/journal.pbio.2005143.

Jette Bertelsen, Claus Thomsen, Christian Christiansen, and Kjeld Hermansen. 1993. "Effect of Meal Frequency on Blood Glucose, Insulin, and Free Fatty Acids in Niddm Subjects." *Diabetes Care* 16 (1). https://doi.org/https://doi.org/10.2337/diacare.16.1.4.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Kennedy-Martin, Tess, Sarah Curtis, Douglas Faries, Susan Robinson, and Joseph Johnston. 2015. "A Literature Review on the Representativeness of Randomized Controlled Trial Samples and Implications for the External Validity of Trial Results." *Trials* 16 (495). https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-015-1023-4.

Labtests Online. 2021. "Hemoglobin A1c." https://labtestsonline.org/tests/hemoglobin-a1c.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. "Assessment, Testing and Comparison of Statistical Models Using R." *Journal of Open Source Software* 6 (59): 3112. https://doi.org/10.31234/osf.io/vtq8f.

Margaret Mitchell, Andrew Zaldivar, Simone Wu. 2019. "Model Cards for Model Reporting." *Association for Computing Machinery*. https://doi.org/https://doi.org/10.1145/3287560.3287596.

Medical News Today. 2019. "How to Manage Diabetes with Basal-Bolus Insulin Therapy." https://www.medicalnewstoday.com/articles/316616#summary.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Nguyen Thanh Ha, Le Thi Thu Ha, Nguyen Thi Phuong. 2019. "How Dietary Intake of Type 2 Diabetes Mellitus Outpatients Affects Their Fasting Blood Glucose Levels?" *AIMS Public Health* 6 (4). https://doi.org/10.3934/publichealth.2019.4.424.

Ooms, Jeroen. 2021. *Magick: Advanced Graphics and Image-Processing in R*. https://CRAN.R-project.org/package=magick.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tandem. 2021. "T:Slim X2 Insulin Pump." https://www.tandemdiabetes.com/en-ca/products/t-slim-x2-insulin-pump.

Teri B. O'Neal, Euil E. Luther. 2020. "Dawn Phenomenon." https://www.ncbi.nlm.nih.gov/books/NBK430893/.

Timnit Gebru, Briana Vecchione, Jamie Morgenstern. 2020. "Datasheets for Datasets." *Cornell University*. arXiv:1803.09010.

Tsalikian, Eva. 2005. "Impact of Exercise on Overnight Glycemic Control in Children with Type 1 Diabetes." *The Diabetes Research in Children Network (DirecNet) Study Group* 147 (4). https://doi.org/10.1016/j.jpeds.2005.04.065.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman; Hall/CRC. https://github.com/rstudio/bookdown.

Yutani, Hiroaki. 2020. *Gghighlight: Highlight Lines and Points in 'Ggplot2'*. https://CRAN.R-project.org/package=gghighlight.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

11 Appendix

11.1 Datasource Datasheet

The following datasource datasheet is created following the template provided by Timnit Gebru (2020).

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The 9 months Type 1 Diabetes dataset was created specifically for personal reasons. It was made so that the creator could examine his fiancee's diabetes data and conduct research to derive personal and academic value.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The original creator of the dataset is Jeremy Chu, a Master of Information student at the University of Toronto at the time of the dataset's release in 2021

What support was needed to make this dataset?

Except for consent from the subject of the dataset, no other support was needed for the dataset's creation

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings;

A single person's blood sugar level readings, food consumption, and insulin intake for 9 months

Is there a label or target associated with each instance?

Data is grouped monthly. Cleaned dataset is aggregated by hour.

Is any information missing from individual instances?

The months January - March 2020 are missing. Dataset goes from December 2019 and jumps to April 2020.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

There are obvious date times for each month's data.

Are there recommended data splits (e.g., training, development/validation, testing)?

No recommended data split. If desired, choose specific months to use as testing months..

Are there any errors, sources of noise, or redundancies in the dataset?

No errors or redundancies. For the cleaned dataset, be aware that data is aggregated by hour. For the raw dataset, there are no errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

Data is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

Data is an individual's diabetes data. Consent has been prior obtained, any further details must be requested.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Does the dataset relate to people?

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)?

Yes.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No direct identifiable information present in the dataset. The study however, speifies the relationship between the individual and the researcher.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

As sensitive as health data publicly disclosed after obtaining consent.

Any other comments?

Collection Process

How was the data associated with each instance acquired?

The data is collected from the Dexcom G6 CGM. One to one translated from the machine to csv.

Over what timeframe was the data collected? D

The data was continously collected by the CGM. Every momth the data is available to be downloaded directly from the machine to a computer. The data collection malfunctioned during the months January - March of 2020, therefore only 9 months of data remain instead of 1 year.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

Dexcom G6 CGM sensor.

What was the resource cost of collecting the data?

No cost.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Sample of the individual's diabetes data.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Just the individual whose data is collected. Obtaining consent was the only requirement.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

Does the dataset relate to people?

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Obtained permission to take directly from CGM.

Were the individuals in question notified about the data collection?

Yes. Individual must give me their CGM for data collection.

Did the individuals in question consent to the collection and use of their data?

Yes. Individual must give me their CGM for data collection.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Yes. Individual has access to the GitHub repository.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Yes. Individual has access to the GitHub repository.

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Yes. The data has been aggregated by hour, with the CGM and insulin intake data combined. Any personal identifiable variables have been removed from both the raw and cleaned versions.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes. In the same GitHub repository.

Is the software used to preprocess/clean/label the instances available?

Any other comments?

Uses

Has the dataset been used for any tasks already?

Yes. The data has been used for a personal research already. The study can be found in the same GitHub repository.

Is there a repository that links to any or all papers or systems that use the dataset?

https://github.com/JeremyJChu/diabetes

What (other) tasks could the dataset be used for?

Combinded with larger repositories of datasets to compile a more comprehensive diabetes dataset. Applicable for meal frequency and/or overnight blood sugar control research.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Know that the cleaned data aggregates by hour, if that is undesired, please turn to the raw dataset.

Are there tasks for which the dataset should not be used?

No.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

No specific distribution. Publicly available on GitHub.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

GitHub.

When will the dataset be distributed?

Sometime in April 2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Creative Commons 4.0. Free to distribute, reuse given proper credits.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

Any other comments?

Maintenance

Who is supporting/hosting/maintaining the dataset?

Jeremy Chu, a Masters of Information student at the University of Toronto at the time of dataset distribution.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Please email jeremychuj@gmail.com for any and all inquiries.

Is there an erratum?

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes. Raw data will be uploaded annually. There will be a changelog on GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data.

No limits on data retenion. Have obtained consent for the data to remain on GitHub until told otherwise.

Will older versions of the dataset continue to be supported/ hosted/maintained?

Yes in the sense that old datasets will not be removed. More new raw datasets will continuously be added.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

No mechanism for adding or building onto datasets.

Any other comments?

While there is no available mechanism for users to build onto the dataset, they are free to take the data and build onto their own datasets. If they would like to include a link to their project/datasets on this GitHub repository, please contact Jeremy Chu @ jeremychuj@gmail.com with the link and it will be added to this project's GitHub repositor

11.2 Model Card

The following model card is created following the template provided by Margaret Mitchell (2019).

Model Card - Predicting Blood Sugar Levels Overnight

Model Details
• Standard logistic and linear regression models

Intended Use
• Intended to determine fit for specific diabetes data relating to an uncontrolled diabetic for the duration of 9 months.
• Initially planned for prediction purposes, ultimately became an evaluation on what factors would fit regression models.

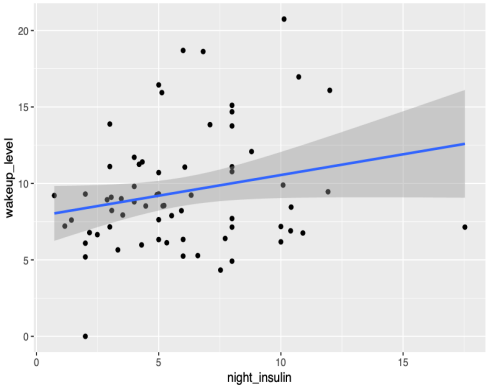
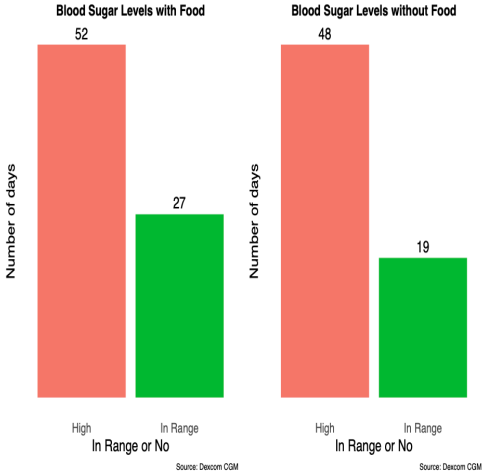


Figure 13: Plotting night insulin amount with morning blood sugar levels

Factors
• Based on available factors in the provided dataset. The dependent variable varies between blood sugar levels and insulin intale.
• Independent variables include carb intake (no specific breakdown on type of carbs, just measured in grams), blood sugar levels (in mmol/L), insulin intake (in mmol/L), time of day



Metrics
• Evaluation metrics included linearity, homogeniety, collinearity, and normality.
• Together they show whether the model fits the data, and where the gaps lie.

Group 1
• Taking insulin with food at night

Group 2
• Taking insulin without food at night

Ethical Considerations
• Data is based on a real individual's health data. Consent has been obtained for free use of the data. Any personal identifiers have been removed and any modelling will not reflect on the individual.

Caveats and Recommendations
• Data aggregates by hour, does not capture minute details.
• Data also captures insulin intake the moment blood sugar is high. Keep that in mind when creating regression analyses. There will always be a positive correlation between insulin iintake and blood sugar levels.