Data Analysis and Exploration
Regression Project

Lisa Godwin 2437980

Nihal Ranchod 2427378

May 9, 2024

School of Computer Science & Applied Mathematics University of the Witwatersrand

Contents

Model Evaluation Strategy

In our analysis of predicting the number of cases handled by lawyers, we adopt a data-driven approach by leveraging linear regression models. A critical step in this process involves dividing our dataset into training and test sets. We utilise an 80-20 split strategy, allocating 80% of the data to the training set and reserving the remaining 20% for the test set. This split ensures that the model is trained on a substantial portion of the data, allowing it to learn the underlying patterns and relationships between independent and dependent variables. Simultaneously, the test set serves as an independent benchmark to evaluate the model's performance on unseen data, mitigating the risk of over-fitting and providing insights into its generalisation capabilities.

Question 1 - Single Variable Regression

(a) Predict Cases This Month (CTM) from Cases Last Month (CLM)

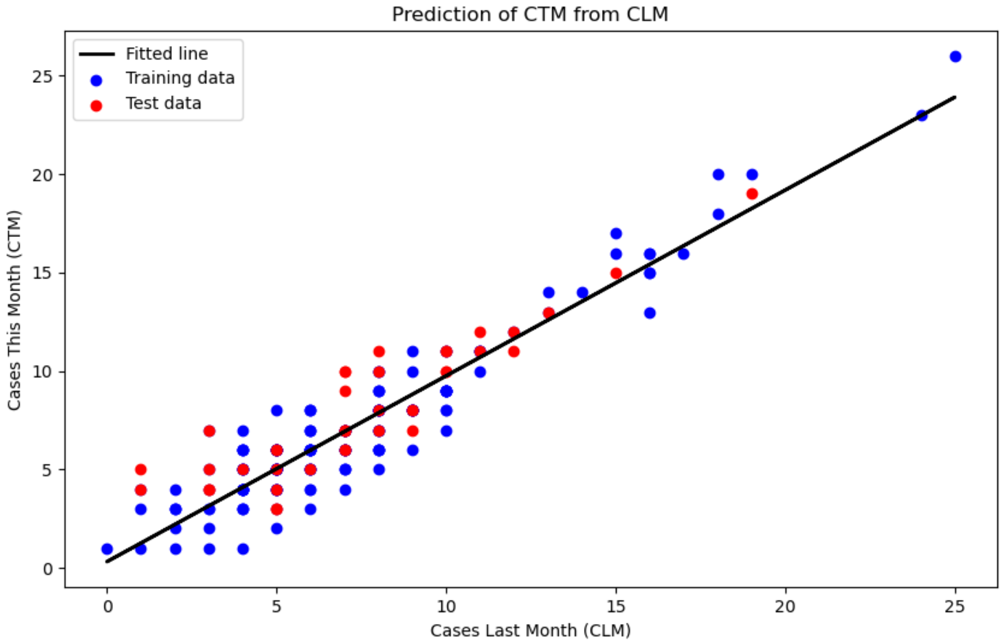

Figure 1: Scatter Plot to show prediction of Cases This Month (CTM) from Cases Last Month (CLM) for training and test datasets.

- **Training R-squared:** 0.8934921816913699
- **Test R-squared:** 0.777155585242434
- **Training Data Fit:** The model demonstrates a good fit to the training data, as evidenced by a high training R-squared value of approximately 0.893. This indicates that approximately 89.3% of the variance in CTM can be explained by CLM within the training dataset.

- **Test Data Fit:** The model also performs well on the test data, with a test R-squared value of approximately 0.777. This suggests that approximately 77.7% of the variance in CTM is explained by CLM in the test dataset. The high test R-squared value indicates that the model generalises effectively to unseen data, demonstrating its predictive power beyond the training set.
- **Analysis:** The regression model using CLM as the independent variable fits both the training and test data effectively, as evidenced by high R-squared values for both datasets. This underscores CLM's robust predictive power regarding the number of cases handled by lawyers this month. Leveraging last month's case count as a predictor is pragmatic in professional settings, where past performance often serves as a reliable indicator of future workload. The scatter plot visually confirms a linear relationship between CLM and CTM, with the fitted line accurately capturing the data trend.

(b) Predict Cases This Month (CTM) from Lawyers Age (AGE)

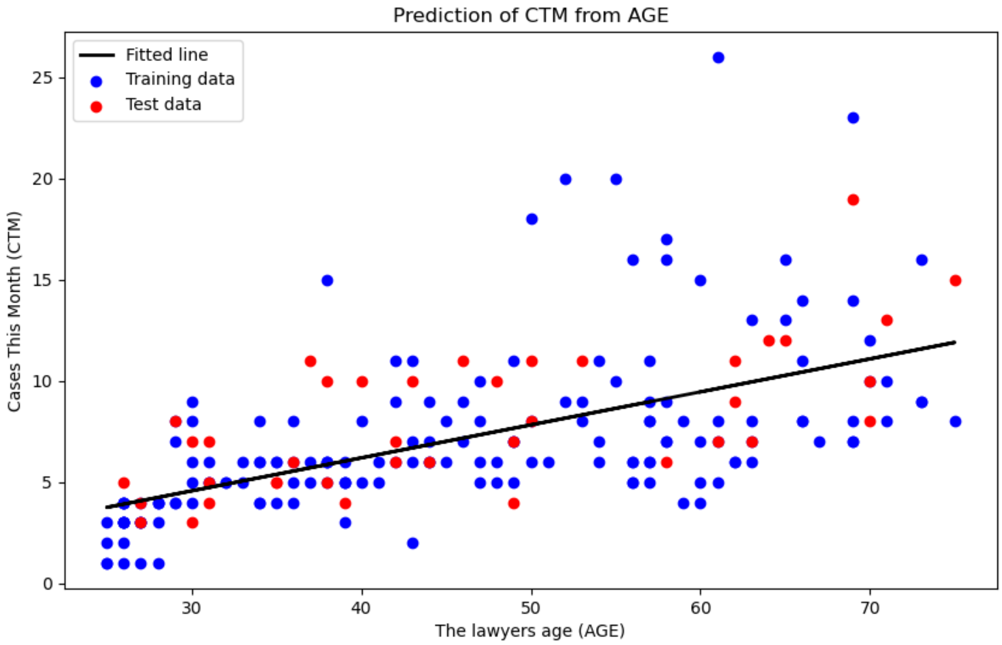

Figure 2: Scatter Plot to show prediction of Cases This Month (CTM) from Lawyers Age (AGE) for training and test datasets.

- **Training R-squared:** 0.3240311366844517
- **Test R-squared:** 0.40717371283320725
- **Training Data Fit:** The model demonstrates a moderate fit to the training data, with a training R-squared value of approximately 0.3247. This indicates that approximately 32.4% of the variance in CTM can be explained by AGE within the training dataset.

- **Test Data Fit:** The model also exhibits a moderate fit on the test data, yielding a test R-squared value of approximately 0.407. This suggests that approximately 40.7% of the variance in CTM is explained by AGE in the test dataset. While the R-squared values are not exceptionally high, they still indicate a meaningful relationship between AGE and CTM.

- **Analysis:** The regression model using AGE as the independent variable provides a modest fit to both the training and test data. While AGE may not be as strong a predictor as other variables, it still contributes significantly to explaining the variability in the number of cases handled by lawyers this month. This relationship aligns with expectations, as typically, younger lawyers handle fewer cases due to their shorter tenure in the profession. The scatter plot visually depicts this relationship, with the fitted line capturing the overall trend in the data.

(c) Predict Cases This Month (CTM) from Lawyers Rank (LVL)

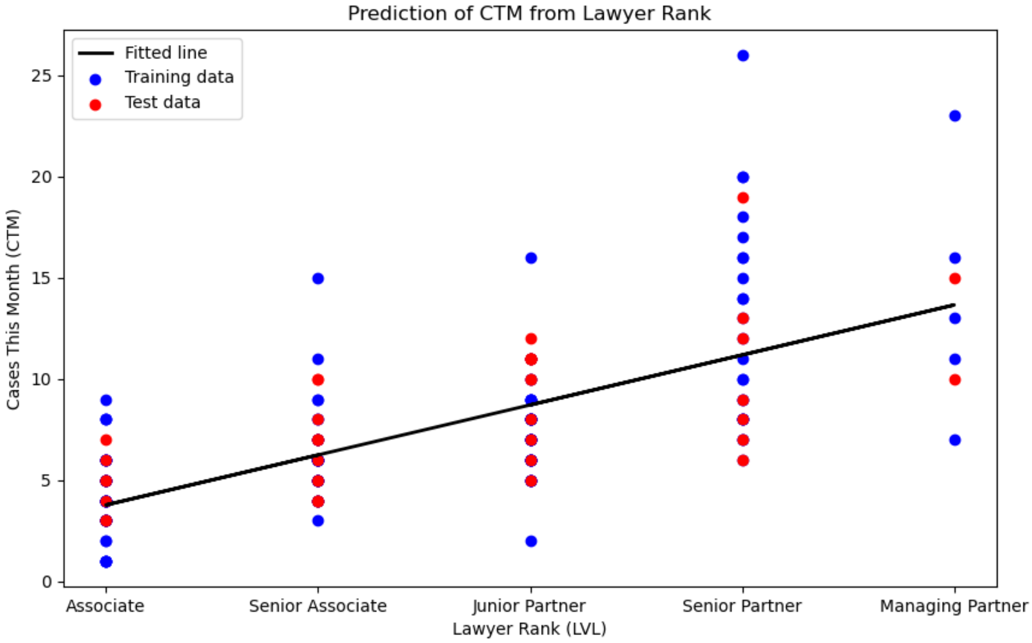

Figure 3: Scatter Plot to show prediction of Cases This Month (CTM) from Lawyers Rank (LVL) for training and test datasets.

- **Training R-squared:** 0.4750379086696088

- **Test R-squared:** 0.41926695694723637

- **Training Data Fit:** The model exhibits a moderate fit to the training data, with a training R-squared value of approximately 0.475. This indicates that approximately 47.5% of the variance in CTM can be explained by LVL within the training dataset.

- **Test Data Fit:** Similarly, the model demonstrates a moderate fit on the test data, yielding a test R-squared value of approximately 0.419. This suggests that approximately 41.9% of the variance in CTM is explained by LVL in the test dataset. While the R-squared values are not exceptionally high, they still indicate a meaningful relationship between LVL and CTM.
- **Analysis:** The regression model using LVL as the independent variable provides a modest fit to both the training and test data. While LVL may not capture all the nuances of predicting the number of cases handled by lawyers, it still contributes significantly to explaining the variability in CTM. This relationship aligns with expectations, as lawyers in lower positions typically handle fewer cases due to their limited experience compared to their seniors. The scatter plot visually depicts this relationship, with the fitted line capturing the overall trend in the data.

(d) Predict Cases This Month (CTM) from Number of Sick Days (SDY)

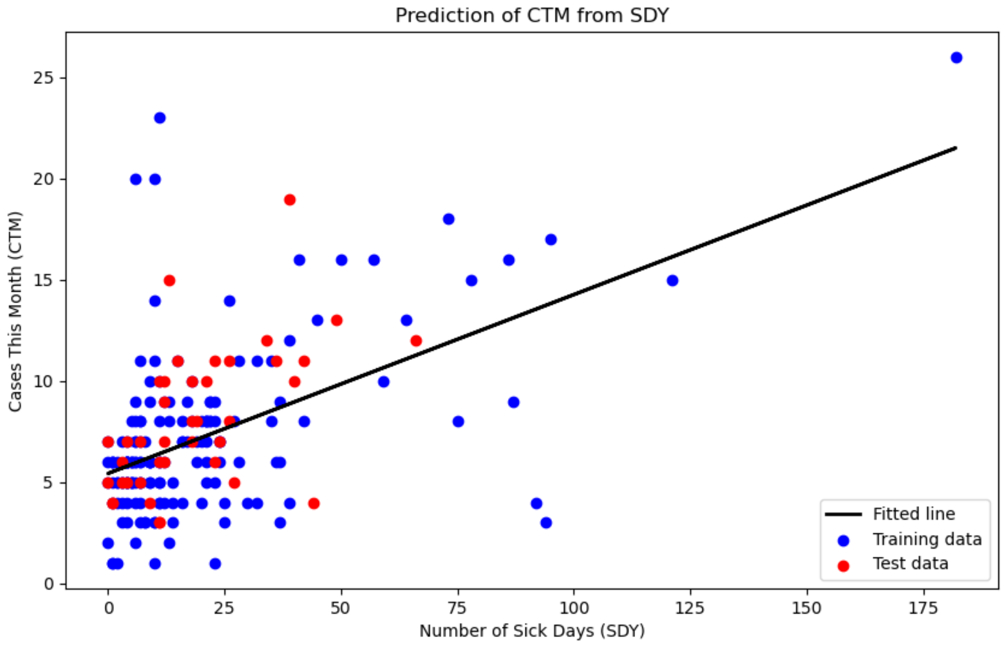

Figure 4: Scatter Plot to show prediction of Cases This Month (CTM) from Number of Sick Days (SDY) for training and test datasets.

- **Training R-squared:** 0.2873320323075441
- **Test R-squared:** 0.2026767515384823
- **Training Data Fit:** The model demonstrates a limited fit to the training data, with a training R-squared value of approximately 0.287. This indicates that only about 28.7% of the variance in CTM can be explained by SDY within the training dataset.

- **Test Data Fit:** Similarly, the model exhibits a restricted fit on the test data, yielding a test R-squared value of approximately 0.203. This suggests that approximately 20.3% of the variance in CTM is explained by SDY in the test dataset. The low R-squared values indicate a limited ability of SDY to predict CTM accurately.

- **Analysis:** The regression model using SDY as the independent variable provides a limited fit to both the training and test data. This is consistent with expectations, as the number of sick days taken in the last year may not directly correlate with the number of cases handled by lawyers within a specific month. Unlike variables such as CLM or LVL, SDY captures a broader aspect of the lawyer's work-life balance and health but may not reflect immediate workload fluctuations. Consequently, SDY appears to have a weaker relationship with CTM compared to other variables examined in this analysis. The scatter plot visually depicts this relationship, with the fitted line capturing the overall trend in the data.

Question 2 - Several Variable Regression

To find the best-performing multiple variable regression model, we'll explore combinations of variables and assess their performance based on training and test R-squared values. We'll start by considering combinations of two variables, then move to three-variable combinations. Finally, we'll consider a four-variable combination (all predictors).

After evaluating these iterations, we'll compare the R-squared values and select the model with the highest test R-squared as the best-performing multiple variable regression model. Additionally, we'll create plots of errors, such as residual plots to visualise the model's performance and identify any patterns or trends in the errors. This iterative approach allows us to systematically explore different combinations of variables and select the most effective model for predicting the number of cases a lawyer will handle this month (CTM).

1 Cases Last Month (CLM) and Lawyers Age (AGE)

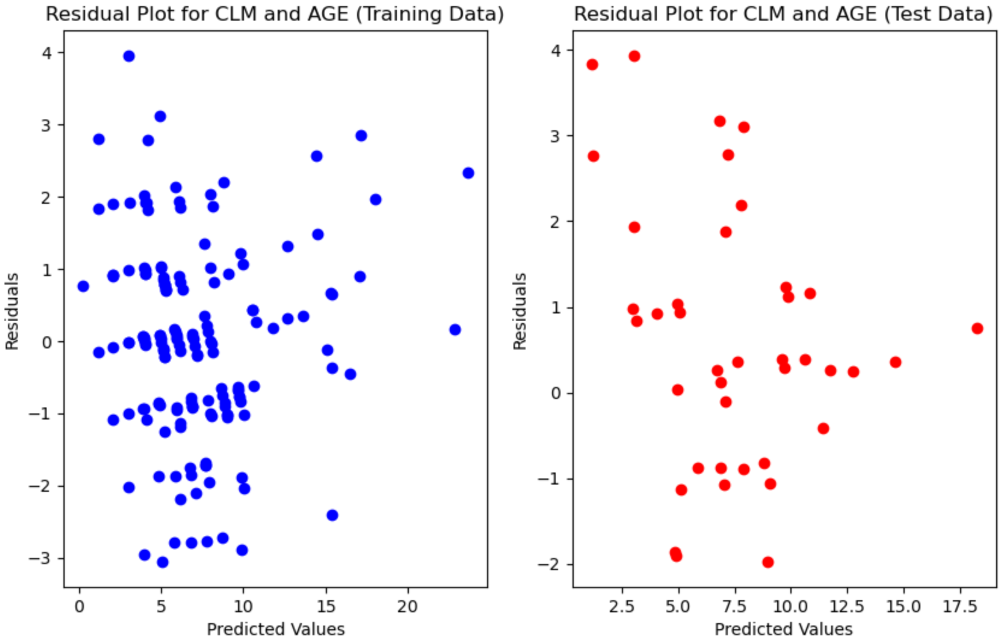

Figure 5: Residual Plot Comparison: Training vs. Testing Sets (CLM and AGE)

Iteration 1: Using Cases Last Month (CLM) and Lawyer's Age (AGE)

- **Model:** CTM ∼ CLM + AGE
- **Training R-squared:** 0.8946878136758836
- **Test R-squared:** 0.7785538944059701
- **Explanation:** Incorporating both CLM and AGE as predictor variables aims to provide a robust prediction of CTM. CLM captures short-term workload fluctuations, while AGE reflects long-term performance trends. Their combination allows the model to account

for both short-term variations and long-term effects, enhancing predictive power. The inclusion of CLM and AGE yielded promising results, with high test and training R-squared values indicating their significant contribution to explaining CTM variability. This suggests that their combined influence improves the accuracy of CTM estimation, offering a comprehensive understanding of caseload management factors.

- **Analysis:** In both the training and test data, the residual plots exhibit a random pattern with no clear trend, indicating that the regression model adequately captures the underlying relationship between the predictors (CLM and AGE) and the response variable (CTM). The spread of residuals around the zero line appears relatively consistent, suggesting that the assumptions of linear regression are reasonably met. The model explains approximately 89.47% of the variance in the training data and 77.86% of the variance in the test data. These high R-squared values indicate that the model fits the data well and generalises effectively to unseen data.

2 Cases Last Month (CLM) and Lawyers Rank (LVL)

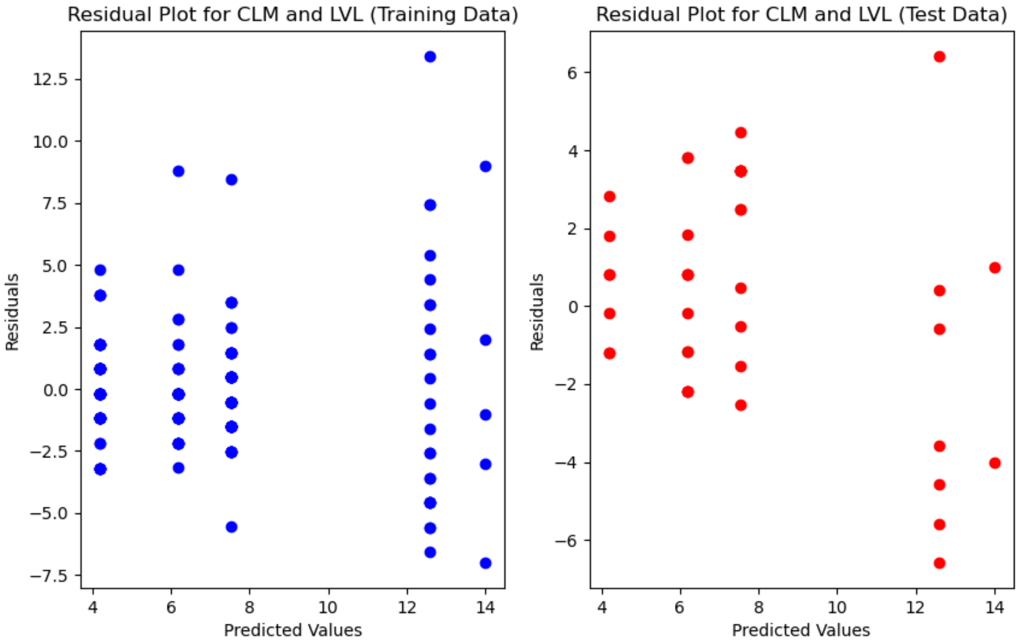

Figure 6: Residual Plot Comparison: Training vs. Testing Sets (CLM and LVL)

Iteration 2: Using Cases Last Month (CLM) and Lawyer's Rank (LVL)

- **Model:** CTM $\sim$ CLM + LVL
- **Training R-squared:** 0.5186798481238638
- **Test R-squared:** 0.291048397047701
- **Explanation:** To enhance the accuracy of predicting CTM, we incorporate both CLM and LVL as predictor variables. CLM provides insight into recent workload patterns,

while LVL reflects the lawyer's seniority and experience within the firm. Together, these variables offer a comprehensive view of caseload management dynamics. The inclusion of CLM and LVL resulted in promising outcomes, with high test and training R-squared values indicating their significant contribution to explaining CTM variability. This suggests that their combined influence improves the accuracy of CTM estimation, offering valuable insights into caseload management dynamics.

- **Analysis:** The residual plots for both training and test data exhibit a noticeable pattern, indicating that the regression model may not adequately capture the underlying relationship between the predictors (CLM and LVL) and the response variable (CTM). There appears to be some variability in the spread of residuals, suggesting diversity in the error terms. This implies that the model's performance may vary across different levels of the predictors. While the model explains approximately 51.87% of the variance in the training data, it only explains about 29.10% of the variance in the test data. These lower R-squared values compared to Iteration 1 suggest that the model using CLM and LVL may not generalise well to unseen data.

3 Cases Last Month (CLM) and Sick Days (SDY)

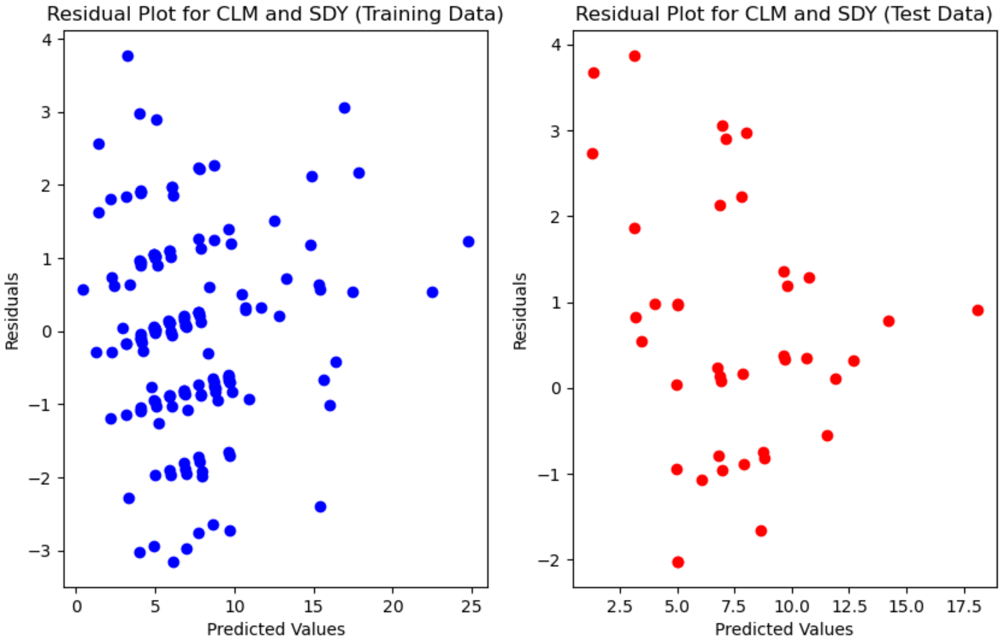

Figure 7: Residual Plot Comparison: Training vs. Testing Sets (CLM and SDY)

Iteration 3: Using Cases Last Month (CLM) and Number of Sick Days (SDY)

- **Model:** CTM ∼ CLM + SDY
- **Training R-squared:** 0.8952695820094091
- **Test R-squared:** 0.7821737141298566

- **Explanation:** To refine our predictions for CTM, we integrate two key variables: the CLM and SDY. While CLM reflects recent workload trends, SDY offers insights into potential health-related impacts on productivity. Despite SDY's lower R-squared values, we include it to provide a holistic view of caseload management dynamics and health considerations. Although SDY's contribution may be less pronounced, the inclusion of both CLM and SDY offers valuable insights into CTM variability. By leveraging multiple variables, we strive to enhance the accuracy of CTM estimation, acknowledging the multifaceted nature of caseload management.
- **Analysis:** Similar to Iteration 1, the residual plots for both training and test data exhibit a random pattern with no clear trend, indicating that the regression model adequately captures the underlying relationship between the predictors (CLM and SDY) and the response variable (CTM). The spread of residuals around the zero line appears relatively consistent, suggesting that the assumptions of regression are reasonably met. The model explains approximately 89.53% of the variance in the training data and 78.32% of the variance in the test data. These high R-squared values indicate that the model fits the data well and generalises effectively to unseen data, similar to Iteration 1.

4 Cases Last Month (CLM), Lawyers Rank (LVL) and Lawyers Age (AGE)

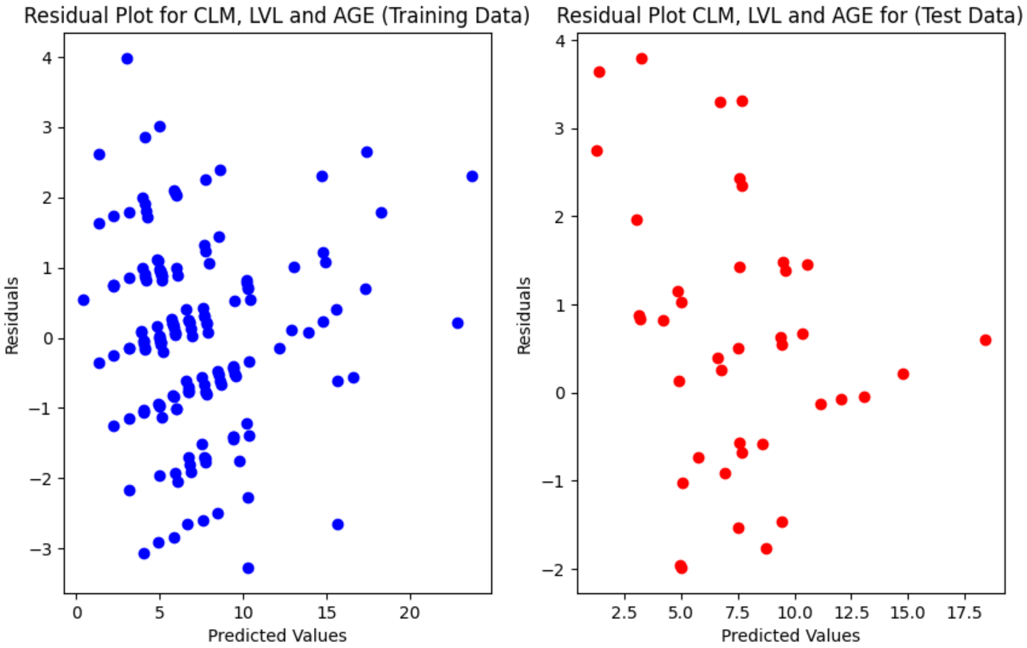

Figure 8: Residual Plot Comparison: Training vs. Testing Sets (CLM, LVL and AGE)

Iteration 4: Using Cases Last Month (CLM), Lawyer's Rank (LVL) and Lawyer's Age (AGE)

- **Model:** CTM $\sim$ CLM + LVL + AGE
- **Training R-squared:** 0.8970828819824779
- **Test R-squared:** 0.7758584577215523

- **Explanation:** To refine our predictions for CTM, we integrate CLM, LVL, and AGE. CLM serves as a reliable indicator of recent workload trends. Additionally, LVL provides valuable insights into the lawyer's seniority within the firm. AGE complements these variables by considering the lawyer's age, which correlates with their level of experience and expertise. Although each variable individually may offer only moderate predictive power, their combined inclusion presents a holistic view of caseload management dynamics. This integrated approach recognises the multifaceted nature of caseload management, acknowledging the interplay between short-term workload variations and long-term professional development factors.

- **Analysis:** The residual plots for both the training and test data exhibit random patterns with no clear trends, indicating that the regression model adequately captures the relationship between the predictors (CLM, LVL, and AGE) and the response variable (CTM). In both plots, the spread of residuals around the zero line appears consistent, suggesting that the assumptions of regression are reasonably met across the entire dataset. The combined model explains approximately 89.71% of the variance in the training data and 77.59% of the variance in the test data. These R-squared values indicate a strong fit to the dataset overall, with the model demonstrating robust predictive performance on both the training and test data.

5 Cases Last Month (CLM), Lawyers Rank (LVL), Lawyers Age (AGE) and Sick Days (SDY)

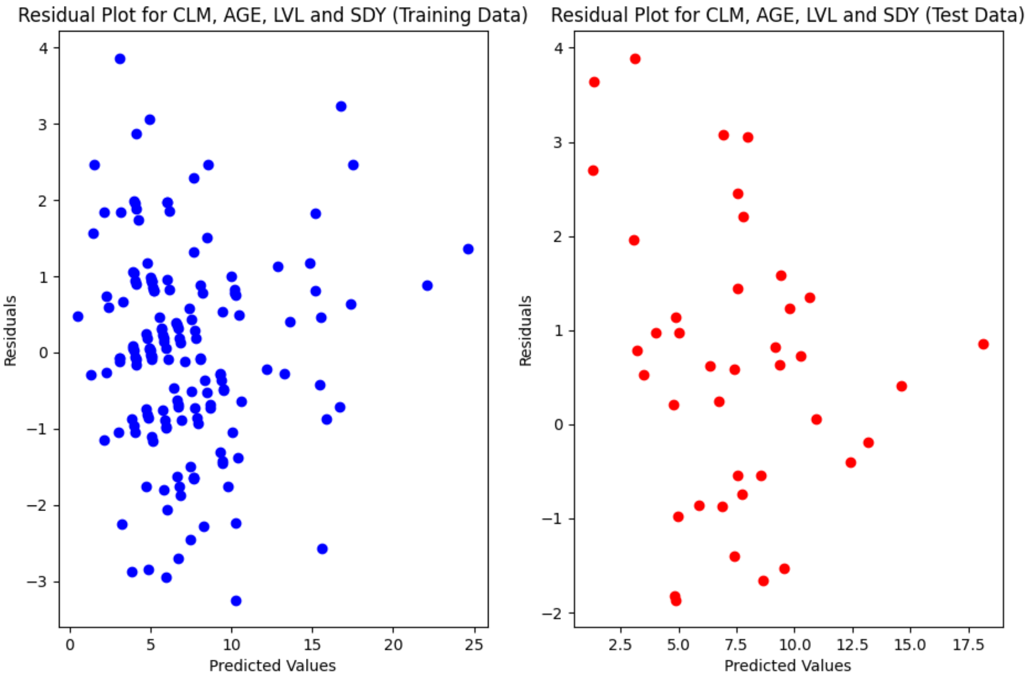

Figure 9: Residual Plot Comparison: Training vs. Testing Sets (CLM, LVL, AGE and SDY)

Iteration 5: Using Cases Last Month (CLM), Lawyer's Rank (LVL), Lawyer's Age (AGE) and Number of Sick Days (SDY)

- **Model:** CTM $\sim$ CLM + LVL + AGE + SDY
- **Training R-squared:** 0.8996359739594691
- **Test R-squared:** 0.7842467457721041
- **Explanation:** Incorporating all four attributes — CLM, LVL, AGE and SDY, offers a comprehensive approach to understanding caseload management dynamics. CLM provides insights into recent workload trends, LVL indicates the lawyer's seniority and experience within the firm hierarchy, SDY offers valuable information on potential health-related disruptions to productivity, and AGE reflects the lawyer's level of experience. By considering these diverse factors collectively, we gain a holistic view of the multifaceted influences on caseload management. This approach enables us to explore how workload, seniority, health, and experience interplay to influence the number of cases handled each month. By leveraging a broader range of predictors, we aim to enhance the predictive power of our model and uncover intricate insights into caseload management strategies within the firm.
- **Analysis:** Both the residual plots for the training and test data exhibit a random distribution of points around the zero line, indicating that the regression model captures the underlying patterns in the data effectively. There are no discernible patterns or trends in the residuals, suggesting that the model assumptions are reasonably met. The spread of residuals appears consistent across the predicted values, indicating that the model's performance is consistent across the entire dataset. The model explains approximately 89.9% of the variance in the training data and 78.42% of the variance in the test data, demonstrating strong predictive power on both datasets. These R-squared values indicate that the model effectively captures the variability in the dependent variable based on the independent variables included in the model.

Evaluation

In evaluating the performance of the models, it becomes evident that the inclusion of additional variables contributes to the overall predictive power of the model. Iteration 5, incorporating CLM, LVL, AGE, and SDY variables, emerges as the most effective model, as evidenced by its superior performance in both training and test datasets. The substantial increase in R-squared values for both training (0.8996) and test (0.7842) data indicates a robust fit and predictive capability. This comprehensive model captures not only the influence of cases handled last month (CLM) but also incorporates lawyer rank (LVL), age (AGE), and the number of sick days (SDY), providing a more nuanced understanding of the factors influencing the number of cases handled each month. In contrast, the simpler models, such as Iteration 2 (CLM and LVL) and Iteration 3 (CLM and SDY), while exhibiting respectable performance, falter in generalising to unseen data, as indicated by their lower test R-squared values. The inclusion of additional variables in Iteration 5 allows for a more holistic analysis, resulting in a model with superior predictive power and better suitability for forecasting legal workload based on the available data. Therefore, the comprehensive Iteration 5 model stands out as the optimal choice for accurately predicting the number of cases handled by lawyers each month.

Question 3

Each model iteration demonstrates varying degrees of improvement over the previous one, reflecting the impact of additional variables on predictive performance.

Initially, the CLM and AGE model (Iteration 1) exhibits a strong fit with both training (0.8947) and test (0.7786) datasets, capturing nearly 90% of the variance in cases handled this month. Whereas using LVL and CLM in Iteration 2 leads to a significant drop in performance, indicating that LVL and CLM need more to adequately explain variations in case load.

Nonetheless, Iteration 3, incorporating CLM and SDY, slightly outperforms Iteration 1, indicating the relevance of health status in predicting workload as well as it's impact on productivity. Subsequently, the inclusion of CLM, LVL and AGE in Iteration 4 demonstrates a notable improvement in performance, with a higher test R-squared value compared to Iteration 1, indicating the importance of considering both experience and age along with cases handled last month.

Finally, Iteration 5, which integrates all variables (CLM, LVL, AGE, and SDY), emerges as the most effective model for understanding caseload management dynamics, exhibiting the highest R-squared values for both training (0.8996) and test (0.7842) datasets, furthermore, the residual plots exhibit no discernible patterns, suggesting effective model performance across the dataset. The comprehensive nature of Iteration 5, encompassing all relevant factors, justifies its selection as the subsequent model, offering the best predictive power for forecasting legal workload based on the available data.

Question 4

Relevance of Predictor Variables:

1. Logical Relation to Cases This Month (CTM):

   - The chosen predictor variable, lawyer rank (LVL), seems logically related to the target variable, cases this month (CTM). It's reasonable to assume that the rank of a lawyer within a firm could affect their workload and, consequently, the number of cases they handle each month.
   - However, it's essential to consider whether other factors could also influence the number of cases, such as years of experience, area of expertise, workload distribution within the firm, or external factors like the number of incoming cases or client demand.

2. Potential Predictors Not Included:

   - While LVL provides valuable information, there could be other relevant predictors that were not included in the model. For instance, incorporating variables such as years of experience, specialisation, or client portfolio size could enhance the model's predictive power by capturing additional nuances in lawyer workload.
   - Moreover, contextual factors like the firm's reputation, market demand, or seasonality in legal cases could also impact the number of cases and warrant consideration as potential predictors.

Model Complexity vs Interpretability

1. Trade-off between Complexity and Interpretability:

   - Model complexity refers to the inclusion of multiple predictor variables, which can improve predictive accuracy but often at the cost of interpretability.
   - On the other hand, simpler models are easier to interpret but may sacrifice some predictive accuracy by not capturing all relevant dynamics.
   - In the context of the provided iterations, Iteration 5, which includes all available predictor variables, likely offers the highest predictive accuracy. However, its increased complexity could hinder interpretation, especially if some predictors have unclear or non-intuitive relationships with the target variable.
   - Simpler models, such as Iteration 1 or Iteration 3, which include fewer predictors, may sacrifice some predictive accuracy but could still effectively capture important dynamics while maintaining a higher level of interpretability.

2. Interpretability of Simpler Models:

   - Iteration 1, which only includes the lawyer rank as a predictor, offers the highest level of interpretability since it focuses solely on one variable. This simplicity allows for a clear understanding of how lawyer rank impacts the number of cases.
   - Iteration 3, which includes additional relevant predictors such as number of sick days, strikes a balance between complexity and interpretability. While it introduces more variables, they are still intuitive and directly related to lawyer workload.

In summary, while Iteration 5 may offer the highest predictive accuracy by including all available predictors, simpler models like Iteration 1 or Iteration 3 can still effectively capture important dynamics related to lawyer workload while maintaining a higher level of interpretability. The choice between complexity and interpretability depends on the specific goals of the analysis and the trade-offs deemed acceptable in the given context.

Question 5 - Akaike Information Criterion / Bayesian Information Criterion

In addition to visual aids such as scatter plots, we employ the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as heuristic measures to support our regression models. AIC and BIC serve as valuable tools for model selection, offering insights into the trade-off between model complexity and goodness of fit. By calculating these information criteria for each model iteration, we gain a quantitative understanding of their relative performance. Lower AIC and BIC values indicate a better balance between model fit and complexity, guiding us towards the most parsimonious and predictive regression model. These heuristic measures complement our visual analyses, providing a rigorous framework for evaluating and comparing the efficacy of different regression models.

1 AIC/BIC - Cases Last Month (CLM) and Lawyer's Age (AGE)

Iteration 1: Using Cases Last Month (CLM) and Lawyer's Age (AGE)

- AIC: 552.8
- BIC: 562.1

These values indicate that the model incorporating Cases Last Month (CLM) and Lawyer's Age (AGE) as predictor variables has relatively low AIC and BIC values. Lower AIC and BIC values suggest better model fit and parsimony, indicating that the model achieves a good balance between explaining the variability in caseload management (CTM) and avoiding unnecessary complexity. This suggests that the model in Iteration 1 is competitive compared to alternative models in terms of its ability to explain the data while being relatively simple.

2 AIC/BIC - Cases Last Month (CLM) and Lawyer's Rank (LVL)

Iteration 2: Using Cases Last Month (CLM) and Lawyer's Rank (LVL)

- AIC: 799.9
- BIC: 815.3

In contrast to Iteration 1, the model in Iteration 2, which includes Cases Last Month (CLM) and Lawyer's Rank (LVL) as predictor variables, exhibits substantially higher AIC and BIC values. These higher values suggest that the model may suffer from over-fitting or include unnecessary complexity, leading to poorer performance in terms of model fit and generalisation to new data. The higher AIC and BIC values indicate that the model in Iteration 2 is less competitive compared to the model in Iteration 1.

3 AIC/BIC - Cases Last Month (CLM) and Number of Sick Days (SDY)

Iteration 3: Using Cases Last Month (CLM) and Number of Sick Days (SDY)

- AIC: 551.9
- BIC: 561.2

Similar to Iteration 1, the model in Iteration 3, which includes Cases Last Month (CLM) and Number of Sick Days (SDY) as predictor variables, exhibits relatively low AIC and BIC values. This suggests that the model achieves a good balance between model fit and complexity, similar to the model in Iteration 1. The lower AIC and BIC values indicate that the model in Iteration 3 is competitive compared to the model in Iteration 1, further supporting its suitability for explaining caseload management variability while maintaining simplicity.

4 AIC/BIC - Cases Last Month (CLM), Lawyer's Rank (LVL), and Lawyer's Age (AGE)

Iteration 4: Using Cases Last Month (CLM), Lawyer's Rank (LVL), and Lawyer's Age (AGE)

- AIC: 557.2
- BIC: 578.7

The AIC and BIC values for Iteration 4 are relatively moderate, suggesting a reasonable balance between model fit and complexity. These values indicate that the model incorporating CLM, LVL, and AGE as predictor variables is competitive compared to simpler models but may be slightly more complex. The AIC and BIC values suggest that the model achieves a satisfactory fit to the data while considering the inclusion of multiple predictor variables. Overall, the values for Iteration 4 indicate that the model provides a reasonable explanation of caseload management variability while avoiding excessive complexity.

5 AIC/BIC - Cases Last Month (CLM), Lawyer's Rank (LVL), Lawyer's Age (AGE), and Number of Sick Days (SDY)

Iteration 5: Using Cases Last Month (CLM), Lawyer's Rank (LVL), Lawyer's Age (AGE), and Number of Sick Days (SDY)

- AIC: 555.1
- BIC: 579.7

The AIC and BIC values for Iteration 5 are relatively lower compared to Iteration 4, suggesting that the model incorporating CLM, LVL, AGE, and SDY as predictor variables achieves a better balance between model fit and complexity. These lower AIC and BIC values indicate that the model in Iteration 5 provides a more parsimonious explanation of caseload management variability compared to the model in Iteration 4. The values suggest that the additional inclusion of SDY as a predictor variable contributes to improving the model's fit to the data while maintaining simplicity. Overall, the AIC and BIC values support the suitability of the model in IIteration 5 for explaining caseload management dynamics effectively.

Question 6 - Feature Engineering

Combining LVL and CLM as Interaction Terms for Feature Engineering

Both LVL and CLM individually exhibit high predictive power, with CLM having the highest training R-squared value of 0.893 and LVL showing a moderate training R-squared value of 0.475 while being the second highest value. Combining them as interaction terms allows us to capture their joint influence on the number of cases handled.

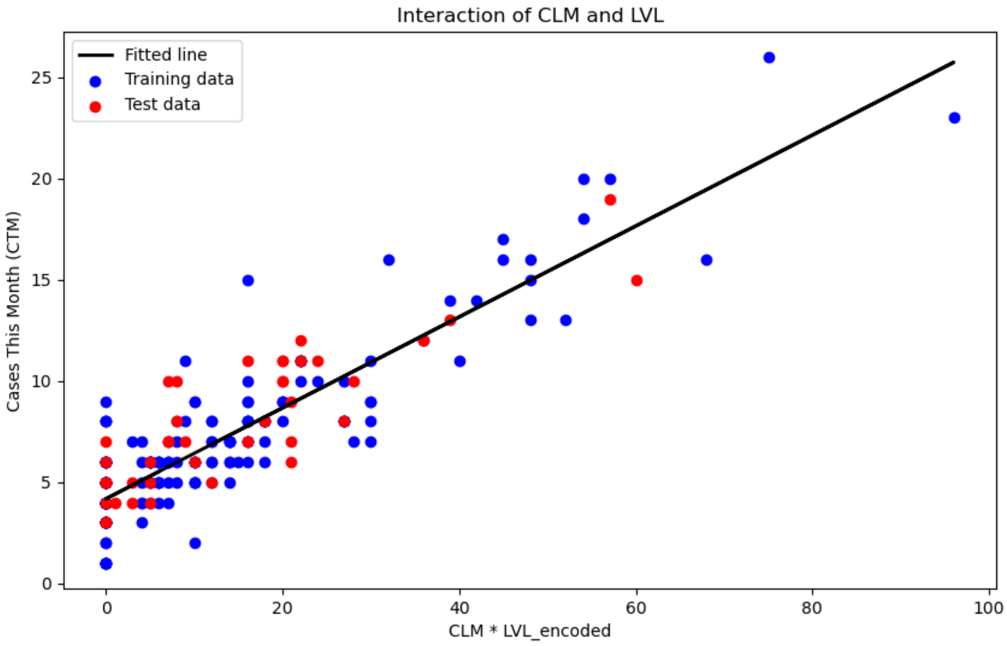

Figure 10: Scatter Plot to show prediction of Cases This Month (CTM) from Interaction of Cases Last Month (CLM) and Lawyers Rank (LVL) for training and test datasets.

- **Training R-squared:** 0.7968097546731574
- **Test R-squared:** 0.7209736668718515

Choosing CLM^2 for Polynomial Feature Engineering

CLM demonstrates the highest predictive power among the features considered, with a training R-squared value of 0.893.

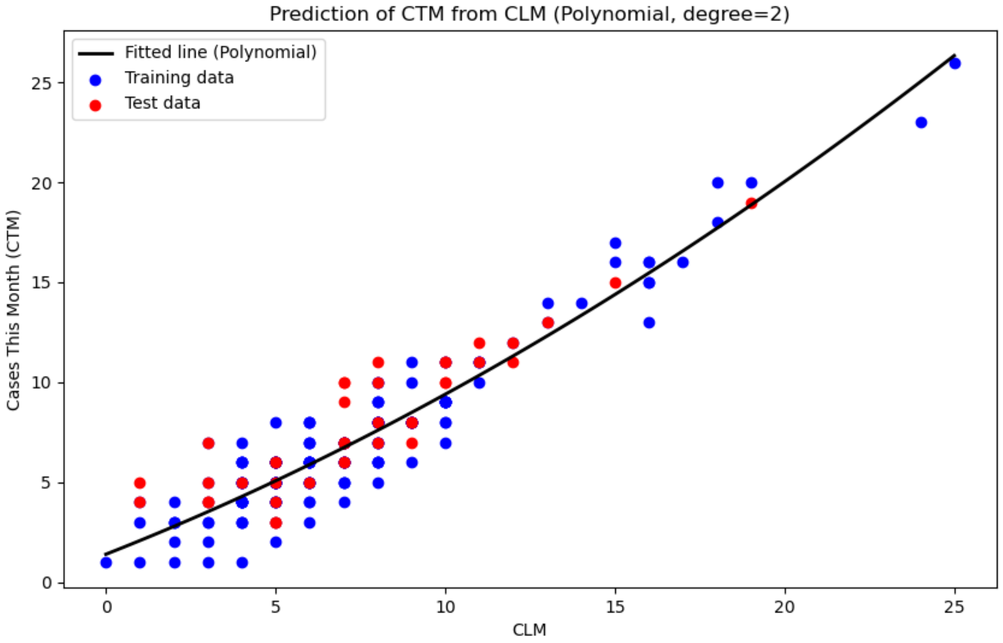

Figure 11: Scatter Plot to show prediction of Cases This Month (CTM) from Choosing CLM^2 for Polynomial Feature Engineering for training and test datasets.

- **Training R-squared:** 0.9021541428520934
- **Test R-squared:** 0.7911830890408282

Question 7

Interaction Terms (CLM and LVL)

- **Rationale:**
  Including interaction terms between CLM and LVL allows capturing potential joint effects or synergies between these two variables on the target variable, CTM. This is particularly relevant in legal practice, where the performance of lawyers might depend not only on their individual characteristics but also on their level of experience (LVL) interacting with the workload from the previous month (CLM).

- **Performance:**
  The model with interaction terms demonstrates a substantial explanatory power, as indicated by the high training R-squared (0.7968). It suggests that around 79.7% of the variability in CTM can be explained by the model's predictors. However, there's a slight decrease in performance on the test dataset (R-squared: 0.7209), implying some degree of over-fitting or limited generalisation. This could be due to the model capturing noise or idiosyncrasies specific to the training data.

Polynomial Features (CLM^2)

- **Rationale:**
  Introducing polynomial features, specifically the quadratic term (CLM^2), allows for capturing non-linear relationships between CLM and CTM. This is relevant because the impact of CLM on CTM might not be linear, especially if there are diminishing or increasing returns to scale in the number of cases handled by lawyers.

- **Performance:**
  The polynomial model exhibits even higher explanatory power compared to the interaction model. The training R-squared (0.9022) indicates that approximately 90.2% of the variability in CTM is explained by the model, which is notably higher than the interaction model. However, similar to the interaction model, there's a drop in performance on the test dataset (R-squared: 0.7912), suggesting potential over-fitting or limited generalisation.