Analyzing Cardiovascular Disease with Linear Regression: OLS and WLS Approach

Cynthia Jin, Liang Yang, and Steven Hsiao

Columbia University in the City of New York

ABSTRACT

Carotid intima-media thickness (cIMT) progression is a marker for cardiovascular disease (CVD) risk, as mentioned in a previous meta-analysis performed by Willeit et al. (Willeit, 2020). However, the prior research did not include the investigation of complete demographic and clinical variations that may influence this association. Building on the foundation, we carried out a meta-regression research to discuss whether specific cofactors affect the link between cIMT progression and CVD risk. Our analysis utilized Ordinary Least Squares (OLS) followed by Weighted Least Squares (WLS) regression models to validate findings across datasets from diverse studies. We examined variations in demographic backgrounds, baseline risk factors, and clinical characteristics to understand their moderating effects. Certain significant results highlight the need to account for study heterogeneity when interpreting cIMT progression as a marker for CVD risk in clinical and public health contexts. Future studies could further evaluate the underlying mechanisms driving these effects and explore the potential benefits they may offer to the healthcare field.

Keywords: Linear Regression, Cardiovascular Disease (CVD), OLS/WLS

1 INTRODUCTION

Cardiovascular disease (CVD) has significantly impacted the human population and affected nearly half of American adults (Martin, 2024). Moreover, cardiovascular disease has consistently been the leading cause of death in the United States, and the number of deaths it causes has been steadily increasing in recent years. The annual total cost of managing CVD is estimated to be $422.3 billion between 2019 and 2020, accounting for 12% of total US health expenditures. It is predicted that the prevalence of CVD will climb from 11.3% to 15.0% for the next 30 years (Joynt, 2024).

The carotid intima-media thickness (cIMT) is defined as the level of the intimal and medial layers' combined thickness in the wall of the carotid artery. It can be measured through ultrasound imaging without any further invasive management, and it is widely recognized as an indicator of atherosclerosis, a common disease where fatty deposits, cholesterol, and other substances build up on the walls of arteries, forming plaques (Libby, 2011). These plaques narrow and stiffen the arteries, impairing blood flow and increasing the risk of critical CVD events. Because CVD is primarily caused by atherosclerosis, monitoring cIMT could be an effective strategy to estimate CVD risks in patients and improve the prevention of CVD morbidity and mortality (Ling, 2023).

Meta-analysis is a study method that combines data from multiple studies to calculate a pooled effect size, providing stronger evidence of the effect size for the mainly discussed factor compared to individual studies (Harrer, 2021). Meta-regression, additionally, identifies possible cofactors that may result in the differences between each study effect size, strengthening the result of the meta-analysis. Besides, meta-regression is generally applied in medical meta-analysis research due to the simplicity and interpretability of linear regression models and their practicality in medical research (2002, Thompson). At the same time, subgroup analysis—a method that stratifies studies into separate groups based on specific factors to compare effect sizes between these groups—is widely used for exploring factors causing differences between effect sizes. Nevertheless, they vary in deployed situations, with subgroup analysis used when the factors are categorical data and chosen with certain interests, and meta-regression used when there are multiple and either continuous or categorical data with a sufficient number of analyzed studies (10 studies per factor) (Higgins, 2011).

In the previous meta-analysis, the pooled effect size of cardiovascular disease (CVD) risk was calculated, and a

significant association between carotid intima-media thickness (cIMT) and CVD risk was identified using a regression model with a single coefficient (Willeit, 2020). However, meta-regression incorporating multiple cofactors was not conducted in that study. Although subgroup analysis was performed, it did not include all continuous and categorical data in the evaluation. Furthermore, continuous data, studied with subgroup analysis by setting thresholds to create groups in the prior research, are often considered better studied with meta-regression. Therefore, our objective is to apply meta-regression to the previous meta-analysis to investigate whether specific cofactors influence the relationship between cIMT and CVD risk. We will first use ordinary least squares (OLS) regression to identify potential cofactors and subsequently apply weighted least squares (WLS) regression, which accounts for the weight of individual studies—a key consideration in meta-regression models assuming there is heteroscedasticity between effect sizes (Harrer, 2021).

2 METHODOLOGY

2.1 Linear Regression

Linear regression is one of the most commonly used methods for decision modeling, it is a powerful model to predict results based on input variables. It can be very helpful for many complex statistical problems (Gallo, 2015).

The model can be expressed in the following form:

$$y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \varepsilon$$

Here:

- y is the dependent variable, representing the outcome whose variations are explained by other factors. For example, in a study of sales trends, y would be the sales figure.
- $x^{(1)}$ and $x^{(2)}$ are independent variables, also referred to as predictors, which explain the variation in y. Predictors can include factors like age, gender, or genetic predisposition.
- β_0, β_1, and β_2 are coefficients that quantify the effect of each predictor on y.
- ε is the random error term that accounts for unobserved factors influencing the outcome. One important assumption is that $\mathbb{E}(\varepsilon) = 0$.

For practical applications, linear regression can take two forms: simple linear regression, which involves only one independent variable, and multiple linear regression, which includes two or more independent variables.

Linear regression tries to find the best coefficients $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ that minimize the gap between actual and estimated values (Altman & Krzywinski, 2015):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)}$$

Linear regression is a reliable choice when the decision outcome variable is continuous. The key assumption in linear regression is that the decision outcome is a linear combination of various decision factors.

2.2 Ordinary Least Squares (OLS)

A common way to recover the coefficients is ordinary least squares (OLS). The goal of OLS is to find the coefficient values β that can minimize the total prediction error (Alto, 2023).

The total prediction error can be expressed as:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Here:

- n: The total number of observations.
- y_i: The actual outcome for the i-th observation.
- $\hat{y}_i$: The predicted outcome for the i-th observation.

And $(y_i - \hat{y}_i)$ is the prediction error for one observation. OLS will minimize the sum of squared errors for all observations.

We can get a unique solution through OLS (Penn State Eberly College of Science, n.d.):

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Here:

- $\hat{\beta}$: A $(k+1) \times 1$ vector. It contains the predicted values of the coefficients for all k independent variables, plus the intercept.
- X: A $n \times (k+1)$ matrix. It contains the values of all independent variables and one column for the intercept.
- y A $n \times 1$ vector. It contains the observed values of the dependent variable.

2.3 Interpreting the OLS Solution

We need to first calculate the inverse of the Gram matrix $X^T X$. This matrix has dimensions $(k+1) \times (k+1)$ where k is the number of predictors. The inverse exists only if the columns of X are linearly independent. If one column is a combination of others, the inverse does not exist.

OLS works by minimizing the sum of squared residuals (SSR/RSS). A residual (ε_i) is the difference between the actual outcome (y_i) and the predicted outcome ($\hat{y}_i$):

$$\varepsilon_i = y_i - \hat{y}_i = y_i - \beta^T x_i$$

Here:

- $\hat{y}_i = X\hat{\beta}$, where X is the predictor matrix.
- x_i is the row of predictors for observation i, including a 1 for the intercept.

2.3.1 Variances

After finding the coefficients and residuals, it's important to understand their variances. Variances show how much the results may vary with new data. It is usually assumed that the errors (ε) follow a normal distribution with mean 0 and constant variance (Penny, 2006):

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

The true variance (σ^2) is unknown but can be estimated as:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \varepsilon^T \varepsilon$$

Here:

- ε is the residual vector.
- $n-k-1$ is the degrees of freedom, accounting for the number of parameters.

Using σ^2, we estimate the covariance matrix of the coefficients ($\hat{\beta}$):

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

The diagonal entries of this matrix give the variances of the coefficients. Their square roots are the standard errors, which are used for hypothesis testing and confidence intervals (CI).

2.3.2 Coefficient of Determination (R^2)

The R^2 value measures how well the model explains variability in the outcome. It ranges from 0 to 1 and is computed as (Newcastle University, n.d.):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where:

- $\sum(y_i - \hat{y}_i)^2$ is the residual sum of squares (RSS).
- $\sum(y_i - \bar{y})^2$ is the total sum of squares (TSS).

If $R^2 = 1$, the model predicts all outcomes perfectly. Lower values mean the model explains less variability. R^2 helps assess the model's fit and predictive accuracy.

2.4 Weighted Least Squares (WLS)

WLS extends OLS by addressing situations where observations vary in reliability or precision. OLS assumes that all data points have equal importance. In contrast, WLS assigns a weight to each observation. So compared to OLS, WLS ensures that observations with greater precision will contribute more to the regression model (Harrer, 2021).

WLS becomes very helpful when dealing with heterogeneous data. And it is also preferred in many meta-regression settings. For example, in meta-regression, studies have uneven spread of effect sizes and often differ in sample size or quality. WLS can include these differences, allowing the model to handle variability. By including weights, WLS can improve the accuracy of the result.

Meta-regression model looks similar to the linear regression model:

$$\hat{\theta}_k = \theta + \beta x_k + \varepsilon_k + \zeta_k$$

Here:

- ζ: The between-study heterogeneity. It shows variability across studies.

2.5 Running with statsmodels

We use the ols function from the statsmodels.formula.api library in our project. The code is:

```
from statsmodels.formula.api import ols

model = ols(formula="A ~ B + C", data=df).fit()
```

Here:

- `A`: The dependent variable (the outcome).
- `B` and `C`: The independent variables (the predictors).
- `df`: The DataFrame containing the data.

Key Outputs:

- `model.summary()`: It shows coefficients, standard errors, R2, and other statistics.
- `model.params`: It returns the estimated coefficients.
- `model.predict()`: Predicts outcomes using the model. To predict new values, provide a DataFrame with input variables.

WLS is used when some observations are more reliable than others. It adjusts the regression by assigning weights to each observation.

```
from statsmodels.api import WLS
```

```
model = wls(formula="A ~ B + C", data=df, weights=weights).fit()
```

Here:

- `A`: The dependent variable (the outcome).
- `B` and `C`: The independent variables (the predictors).
- `weights` is a vector assigning importance to each observation.

3 IMPLEMENTATION: APPLYING ACTUAL DATA TO OLS AND WLS MODEL

We first collected data related to CVD factors from several different studies. And then used these data to build our database. The dataset has key information extracted from research reports. It includes: Name of Study, Year of the Study, Country, Number of Patients per Trial Arm, Type of Population, Mean Age, Percentage of Females in the Observation, CVD Risk Median Follow-Up Years, cIMT Progression Maximum Follow-Up Years, Percentage of Observation with cIMT Data, Effect Size (RR): Relative Risk, Lower Confidence Interval (CI) and Upper Confidence Interval (CI).

We started by loading the dataset and also "statsmodels" library. We used this library to perform OLS and WLS analyses.

```python
from statsmodels.formula.api import ols
import pandas as pd
import numpy as np
import statsmodels.api as sm
```

```python
Data = pd.read_csv("all_study_data_combined.csv")
Data
```

Figure 1. Importing libraries and loading the dataset.

	Name of Study	Year	Country	No. of Patients per trial arm	Type of Population	Mean Age	% Female	CVD Risk Median Follow-Up Years	cIMT Progression Maximum Follow-Up Years	% with cIMT Data	cIMT progression per year	effect size (RR)	lower CI	upper CI
0	ACAPS (1)	1990	USA	230.0	Elevated CVD risk	62	48	5.0	6.0	100	-3	0.38	0.14	1.07
1	ACAPS (2)	1990	USA	230.0	Elevated CVD risk	62	48	5.0	6.0	100	1	0.64	0.25	1.65
2	ACT NOW	2006	USA	301.0	Dysglycemia	52	58	2.2	4.0	63	-5	0.44	0.14	1.41
3	ALL-IMT	2010	UK	40.0	Pre-existing CVD	68	43	1.0	1.2	100	-68	0.36	0.09	1.34
4	AMAR	2005	Russia	129.0	Elevated CVD risk	61	0	2.0	2.0	76	-37	0.58	0.24	1.39
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
131	VITAL	2004	Netherlands	100.0	Elevated CVD risk	49	41	1.5	2.5	99	16	1.41	0.45	4.46
132	WISH	2007	USA	175.0	General population	61	100	2.7	3.0	93	-1	3.00	0.12	73.14
133	Yang et al.	2017	China	60.0	Elevated CVD risk	54	72	0.5	0.5	100	-48	0.98	0.02	48.75
134	Yun et al.	2013	China	68.0	Pre-existing CVD	62	40	2.3	4.5	93	-38	0.47	0.21	1.05
135	Zou et al.	2010	China	48.0	Elevated CVD risk	57	59	1.0	1.0	89	-38	1.00	0.02	49.38

Figure 2. Dataset.

3.1 OLS Model

First, we used a simple OLS model to study the relationship between the Effect Size (RR) and chosen independent factors. The factors are: Number of Patients per Trial Arm, Mean Age, Percentage of Females, CVD Risk Median Follow-Up Years, cIMT Progression Maximum Follow-Up Years, Percentage of Observation with cIMT Data, and cIMT Progression per Year.

```
formula = """
Q("effect size (RR)") ~ Q("No. of Patients per trial arm") + Q("Mean Age") +
Q("% Female") + Q("CVD Risk Median Follow-Up Years") +
Q("cIMT Progression Maximum Follow-Up Years") + Q("% with cIMT Data") +
Q("cIMT progression per year")
"""

model_ols = ols(formula, data=Data).fit()
model_ols.summary()
```

Figure 3. OLS code

The results provided a basic understanding of the relationships between the variables.

OLS Regression Results

Dep. Variable:	Q("effect size (RR)")	**R-squared:**	0.069
Model:	OLS	**Adj. R-squared:**	0.018
Method:	Least Squares	**F-statistic:**	1.352
Date:	Tue, 26 Nov 2024	**Prob (F-statistic):**	0.232
Time:	00:50:42	**Log-Likelihood:**	-525.31
No. Observations:	136	**AIC:**	1067.
Df Residuals:	128	**BIC:**	1090.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.1367	10.492	0.585	0.560	-14.623	26.896
Q("No. of Patients per trial arm")	-0.0005	0.001	-0.416	0.678	-0.003	0.002
Q("Mean Age")	0.0487	0.153	0.318	0.751	-0.254	0.351
Q("% Female")	-0.0676	0.042	-1.628	0.106	-0.150	0.015
Q("CVD Risk Median Follow-Up Years")	-0.1492	1.446	-0.103	0.918	-3.010	2.711
Q("cIMT Progression Maximum Follow-Up Years")	-0.8506	1.478	-0.575	0.566	-3.776	2.075
Q("% with cIMT Data")	0.0064	0.048	0.132	0.895	-0.089	0.102
Q("cIMT progression per year")	0.1006	0.037	2.705	0.008	0.027	0.174

Omnibus:	224.037	**Durbin-Watson:**	1.900
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	17915.945
Skew:	7.204	**Prob(JB):**	0.00
Kurtosis:	57.351	**Cond. No.**	1.15e+04

Figure 4. OLS summary

3.2 WLS Model

Then, because the standard error of each effect size is calculated as the difference between the upper and lower bounds of the confidence interval divided by 2 times z (where z = 1.96 for a 95% confidence level), and the weight is the inverse of the variance, we used this formula to calculate the weights:

$$\text{Weight} = \left(\frac{2 \times 1.96}{\text{Upper CI} - \text{Lower CI}} \right)^2$$

We added a new column to the dataset to store these weight values.

```
data_copy = Data.copy()

z = 1.96
data_copy["weight"] = ((2 * z) / (data_copy["upper CI"] - data_copy["lower CI"]))**2

X = data_copy[["No. of Patients per trial arm", "Mean Age", "% Female",
               "CVD Risk Median Follow-Up Years", "cIMT Progression Maximum Follow-Up Years",
               "% with cIMT Data", "cIMT progression per year"]]

X = sm.add_constant(X)

y = data_copy["effect size (RR)"]

wls_model = sm.WLS(y, X, weights=data_copy["weight"]).fit()

print(wls_model.summary())
```

Figure 5. WLS code

```
                            WLS Regression Results
==============================================================================
Dep. Variable:       effect size (RR)   R-squared:                       0.444
Model:                            WLS   Adj. R-squared:                  0.413
Method:                 Least Squares   F-statistic:                     14.58
Date:                Tue, 26 Nov 2024   Prob (F-statistic):           7.37e-14
Time:                        00:50:43   Log-Likelihood:                -166.14
No. Observations:                 136   AIC:                             348.3
Df Residuals:                     128   BIC:                             371.6
Df Model:                           7
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                        0.5605      0.237      2.364      0.020       0.091       1.030
No. of Patients per trial arm             9.162e-06   8.47e-06      1.082      0.281   -7.59e-06    2.59e-05
Mean Age                                     0.0022      0.004      0.550      0.583      -0.006       0.010
% Female                                     0.0003      0.001      0.372      0.710      -0.001       0.002
CVD Risk Median Follow-Up Years             -0.0004      0.016     -0.025      0.980      -0.031       0.030
cIMT Progression Maximum Follow-Up Years     0.0386      0.024      1.635      0.104      -0.008       0.085
% with cIMT Data                          1.491e-05      0.001      0.023      0.982      -0.001       0.001
cIMT progression per year                    0.0070      0.001      5.225      0.000       0.004       0.010
==============================================================================
Omnibus:                       30.258   Durbin-Watson:                   1.909
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              120.471
Skew:                           0.676   Prob(JB):                     6.92e-27
Kurtosis:                       7.408   Cond. No.                     9.55e+04
==============================================================================
```

Figure 6. WLS summary

3.3 Prediction

We also applied the models to predict the Effect Size (RR) for new data. By inputting new observations into the trained model, we can get predictions for the Effect Size (RR).

```
new_data = pd.DataFrame({
    "No. of Patients per trial arm": [200, 300],
    "Mean Age": [55, 65],
    "% Female": [50, 60],
    "CVD Risk Median Follow-Up Years": [2.5, 3.5],
    "cIMT Progression Maximum Follow-Up Years": [4.0, 5.0],
    "% with cIMT Data": [80, 90],
    "cIMT progression per year": [10, 20]
})

new_data_with_const = sm.add_constant(new_data)

predictions = wls_model.predict(new_data_with_const)

new_data["Predicted Effect Size (RR)"] = predictions

new_data
```

Figure 7. Prediction code

	No. of Patients per trial arm	Mean Age	% Female	CVD Risk Median Follow-Up Years	cIMT Progression Maximum Follow-Up Years	% with cIMT Data	cIMT progression per year	Predicted Effect Size (RR)
0	200	55	50	2.5	4.0	80	10	0.926098
1	300	65	60	3.5	5.0	90	20	1.061156

Figure 8. Prediction

4 DISCUSSION

4.1 Key Findings and Observations

Our study explored the relationship between carotid intima-media thickness (cIMT) progression and cardiovascular disease (CVD) risk using Ordinary Least Sqauares (OLS) and Weighted Least Square (WLS) regression to identify significant predictors. Results outputted by the OLS and WLS model, shown in Figure 4 and 6, identified cIMIT progression per year as the most and only significant predictor, with a positive and statistically significant relationship ($p = 0.001$ in OLS and < 0.001 in WLS). This indicates that an increase in cIMIT progression correlates independently and strongly with elevated CVD risk after adjustment with various cofactors involved in the regression model. Our findings align with the earlier meta-analysis that highlights cIMT as a reliable surrogate marker for CVD risk (Willeit, 2020), but with additional robustness due to the weighted methodology. While other predictors, such as mean age, gender distribution, and CVD Risk Median Follow-Up Years, showed negligible contributions and lacked statistical significance, this may indicate that these demographic factors are less directly impactful or are eclipsed by the dominant influence of cIMT progression.

The explanatory power of the WLS model was notable, with an R-squared value of 0.444, indicating that 44.4% of the variability in the effect size was explained by the predictors. This represented a significant improvement over the Ordinary Least Squares (OLS) model, which had an R-squared value of only 0.069. The enhanced performance of the WLS model underscores the importance of accounting for study heterogeneity, such as differences in sample size, study quality, and measurement protocols, which are common in medical meta-analyses. The trend observed with cIMT Progression Maximum Follow-Up Years ($p = 0.104$), while not reaching conventional significance levels, suggests that longer follow-up periods might also have relevance in understanding the long-term impact of cIMT on CVD risk, warranting further investigation.

4.2 Challenges and limitations

Despite the significant findings, the study faced challenges. A high condition number (Cond. No = 9.55e+04) indicated potential multicollinearity, particularly among demographic variables such as mean age and % female. This could have destabilized coefficient estimates, making it harder to disentangle the independent effects of these predictors. In addition, data heterogeneity, stemming from differences in population characteristics, follow-up durations, and cIMT measurement protocols, introduced variability that could not be fully addressed, even with WLS. The exclusion of key covariates, such as lifestyle factors (e.g., smoking or diet), genetic predispositions, and comorbidities, further limited the scope of the analysis and its ability to provide a comprehensive risk assessment. These may be the reasons why a significant portion of the variability in CVD risk remained unexplained (55.6%).

4.3 Future Improvements

To address these limitations, future studies could incorporate additional covariates, such as behavioral and genetic factors, to provide a more holistic view of CVD risk. Non-linear regression models, such as splines or polynomial regression, could be employed to better capture complex relationships, particularly between cIMT progression and long-term outcomes. Expanding the dataset to include diverse populations would improve the generalizability of findings and reveal population-specific risk factors. Finally, standardizing cIMT measurement protocols and follow-up assessments across studies would reduce variability and improve comparability.

5 CONCLUSION

This study reaffirmed the role of cIMT progression per year as a key predictor of cardiovascular disease risk, providing robust evidence of its utility as a surrogate marker. The findings indicate that for every unit increase in cIMT progression per year, there is a statistically significant and meaningful increase in CVD risk. These results align with prior research but go further by applying WLS regression, which accounts for heterogeneity among studies and enhances the reliability of the findings. While demographic variables and other predictors showed limited impact, this highlights the dominant role of cIMT progression in driving CVD risk.

The implications of these findings are significant for clinical practice and public health. Monitoring cIMT progression could enable earlier detection of patients at elevated risk of CVD, facilitating timely intervention and personalized treatment strategies. However, to maximize its predictive power, future research must address the limitations of this study by incorporating additional covariates, expanding dataset diversity, and exploring advanced analytical methods. With these improvements, the field can further refine risk assessment models and contribute to the development of more effective strategies for CVD prevention and management, ultimately reducing morbidity, mortality, and healthcare costs associated with cardiovascular disease.

APPENDIX

Derivation of $\hat{\beta}$

When in vector form, the linear regression model can be written as:

$$y = X\beta + \varepsilon$$

The goal of OLS is to minimize the sum of squared residuals. It can be shown as:

$$\min(y - X\beta)^T (y - X\beta)$$

We can expand this expression:

$$(y - X\beta)^T (y - X\beta) = y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta$$

We notice that the middle two terms, $-y^T X\beta$ and $-\beta^T X^T y$ are transposes of each other. So we can simplify it:

$$(y - X\beta)^T (y - X\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

To find the value of β that minimizes this expression, we take the derivative with respect to β and set it to zero:

$$\frac{\partial}{\partial \beta} \left(y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \right) = 0$$

This simplifies to:

$$-2X^T y + 2(X^T X)\beta = 0$$

Rearranging terms gives:

$$(X^T X)\beta = X^T y$$

Finally, solving for β yields:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Derivation of Var$(\hat{\beta})$

Starting from the solution for $\hat{\beta}$, we can compute its variance. Substituting β into the model:

$$\text{Var}(\hat{\beta}) = \text{Var}((X^T X)^{-1} X^T \varepsilon)$$

Here, β drops out of the variance operation because it is a constant. Applying the variance operator:

$$\text{Var}((X^T X)^{-1} X^T \varepsilon) = (X^T X)^{-1} X^T \text{Var}(\varepsilon) X (X^T X)^{-1}$$

The errors ε are assumed to have a constant variance σ^2, so:

$$\text{Var}(\varepsilon) = \sigma^2 I$$

Substituting this into the formula:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$$

Simplifying further:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Thus, the variance of $\hat{\beta}$ depends on the variance of the residuals (σ^2) and the inverse of the Gram matrix $(X^T X)$. This result is essential for hypothesis testing and constructing confidence intervals for the coefficients.

REFERENCES

Altman, N., & Krzywinski, M. (2015). Simple linear regression. Nature Methods, 12(11), 999–1000. https://doi.org/10.1038/nmeth.3627

Alto, V. (2023, February 14). Understanding Ordinary Least Squares (OLS) regression. Towards Data Science. https://builtin.com/data-science/ols-regression

Gallo, A. (2015, November 4). A refresher on regression analysis: Understanding one of the most important types of data analysis. Harvard Business Review. https://hbr.org/2015/11/a-refresher-on-regression-analysis

Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2021). Doing Meta-Analysis with R: A Hands-On Guide. Boca Raton, FL and London: Chapmann & Hall/CRC Press. ISBN978-0-367-61007-4

Higgins, J. P. T., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions. Wiley-Blackwell

Joynt Maddox, K. E., Elkind, M. S. V., Aparicio, H. J., Commodore-Mensah, Y., de Ferranti, S. D., Dowd, W. N., Hernandez, A. F., Khavjou, O., Michos, E. D., Palaniappan, L., Penko, J., Poudel, R., Roger, V. L., Kazi, D. S., . . . American Heart Association (2024). Forecasting the Burden of Cardiovascular Disease and Stroke in the United States Through 2050-Prevalence of Risk Factors and Disease: A Presidential Advisory From the American Heart Association. Circulation, 150(4), e65–e88. https://doi.org/10.1161/CIR.0000000000001256

Libby, P., Ridker, P. M., & Hansson, G. K. (2011). Progress and challenges in translating the biology of atherosclerosis. Nature, 473(7347), 317–325. https://doi.org/10.1038/nature10146

Ling, Y., Wan, Y., Barinas-Mitchell, E., Fujiyoshi, A., Cui, H., Maimaiti, A., Xu, R., Li, J., Suo, C., & Zaid, M. (2023). Varying Definitions of Carotid Intima-Media Thickness and Future Cardiovascular Disease: A Systematic Review and Meta-Analysis. Journal of the American Heart Association, 12(23), e031217. https://doi.org/10.1161/JAHA.123.031217

Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Barone Gibbs, B., Beaton, A. Z., Boehme, A. K., Commodore-Mensah, Y., Currie, M. E., Elkind, M. S. V., Evenson, K. R., Generoso, G., Heard, D. G., Hiremath, S., Johansen, M. C., Kalani, R., Kazi, D. S., . . . American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee (2024). 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation, 149(8), e347–e913. https://doi.org/10.1161/CIR.0000000000001209

Newcastle University. (n.d.). Coefficient of determination, R-squared. Retrieved November 26, 2024, from https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources

Penn State Eberly College of Science. (n.d.). 5.4 - A matrix formulation of the multiple regression model. STAT 462: Applied Regression Analysis. Retrieved November 26, 2024, from https://online.stat.psu.edu/stat462/node/132/

Penny, W. (2006). Finding the uncertainty in estimating the slope. In Mathematics for brain imaging (Chapter 1.2.4, pp. 18–20, Eq. 1.37). Retrieved from https://ueapsylabs.co.uk/sites/wpenny/mbi/mbi_course.pdf

Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted?. Statistics in medicine, 21(11), 1559–1573. https://doi.org/10.1002/sim.1187

Willeit, P., Tschiderer, L., Allara, E., Reuber, K., Seekircher, L., Gao, L., Liao, X., Lonn, E., Gerstein, H. C., Yusuf, S., Brouwers, F. P., Asselbergs, F. W., van Gilst, W., Anderssen, S. A., Grobbee, D. E., Kastelein, J. J. P., Visseren, F. L. J., Ntaios, G., Hatzitolios, A. I., Savopoulos, C., . . . PROG-IMT and the Proof-ATHERO Study Groups (2020). Carotid Intima-Media Thickness Progression as Surrogate Marker for Cardiovascular Risk: Meta-Analysis of 119 Clinical Trials Involving 100 667 Patients. Circulation, 142(7), 621–642. https://doi.org/10.1161/CIRCULATIONAHA.120.046361