Regression Analysis of C-Peptide Dependence on Age and Base Deficit

Summary

This study analyzes a dataset of children who were recently diagnosed with Type 1 diabetes to understand what factors affect how much insulin their bodies can still produce. It uses a log transformation of C-peptide concentration as a measure of remaining insulin production and examines how it relates to child's age and a measure of blood acidity called base deficit. Using regression analysis, the study explores whether older children and those who had less base deficit at diagnosis retained more natural insulin production. Using a linear regression analysis, we found that age and base deficit are both factors that are linked to C-peptide levels.

Introduction

Diabetes is a chronic disease in which the body is unable to properly regulate blood sugar. According to the World Health Organization (WHO), more than 830 million people worldwide, around one in ten people, live with diabetes (2024). It is also one of the top ten leading causes of death globally (CDC, 2023).

WHO says: "Diabetes causes blindness, kidney failure, heart attacks, stroke and lower limb amputation [...] Neither its cause nor the means to prevent it are known." (2024)

Type 1 diabetes, also called insulin-dependent diabetes, is a more serious form that usually starts in childhood. In Type 1 diabetes, the pancreas stops producing insulin, a hormone that allows sugar to move from the blood into the body's cells. Without insulin, blood sugar rises to dangerous levels, and people need insulin injections to survive.

When children are first diagnosed, not all of them have completely lost the ability to produce insulin. Children who retain some natural insulin production have easier time managing the disease. This raises an important question: What factors influence how much insulin a child's body is still able to produce at diagnosis?

For this study, we will specifically investigate the factors of age and base deficit. To answer this, we analyzed a dataset containing three measurements taken at or around the time of diagnosis:

- C-peptide level: A substance released in equal amounts to insulin. If C-peptide is present, it means the body is still producing some insulin.

- Age: According to previous research, age may be an important factor in explaining why some retain ability to produce small amounts of insulin.
- Base deficit: A measure of how acidic the blood is. Higher acidity means the child was more ill at diagnosis.

The goal of this study is to determine how age and base deficit relate to C-peptide levels, which reflect the amount of insulin the pancreas is still making. Regression analysis is used to model how these predictors influence the log-transformed C-peptide concentration. This may help to understand the factors behind residual insulin production in children.

Question:

When studying patterns of insulin-dependent diabetes mellitus in children, is there a relationship between levels of serum C-Peptide and the factors of patient age and base deficit?

Data Source:

KEEL Diabetes Data Set

The data was sourced from KEEL (Knowledge Extraction based on Evolutionary Learning), which is an open source software tool containing datasets that can be used for knowledge data discovery. The chosen dataset is designed for regression analysis, taken from LIACC's repository. The objective of this data is to investigate how factors such as patient age and base deficit (a measure of metabolic acidosis) are associated with concentration of C-peptide, in order to further understand patterns of insulin-resistant Diabetes Mellitus in children.

Data Structure:

`Age` : Patient Age, domain: [0.9,15.6]

`Deficit` : Base Deficit, measure of acidity, domain: [-29.0,-0.2]

`C-peptide` : Logarithm of C-Peptide concentration (pmol/ml), domain: [3.0,6.6]

Methods

Multiple linear regression was used to model the relationship of age and base deficit to logarithm of C-peptide concentration in patients. The Python programming language (Van Rossum and Drake 2009) was used to perform the analysis.

Imports

```
In [1]: import pandas as pd
        import numpy as np
        import requests
        import zipfile
        import os
        import io
        import seaborn as sns
        import matplotlib.pyplot as plt
        import altair as alt
        import pingouin as pg
```

Download Data into data/ directory

This code was adapted from Microsoft Copilot prompts: 'Use Python to download a data file from a download link into a directory, as a csv'

```
In [2]: url = "https://sci2s.ugr.es/keel/dataset/data/regression/diabetes.zip"

        data_dir = os.path.join("..", "data")
        os.makedirs(data_dir, exist_ok=True)

        file_path = os.path.join(data_dir, "diabetes.zip")

        response = requests.get(url)
        zip_bytes = io.BytesIO(response.content)

        with open(file_path, "wb") as f:
            f.write(response.content)
```

```
In [3]: with zipfile.ZipFile(zip_bytes, "r") as zip_ref:
            dat_files = [f for f in zip_ref.namelist() if f.endswith(".dat")]
            dat_content = zip_ref.read(dat_files[0]).decode("utf-8")
```

Data Wrangling and Cleaning

1. original file downloads as .dat, ensure it fits with a .csv format by splitting with the ','
   delimiter and stripping lines beginning with '@'
2. re-name columns since names were stripped away
3. ensure data is the correct dtype

```
In [4]: lines = dat_content.splitlines()
        data_lines = [line for line in lines if not line.startswith("@") and line.st
```

```
In [5]: rows = [line.strip().split(",") for line in data_lines]
        diabetes_df = pd.DataFrame(rows)
```

```
diabetes_df.columns = ["Age", "Deficit", "C_peptide"]
diabetes_df = diabetes_df.astype(float)
```

Note: dtypes were set to float manually as they were downloaded as dtype: object

In [6]:
```
csv_path = os.path.join("..", "data", "diabetes.csv")
diabetes_df.to_csv(csv_path, index=False)
```

EDA

In [7]:
```
diabetes_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43 entries, 0 to 42
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Age        43 non-null     float64
 1   Deficit    43 non-null     float64
 2   C_peptide  43 non-null     float64
dtypes: float64(3)
memory usage: 1.1 KB
```

Figure 1: Dataset summary

No null values, data types were set to float earlier

In [8]:
```
diabetes_df.describe()
```

Out[8]:

	Age	Deficit	C_peptide
count	43.000000	43.000000	43.000000
mean	9.032558	-8.148837	4.746512
std	4.022539	7.123080	0.720565
min	0.900000	-29.000000	3.000000
25%	5.500000	-12.700000	4.450000
50%	10.400000	-7.800000	4.900000
75%	11.850000	-2.000000	5.100000
max	15.600000	-0.200000	6.600000

Figure 2: Dataset descriptive statistics

The mean patient age is 9 years old, with the maximum age being 15.6 and the minimum being 0.9. The C_peptide concentration average is about 4.75, mean Base Deficit value is -8.15 mEq/L. The standard deviation of Base Deficit is quite high, at about 7.1 mEq/L, while the standard deviations of the other two variables are within a reasonable range for their domains. This high variance could be due to the small size of the dataset.

```
In [9]:  sns.histplot(diabetes_df["C_peptide"], kde=True)
         plt.title("Distribution of C_peptide")
         plt.show()
```
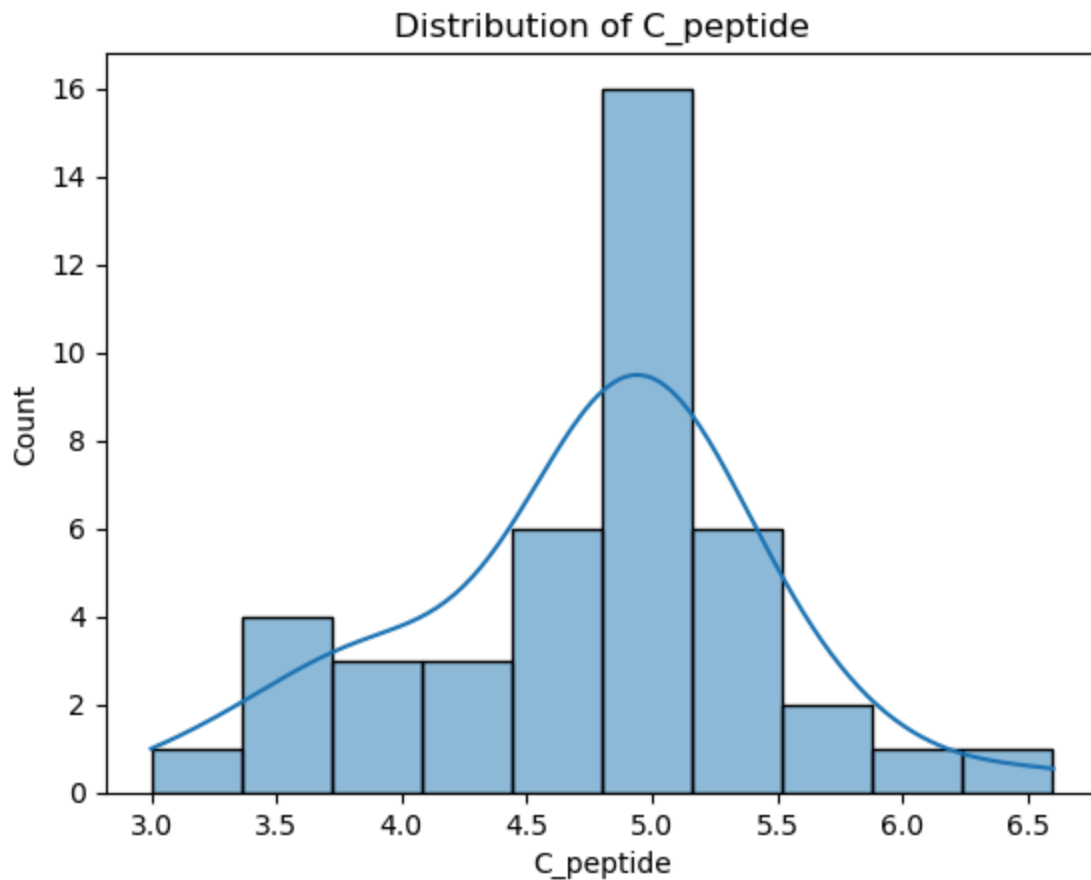

Figure 3: Distribution of target variable, C_peptide

```
In [10]:  alt.Chart(diabetes_df).mark_circle().encode(
              x=alt.X(alt.repeat("column"), type="quantitative"),
              y=alt.Y(alt.repeat("row"), type="quantitative")
          ).properties(
              width=150,
              height=150
          ).repeat(
              row=["Age", "Deficit", "C_peptide"],
              column=["Age", "Deficit", "C_peptide"]
          )
```
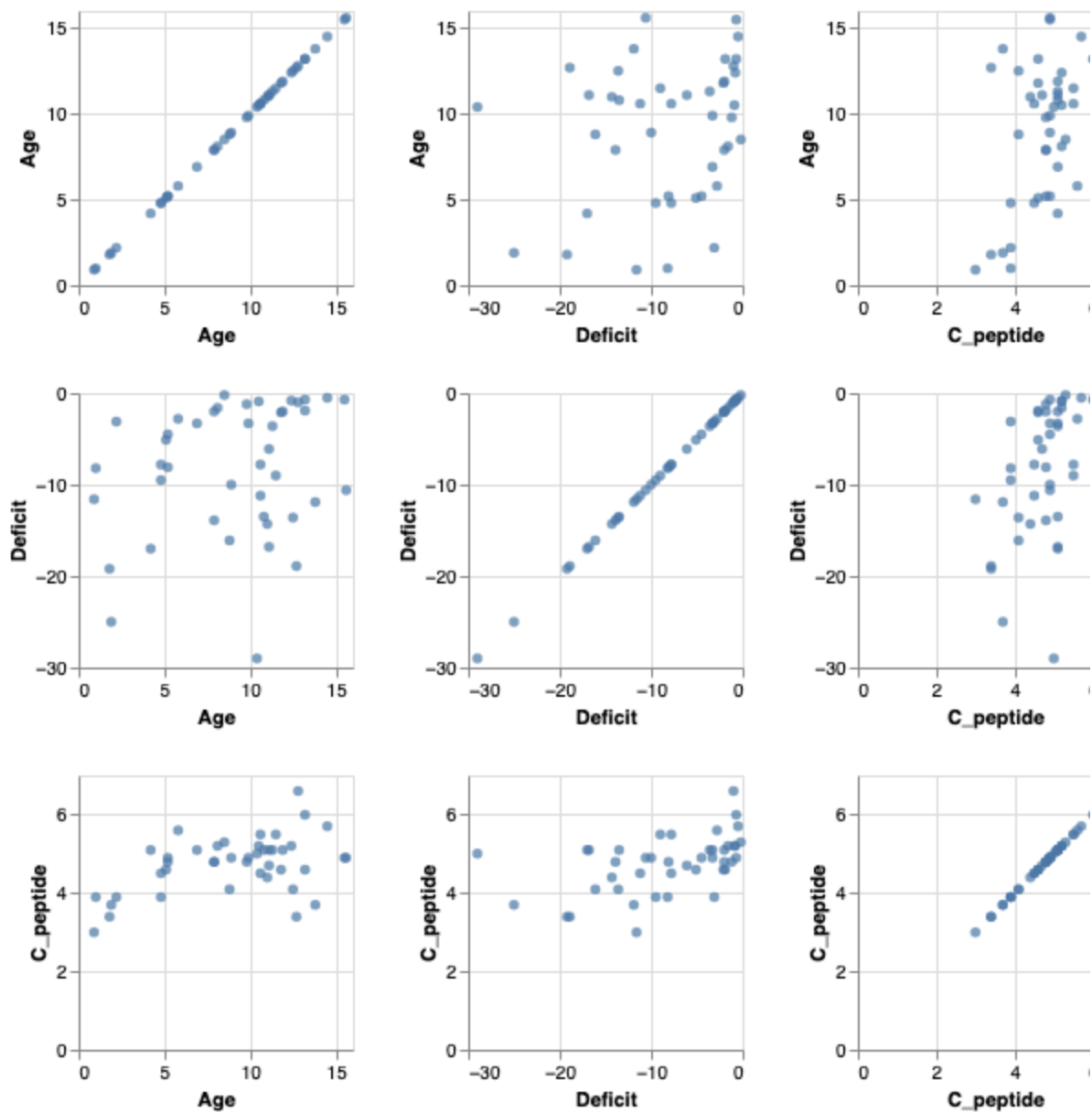
`Out[10]:`

Figure 4: Scatterplot matrix of all variables

`In [11]:`
```python
corr = diabetes_df.corr()


plt.figure(figsize=(6, 4))
sns.heatmap(corr, annot=True, cmap="coolwarm", center = 0, fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```
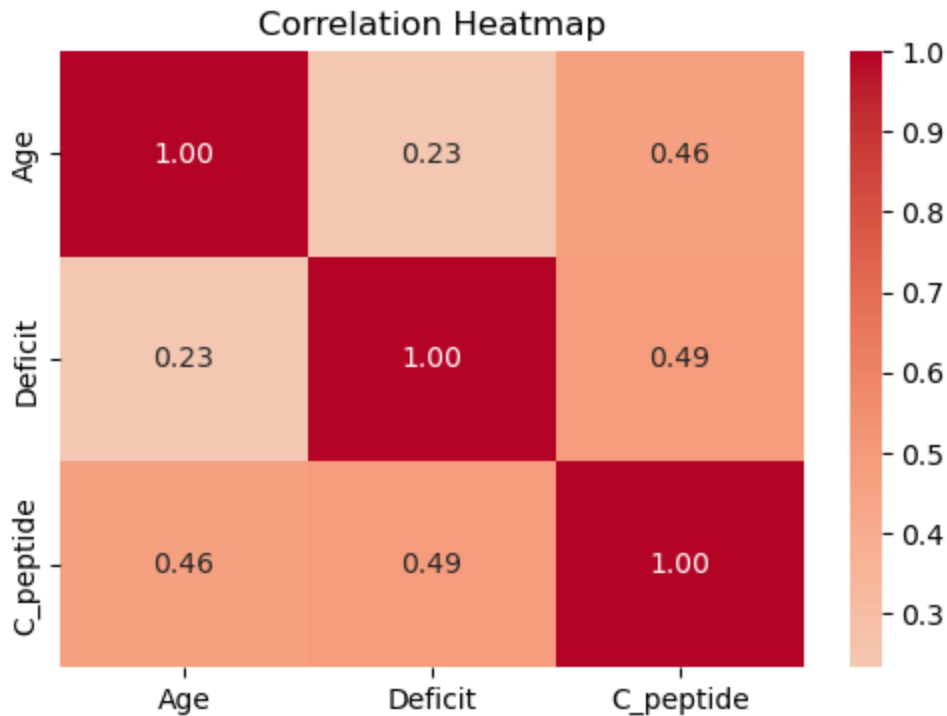
Figure 5: Correlation heatmap of all variables

EDA Summary:

Both features are positively associated with the target variable, `C_peptide`, at approximately equal magnitudes (0.46 and 0.49). The distribution of the target variable is approximately normal, with most values clustered around 5.0. There does seem to be one or two outliers in scatterplots comparing the relationship between `Age` and `Deficit`, but as these are both non-target features this can be ignored for the sake of this analysis. There are no obvious outliers in the relationship between the target and its predictors.

Modelling

We use linear regression to model the relationship between `C_peptide` and covariates `Age` and `Deficit`.

```
In [12]:  X = diabetes_df.drop(columns=["C_peptide"])
          y = diabetes_df["C_peptide"]
```

```
In [13]:  model = pg.linear_regression(X, y)
          model
```

	names	coef	se	T	pval	r2	adj_r2	CI[2.5%
0	Intercept	4.479400	0.270929	16.533455	1.765281e-19	0.368452	0.336874	3.93183
1	Age	0.066314	0.023143	2.865386	6.608034e-03	0.368452	0.336874	0.01954
2	Deficit	0.040726	0.013069	3.116167	3.384871e-03	0.368452	0.336874	0.01431

Figure 6: Linear Regression Intercept and Coefficients

Model Diagnostics

In [14]:
```
resid = model.residuals_
```

To check for Normality of residuals, we employ the Shapiro-Wilk test and Q-Q plot.

In [15]:
```
pg.normality(resid, method = "shapiro")
```

Out[15]:

	W	pval	normal
0	0.987989	0.927343	True

Figure 7: Shapiro-Wilk Test Results

In [16]:
```
pg.qqplot(resid)
```

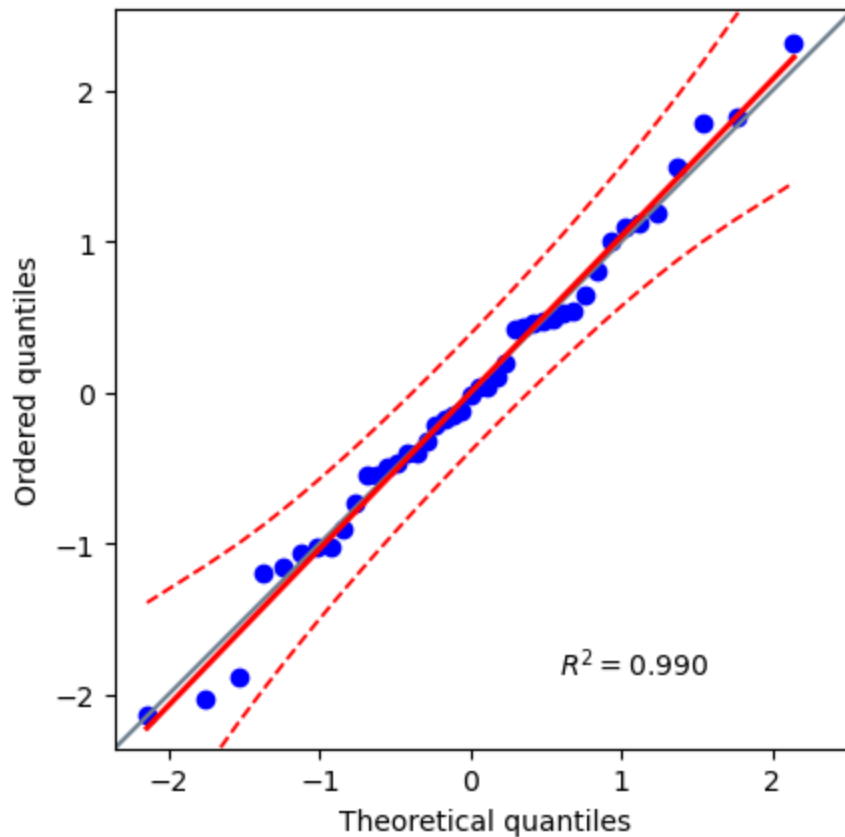Out[16]: `<Axes: xlabel='Theoretical quantiles', ylabel='Ordered quantiles'>`

Figure 8: Q-Q Plot of Residuals

To test for equal variance, we plot the residuals against the fitted values for the model.

```
In [17]:  fitted_df = pd.DataFrame({
              "observed": y,
              "fitted": y - resid,
              "residuals": resid
          })

          alt.Chart(fitted_df).mark_point().encode(
              x=alt.X(
                  "fitted",
                  scale=alt.Scale(zero=False)
              ).title("Fitted Values"),
              y=alt.Y(
                  "residuals",
                  scale=alt.Scale(zero=False)
              ).title("Residuals")
          )
```
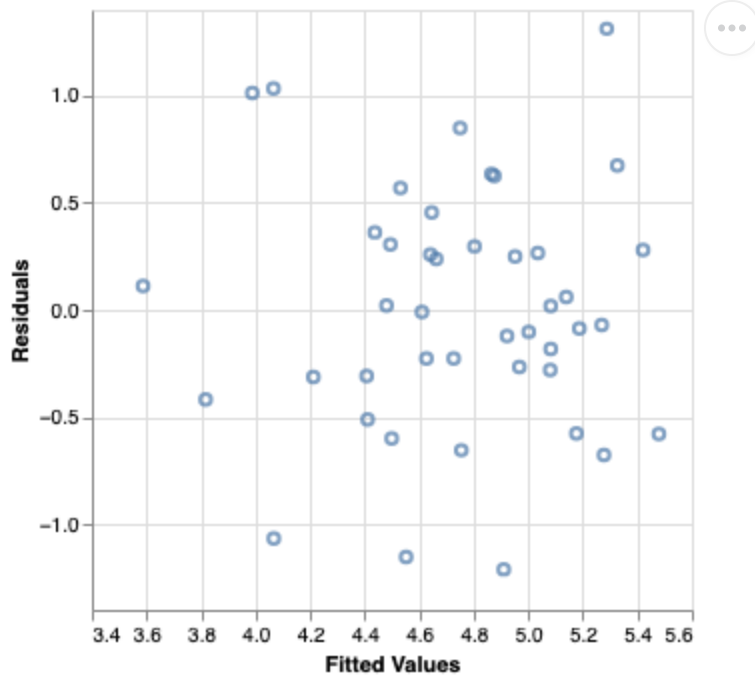
Out[17]:

Figure 9: Residuals vs Fitted Values

DISCUSSION

Summary of Findings

The objective of this analysis was to determine if there is a linear relationship between serum C-peptide (log-transformed concentration, pmol/ml) and the factors of patient age (Age, years) and base deficit (Deficit, mEq/L) in children with insulin-dependent diabetes mellitus.

A multiple linear regression model yielded the following key findings (Figure 6):

Both predictor variables, Age ($p = 0.0066$) and Deficit ($p = 0.0034$), were found to be statistically significant in predicting the C-peptide level (using a typical $\alpha = 0.05$). Age has a positive coefficient ($\mathrm{coef} = 0.066$), indicating that for every one-year increase in age, the log C-peptide concentration is expected to increase by 0.066 units, holding Deficit constant. Base deficit has a positive coefficient ($\mathrm{coef} = 0.041$), meaning that for every one $\mathrm{mEq/L}$ increase in Deficit, the log C-peptide concentration is expected to increase by 0.041 units, holding Age constant. (Note: Since Deficit values are negative, an increase (closer to 0) signifies less severe metabolic acidosis.)

The model achieved an adjusted R^2 of 0.337. This means that approximately 33.7% of the variance in the log C-peptide concentration is explained by the patient's age and base deficit.

When examining model assumptions, the diagnostic tests (Shapiro-Wilk $p = 0.927$ and Q-Q plot/Residuals vs. Fitted plot, Figures 7-9) shows that the key linear regression assumptions of normally distributed residuals and equal variance are reasonably met. While it is more difficult to confirm whether the observations are independent, we will assume so here for the purposes of out analysis.

Expected Findings

The results showing that Age and Base Deficit are both positive predictors of C-peptide levels generally match what we expect in children with Type 1 Diabetes.

Since C-peptide reflects how much insulin the body can still produce, the positive relationship with age suggests that older children in this dataset tend to have more remaining beta-cell function at the time of measurement. This is reasonable because some older children may lose beta-cells more slowly, or may still be in an early stage of the disease when some insulin production is preserved.

The positive effect of base deficit here means that higher C-peptide levels are linked to a less severe Deficit (meaning a less severe case of DKA). This fits existing medical understanding: children who still produce some insulin are less likely to arrive with very severe metabolic disturbances compared to those whose beta-cells are almost fully destroyed.

The one surprising part of the results is the low R² value (0.337). Even though the model is statistically significant, it explains only a portion of the differences in C-peptide levels. This suggests that other factors, such as how long the child has had symptoms, their genetic markers, autoantibody levels, or initial glucose and HbA1c—also play important roles and should be included in future analyses.

Impact of Findings

These results have several important implications for children with Type 1 diabetes:

One is prognostic value: Because C-peptide is a strong indicator of long-term diabetes outcomes (such as the risk of low blood sugar and future complications), this analysis suggests that a child's Age and Deficit at diagnosis can serve as easy-to-use markers for predicting how well they might manage the disease over time.

The findings also show that the onset of Type 1 diabetes is not the same for all children. Those who are a bit older and arrive with less severe metabolic problems (a less negative Deficit) appear to belong to a group that still has better remaining insulin-producing ability.

Knowing how Age and Deficit relate to C-peptide can help doctors tailor treatments more effectively. For example, children with higher predicted C-peptide levels might be

better candidates for therapies designed to protect or support the remaining beta cells, since they already have some preserved function.

Finally, the strong link between Deficit and C-peptide supports the idea that C-peptide is a reliable indicator of how severe insulin loss is at the time a child is diagnosed.

Future Questions for Further Research

The results from this model raises several important follow-up questions:

1. **Non-Linear Effects and Interactions:** Is the relationship between Age, Deficit, and C-peptide more complicated than a straight line? For example, does the effect of Deficit on C-peptide change depending on the child's age, or would adding curved (quadratic) terms improve the model?

2. **Longitudinal Data:** Since this dataset is cross-sectional, we only see one point in time. How would the predictive power of Age and Deficit change if we tracked children over several years to see how quickly their C-peptide levels decline after diagnosis?

3. **Adding More Predictors:** Considering that the current model explains only about 34% of the variation, what extra value would we gain by including other medical factors—such as genetic markers (HLA types), autoantibodies, or initial blood glucose/HbA1c levels?

4. **Effect of Transformations:** Since our current model uses a log transformation of C-peptide. How would the results change if we modeled the raw (untransformed) C-peptide values?

References

Centers for Disease Control and Prevention. (2025, September 17). FASTSTATS - leading causes of death. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Cleveland Clinic. (2025, July 22). C-peptide test: What it is, purpose, procedure & results. https://my.clevelandclinic.org/health/diagnostics/24242-c-peptide-test

KEEL Diabetes Dataset (By KEEL). (n.d.). [Dataset]. https://sci2s.ugr.es/keel/dataset.php?cod=45

Van Rossum, Guido, and Fred L. Drake. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

World Health Organization. (n.d.). Diabetes. https://www.who.int/news-room/fact
sheets/detail/diabetes