Modelling Heteroscedasticity for Fair Regression using Polynomial Models

Mana Douma[1] and Emma Beauxis-Aussalet[1[0000−0002−4657−892X]]

Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands
{m.a.douma, e.m.a.l.beauxisaussalet}@vu.nl

Abstract. Ensuring an even distribution of AI errors across social groups is a key aspect of fairness: AI errors should not be systematically larger or more frequent for specific social groups. For AI systems predicting numeric values (i.e., regression problems), residuals are the primary error metrics. Even or uneven distributions of residuals is a problem of homo- or heteroscedasticity: a regression system is fair if its residuals are randomly and homogeneously distributed (i.e., homoscedastic) for all social groups (e.g., defined by sensitive features such as race or gender). Thus modeling heteroscedasticity is important to identify fairness issues. We focus on parametric methods for modeling the heteroscedasticity of residuals, as their parameters facilitate human interpretation when identify heteroscedasticity. State-of-the-art parametric methods model residuals' heteroscedasticity use simple linear models. We demonstrate key limitations of such approach, and how these limitations are can be addressed by using polynomial models, and signed residuals for the most complex cases. Polynomial models can address complex cases of heteroscedasticity that would remained undetected using simple linear models. However, interpreting their results is more complex, and future work is needed to assess the impact of outliers and overfitting.

Keywords: Fair AI · Regression Problem · Heteroscedasticity.

1 Introduction

Algorithmic bias in artificial intelligence (AI) has become a prominent issue, as cases discrimination have been arising in many applications [7, 12, 14]. In regression problems, a form of bias is the heteroscedasticity of residuals: e.g., if residuals are not uniformly distributed over the AI output, heteroscedasticity is present. It means that the AI results are more inaccurate for specific data points. Heteroscedasticity may present for data points belonging to a specific social group, but absent for other social groups. This would create a fairness issue and a risk of discrimination, as specific social group are impacted by larger AI errors. Modelling heteroscedasticity with parametric models can be used for identifying heteroscedasticity, by analysing the magnitude and statistical significance of their parameters.

Current parametric methods use simple linear methods to identify heteroscedasticity [2–6, 8]. Given AI results $\hat{\mathbf{y}}$, and groundtruth values $\mathbf{y}$, state-of-the-art methods model the squared or absolute residuals as:

$$(\hat{\mathbf{y}} - \mathbf{y})^2 = \beta_0 + \beta_1 \mathbf{x} \tag{1}$$

$$|\hat{\mathbf{y}} - \mathbf{y}| = \beta_0 + \beta_1 \mathbf{x} \tag{2}$$

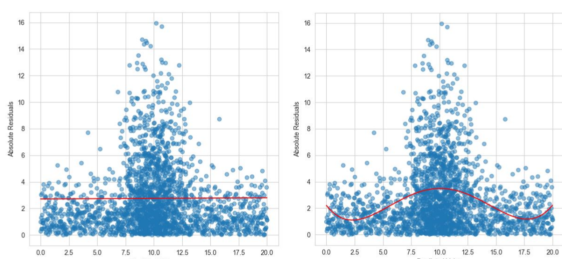
where β_0 is the intercept, β_1 the coefficient (also called slope), and $\mathbf{x}$ can be the AI results $\hat{\mathbf{y}}$, the groundtruth value $\mathbf{y}$, a numeric sensitive feature (e.g., age), or any of the numeric input features of the AI system. The presence of significant non-zero slope indicates heteroscedasticity in the residuals' distribution over feature $\mathbf{x}$.

However, this approach can overlook complex cases of heteroscedasticity (e.g., Fig. 1) resulting in compromised validity of the heteroscedasticity analyses [13]. We thus investigate the use of polynomial models compared to simple linear models. Polynomial models are particularly suited for capturing complex, non-linear relationships between variables. We thus argue that we better model squared or absolute residuals as:

$$(\hat{\mathbf{y}} - \mathbf{y})^2 = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{x}^2 + ... + \beta_n \mathbf{x}^n \tag{3}$$

$$|\hat{\mathbf{y}} - \mathbf{y}| = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{x}^2 + ... + \beta_n \mathbf{x}^n \tag{4}$$

where β_0 is the intercept, β_1 the coefficient of the linear term, $\beta_2, ..., \beta_n$ the parameters β of the polynomial terms. With polynomial models, any non-zero coefficient indicate heteroscedasticity. To assume homoscedastic residuals, all parameters β must be zero. It provides a stricter set of criteria that are sensitive to more complex patterns of variability in the residuals.

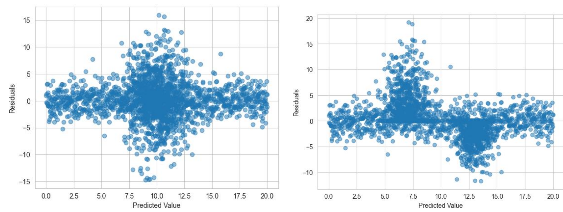


(a) Degree 1 (b) Degree 4

Fig. 1: Example of heteroscedasticity in (simulated) residuals that is not identified with linear models (a) but can be identified with polynomial models (b).

Analysing complex cases of heteroscedasticity in AI errors with polynomial models allows to better identify fairness issues, and to better estimate the uncertainty of AI results. Thus it provides insights into the ethical implications of biased AI systems, and the robustness of AI systems.

While homoscedastic residuals might be reasonably expected (e.g., as a constraint that guides the convergence of robust AI algorithms), it may not be the case for specific data subsets. Heteroscedastic residuals may occur for specific



(a) Symmetric residuals (b) Asymmetric residuals

Fig. 2: Example of heteroscedasticity in (simulated) residual that can be modelled using squared or absolute residuals (a), or requires signed residuals (b).

data subsets (e.g., for data points with specific sensitive features), while remaining generally homoscedastic for the whole dataset.

Complex patterns of heteroscedasticity may also remain undetected if squared or absolute residuals are used (e.g., Fig. 2). Signed residuals allow to detect heteroscedasticity when positive and negative residuals have different distributions. We thus model signed residuals as:

$$\hat{\mathbf{y}} - \mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{x}^2 + ... + \beta_n \mathbf{x}^n \quad (5)$$

Selecting an appropriate heteroscedasticity model is essential to guarantee the validity and completeness of AI bias analysis. To contribute to understanding the applicability bias analysis methods, we study the applicability of linear and polynomial models with the following research questions:

RQ1 How efficient are polynomial models for modelling heteroscedasticity?
RQ2 How to interpret polynomial models to identify fairness issues due to heteroscedasticity?
RQ3 How does using signed or absolute residuals impact the detection of heteroscedasticity?

We will assess the performance of polynomial regression models across different levels of complexity (i.e., polynomials of degrees 1 to 6), applied to the dataset from Obermeyer et al. [12] where racial discrimination was identified. We will review a wide array of evaluation metrics for each heteroscedasticity model, and each of their coefficient (e.g., p-values, deviance). While our results clearly demonstrate the advantages of using polynomial models, we cannot assume that these models are sufficient to detect all forms of heteroscedasticity. Other approaches to modelling complex heteroscedasticity remain unexplored, e.g., parametric, non-parametric, or multivariate models with weighted sensitive features.

2 Related Work

Fairness metrics in regression problems focus either on statistical parity (which is not concerned with residuals) or loss-based (e.g., bounded group loss) which uses squared residuals [1]. These do not account for heteroscedasticity, i.e., fairness criteria may be satisfied while a protected group is more impacted by heteroscedasticity. Thus, with such approach an AI model may be considered fair while a protected group still suffers from systematically higher AI errors (e.g., for predicted values in [7.5, 12.5] in Fig. 1).

Heteroscedasticity can be identified with statistical tests of normality applied to the residuals. However, this approach does not model the magnitude of heteroscedasticity, nor how it can impact protected groups differently, e.g., for data points with specific predicted or actual values ($\hat{y}$ or y).

Existing methods for modeling heteroscedasticity use linear models [2, 3, 8, 10]. A significant slope β_1 (1-2) indicates heteroscedasticity. These methods are not suited to identify all patterns of heteroscedasticity (e.g., Fig.1-2). These methods are also impacted by the choice of residual measurement, e.g., absolute, squared, or signed residuals (1-2). Squared residuals are frequently used, but are sensitive to outliers and distort small residuals. To address some of these limitations, other methods use absolute residuals as the dependent variable [4, 6]. These methods mitigate excessive outlier influence, and provide an alternative perspective on the residual distribution by focusing on the magnitude of differences between observed and predicted values. However, complex patterns of residuals may not be identifiable using neither absolute or squared residuals (Fig. 2).

Su et al. [16] apply local polynomial regressions, fitting multiple polynomial equations to different regions of the data. This method is non-parametric, thus does not require prior knowledge of the form of the heteroscedasticity function. Yet, it is applied exclusively to squared residuals, and only provide local models. In contrast, our approach provides global models of heteroscedasticity, and address asymmetric patterns by using absolute residuals. Out approach also decreases the computational complexity, requires fewer data points, and is easier to interpret.

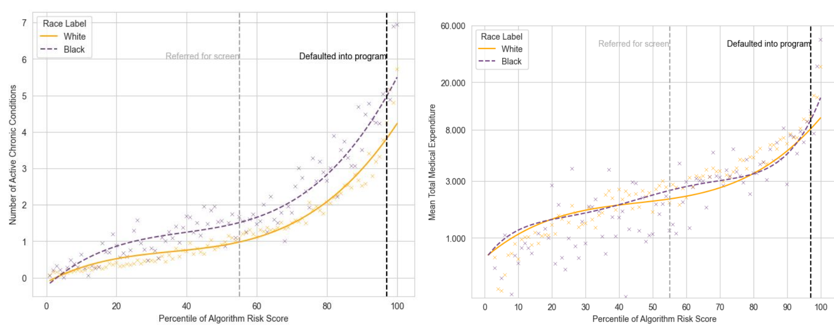
Hsiao et al. [6] present a model to analyze complex financial data, incorporating various forms of heteroscedasticity such as linear, nonlinear, curvilinear (e.g., polynomial regressions), and composition functions. However, this method only uses squared residuals and its goal is to incorporating heteroscedasticity into the initial AI model. In contrast, our work also is also applied to signed residuals, and aims at identifying fairness issues with specific interpretations of polynomial parameters β.

3 Method

We used the dataset studied by Obermeyer et al. [12] because it contains racial bias creating fairness issues. It consists of a comprehensive collection of medical information, and AI-generated scores used throughout the USA to determine patient eligibility for a specialized healthcare program. Its data points represent patients from two racial groups: black and white. Each patient has a detailed set of medical information, including the variables *number of active chronic illness*, and *total medical expenditure* spent in hospital treatments. Their AI-generated risk scores range from 0 (low risk) to 100 (high risk), and are used to assess their health status. When an individual's risk score exceeds the 55^{th} percentile, they are referred for screening by a medical professional who determines their eligibility for the healthcare program. Patients with risk scores above the 97^{th} percentile are automatically admitted to the program.

Obermeyer et al. found significant biases when comparing signs of illnesses (*number of active chronic conditions*) at a given risk score, showing that black patients are considerably sicker than white patients at the same risk score. The AI bias is due to using *total medical expenses* as a predictor, as black patients often have less financial resources thus cannot afford to spend as much as white patients.

We used this dataset to create 2 experimental use cases: regression problems predicting (1) *number of active chronic conditions* and (2) *total medical expenses*, using risk score percentiles (Fig. 3). We analyzed the residuals from these regression problems, to assess potential heteroscedasticity. If heteroscedasticity is present, and impacts one race more than the other, a fairness issue is identified. We researched the effectiveness of polynomial models to identify such issues with heteroscedasticity.



(a) Chronic Illnesses case (b) Medical Expenditure case

Fig. 3: Experimental data drawn from Obermeyer et al. (2019)

To ensure clarity in graphical representations, we opted for a data aggregation method akin to Obermeyer et al.'s approach, which entails the use of mean values per risk score percentile. It is imperative to clarify that the results presented in this paper do not directly mirror the levels of heteroscedasticity within the original dataset. Rather, they offer an overview of polynomial regression models' performance in addressing heteroscedasticity.

Experimental Data We calculated the mean *active chronic illnesses* and the *total medical expenditure* per race and risk score percentile. This reduced the dataset from 48785 to 400 data points: 100 data points per race and variable (*active chronic illnesses* and the *total medical expenditure*). We then fit a polynomial model with degree 3 to the data points (Fig. 3).

We modelled the heteroscedascity in our two experimental use cases with polynomial regression of degrees 1 to 6, fitted with Ordinary Least Squares

(OLS). We assessed the heteroscedasticity modelling using well established metrics: mean squared error (MSE), p-values for each coefficient, deviance, and change in deviance between linear and polynomial models [11].

Deviance is calculated from the likelihood function, representing the probability of observing the residuals given the model's parameters $(\beta_0, ..., \beta_n)$. Higher likelihood means that the observed residuals are more probable under that regression model. The likelihood function $L(\beta; r)$ is defined as the joint probability of the observed residuals given the parameters β:

$$L(\beta; r) = \prod_{i=1}^{n} f(r_i, \hat{r_i}, \beta)$$

where $f(r_i, \hat{r_i}, \beta)$ is the probability density function of the observed residuals r_i given the predicted residual $\hat{r_i}$ and the model parameters β. We eventually compare two models' difference in deviance, thus deviance can be simplified as:

$$D = -2F(\beta; r)$$

where $F(\beta; r) = \sum_{i=1}^{n} \ln f(r_i, \hat{r_i}, \beta)$ is the log-likelihood of the regression model. The negative twice log-likelihood ensures that larger deviance indicates poorer fit of the regression model. We examine changes in deviance [11] between a simple linear model and a polynomial models as:

$$\Delta D = D_{polynomial} - D_{linear}$$

where D_{linear} is the deviance of the baseline linear model and $D_{polynomial}$ the deviance of a polynomial model. A negative ΔD suggests that the polynomial model is capturing the distribution of residuals more efficiently. This analysis helps select a model that strikes a balance between capturing the distribution of the residuals and avoiding unnecessary complexity. When the difference in deviance between two models is minimal, the simpler model may be preferable due to its potential for improved generalisability.

Table 1: Deviance D and difference with baseline ΔD for Black and White races.

			degree 1	2	3	4	5	6
Chronic Illness	Signed Residuals	D	$1.2\cdot10^{1}$ $1.1\cdot10^{2}$	8.5 $1.1\cdot10^{2}$	8.1 $1.1\cdot10^{2}$	$-4.2\cdot10^{1}$ $9.9\cdot10^{1}$	$-7.7\cdot10^{1}$ $9.9\cdot10^{1}$	$-9.5\cdot10^{1}$ $9.7\cdot10^{1}$
		ΔD	0 0	-3.5 -2.7	-3.9 -3.0	$-5.4\cdot10^{1}$ $-1.1\cdot10^{1}$	$-9.0\cdot10^{1}$ $-1.1\cdot10^{1}$	$-1.1\cdot10^{2}$ $-1.3\cdot10^{1}$
	Absolute Residuals	D	$-4.4\cdot10^{1}$ $1.1\cdot10^{1}$	$-6.2\cdot10^{1}$ 7.7	$-1.0\cdot10^{2}$ 5.0	$-1.2\cdot10^{2}$ 4.7	$-1.3\cdot10^{2}$ 2.7	$-1.5\cdot10^{2}$ -5.6
		ΔD	0 0	$-1.8\cdot10^{1}$ -3.4	$-5.7\cdot10^{1}$ -6.1	$-7.5\cdot10^{1}$ -6.5	$-8.7\cdot10^{1}$ -8.5	$-1.1\cdot10^{2}$ $-1.7\cdot10^{1}$
Total Expenditure	Signed Residuals	D	$1.8\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.8\cdot10^{3}$
		ΔD	0 0	$-1.5\cdot10^{1}$ -9.0	$-3.0\cdot10^{1}$ $-1.9\cdot10^{1}$	$-4.9\cdot10^{1}$ $-2.8\cdot10^{1}$	$-7.3\cdot10^{1}$ $-4.1\cdot10^{1}$	$-9.7\cdot10^{1}$ $-6.4\cdot10^{1}$
	Absolute Residuals	D	$1.8\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.9\cdot10^{3}$	$1.7\cdot10^{3}$ $1.8\cdot10^{3}$	$1.7\cdot10^{3}$ $1.8\cdot10^{3}$	$1.7\cdot10^{3}$ $1.8\cdot10^{3}$
		ΔD	0 0	$-2.0\cdot10^{1}$ $-1.1\cdot10^{1}$	$-3.4\cdot10^{1}$ $-2.7\cdot10^{1}$	$-5.5\cdot10^{1}$ $-4.6\cdot10^{1}$	$-7.7\cdot10^{1}$ $-6.6\cdot10^{1}$	$-1.1\cdot10^{2}$ $-8.8\cdot10^{1}$

4 Results

Table 1 shows the deviance values, and the difference in deviance (Δ D) with to the baseline model (Degree 1), for both the Black and White races. Deviance descreases when increasing the polynomial degree, indicating an improvement in goodness of fit. However, the decrease in deviance for the Black race is smaller than for the White race (especially for the Chronic Illness case). Thus the relationship between polynomial degree and deviance reduction varies between races. It indicates a variability in the predictive accuracy of the regression models across different racial groups. Regarding difference in deviance with baseline model (ΔD), it consistently decreases as the polynomial degree increases. This indicates that the polynomial models are more accurate for modelling the residuals and their heteroscedasticity.

4.1 Signed Residuals

Chronic Illness Case: Table 2 shows the regression parameters (with their p-values) and MSE for polynomial models of degrees 1 to 6.

For the Black race, no model achieves statistical significance ($p<0.05$) for all their parameters β. All models have equivalent mean squared error (MSE), although models with degrees 4 to 6 having slightly lower MSE. With statistical significance for all parameters β except the intercept, and the lowest MSE, the model of degree 4 seems preferable. The non-significant intercept indicates that the model is less accurate when risk scores are close to 0.

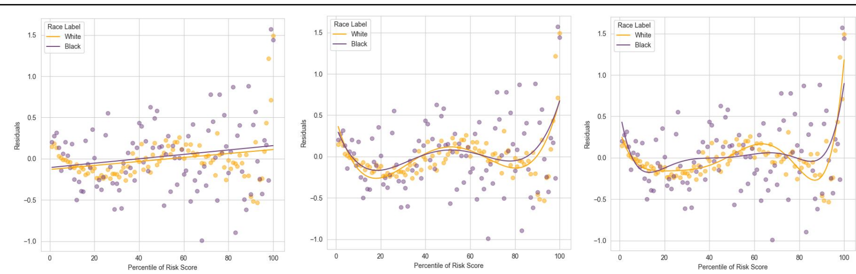
For the White race, MSE decreases as the model degree increases. It shows that more complex polynomial models are more accurate. Only the models of degree 4 and 6 achieve statistical significance for all parameters β. With degree 4, p-values are the lowest, and this is the preferred model. With degree 6, MSE is lower by 4.15% but this may not justify to increase the model complexity.

These results indicate that modelling heteroscedasticity may require different models for different populations, which is crucial for assessing fairness. The varying MSE and significance of parameters β across different polynomial degrees, and across races, underscore the importance of tailoring models to specific demographic groups. Thus, in practice, a one-size-fits-all approach to heteroscedasticity modelling may not be appropriate for fairness assessments.

Figure 4 shows the distribution of signed residuals, the baseline model, and the preferred models. Both races have outliers when risk scores are around the maximum value (100), which slightly skews the models. For the Black race, we observe a cone-shaped pattern of heteroscedasticity, where the range of residuals widens as risk scores increase. This pattern may be better modelled using absolute or squared residuals, since it is symmetrical around the zero line. For the White race, residuals are asymmetrical around the zero, showing a complex pattern that may not be identifiable using absolute or squared residuals. Thus modelling heteroscedasticity may require not only different model complexity (e.g., polynomial of degree 4 or 6), but also different types of residuals measurements (e.g., signed or absolute residuals).

Table 2: Modelling of signed residuals for the Chronic Illness case for Black and White races. Preferred model is in bold, parentheses show p-values.

degree	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE
1	$-1.3\cdot10^{-1}$ ($8.5\cdot10^{-3}$)	$2.4\cdot10^{-3}$ ($8.5\cdot10^{-3}$)						$6.7\cdot10^{-2}$
	$-1.1\cdot10^{-1}$ ($7.3\cdot10^{-2}$)	$2.7\cdot10^{-3}$ ($7.3\cdot10^{-2}$)						$1.8\cdot10^{-1}$
2	$-2.0\cdot10^{-2}$ ($2.6\cdot10^{-1}$)	$-4.0\cdot10^{-3}$ ($2.6\cdot10^{-1}$)	$6.4\cdot10^{-5}$ ($6.7\cdot10^{-2}$)					$6.6\cdot10^{-2}$
	$5.0\cdot10^{-2}$ ($2.7\cdot10^{-1}$)	$-6.6\cdot10^{-3}$ ($2.7\cdot10^{-1}$)	$9.1\cdot10^{-5}$ ($1.1\cdot10^{-1}$)					$1.8\cdot10^{-1}$
3	$-6.5\cdot10^{-2}$ ($9.0\cdot10^{-1}$)	$1.1\cdot10^{-3}$ ($9.0\cdot10^{-1}$)	$-6.3\cdot10^{-5}$ ($7.6\cdot10^{-1}$)	$8.3\cdot10^{-7}$ ($5.4\cdot10^{-1}$)				$6.6\cdot10^{-2}$
	$-1.5\cdot10^{-2}$ ($9.5\cdot10^{-1}$)	$9.9\cdot10^{-4}$ ($9.5\cdot10^{-1}$)	$-9.5\cdot10^{-5}$ ($7.8\cdot10^{-1}$)	$1.0\cdot10^{-6}$ ($5.8\cdot10^{-1}$)				$1.8\cdot10^{-1}$
4	**$4.6\cdot10^{-1}$ ($4.6\cdot10^{-5}$)**	**$-9.8\cdot10^{-2}$ ($1.3\cdot10^{-9}$)**	**$4.3\cdot10^{-3}$ ($5.7\cdot10^{-11}$)**	**$-6.6\cdot10^{-5}$ ($1.6\cdot10^{-11}$)**	**$3.3\cdot10^{-7}$ ($7.4\cdot10^{-12}$)**			**$4.1\cdot10^{-2}$**
	$3.6\cdot10^{-1}$ ($9.8\cdot10^{-2}$)	**$-7.0\cdot10^{-2}$ ($1.9\cdot10^{-2}$)**	**$3.1\cdot10^{-3}$ ($1.1\cdot10^{-2}$)**	**$-4.7\cdot10^{-5}$ ($8.3\cdot10^{-3}$)**	**$2.4\cdot10^{-7}$ ($6.4\cdot10^{-3}$)**			**$1.7e^{-1}$**
5	$4.2\cdot10^{-2}$ ($7.1\cdot10^{-1}$)	$1.8\cdot10^{-2}$ ($4.2\cdot10^{-1}$)	$-3.6\cdot10^{-3}$ ($8.6\cdot10^{-3}$)	$1.4\cdot10^{-4}$ ($5.6\cdot10^{-5}$)	$-2.0\cdot10^{-6}$ ($4.5\cdot10^{-7}$)	$9.1\cdot10^{-9}$ ($6.7\cdot10^{-9}$)		$2.9\cdot10^{-2}$
	$3.0\cdot10^{-1}$ ($2.6\cdot10^{-1}$)	$-5.4\cdot10^{-2}$ ($3.1\cdot10^{-1}$)	$1.9\cdot10^{-3}$ ($5.5\cdot10^{-1}$)	$-1.8\cdot10^{-5}$ ($8.3\cdot10^{-1}$)	$-8.9\cdot10^{-8}$ ($9.2\cdot10^{-1}$)	$1.3\cdot10^{-9}$ ($7.1\cdot10^{-1}$)		$1.7\cdot10^{-1}$
6	$3.6\cdot10^{-1}$ ($8.0\cdot10^{-3}$)	$-9.3\cdot10^{-2}$ ($5.8\cdot10^{-3}$)	$7.1\cdot10^{-3}$ ($1.3\cdot10^{-2}$)	$-2.8\cdot10^{-4}$ ($8.4\cdot10^{-3}$)	$5.7\cdot10^{-6}$ ($2.5\cdot10^{-3}$)	$-5.8\cdot10^{-8}$ ($4.3\cdot10^{-4}$)	$2.2\cdot10^{-10}$ ($5.4\cdot10^{-5}$)	$2.4\cdot10^{-2}$
	$5.8\cdot10^{-1}$ ($7.7\cdot10^{-2}$)	$-1.6\cdot10^{-1}$ ($7.1\cdot10^{-2}$)	$1.2\cdot10^{-2}$ ($1.1\cdot10^{-1}$)	$-4.1\cdot10^{-4}$ ($1.3\cdot10^{-1}$)	$7.1\cdot10^{-6}$ ($1.4\cdot10^{-1}$)	$-6.1\cdot10^{-8}$ ($1.4\cdot10^{-1}$)	$2.1\cdot10^{-10}$ ($1.3\cdot10^{-1}$)	$1.7\cdot10^{-1}$



(a) Degree 1 (b) Degree 4 (c) Degree 6

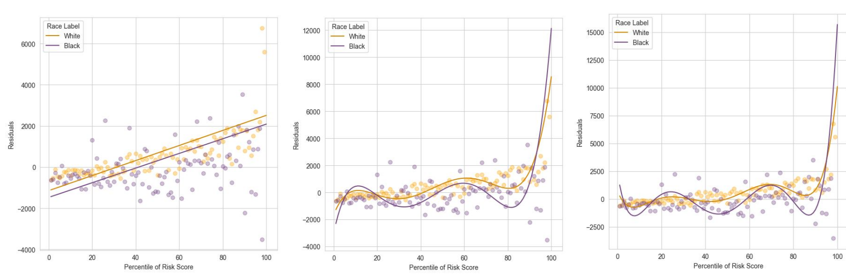
Fig. 4: Signed residuals vs risk score, for the Chronic Illness case.

Total Expenditure Case: Table 3 shows the regression parameters (with their p-values) and MSE for polynomial models of degrees 1 to 6.

For both the Black and White races, MSE significantly decreases as the polynomial degree increases. For the White race, polynomial degrees 1, 3 and 5 exhibit parameters β that are all statistically significant. The model with degree 5 has the lowest MSE (50% lower than than the baseline with degree 1). For the Black race, the only model where all parameters β are statistically significant is the baseline (degree 1). However, when disregarding the intercept, all the parameters β of models with degrees 2, 3, 5, and 6 are statistically significant. Since their MSEs are much lower than the baseline, polynomial models remain preferable. This suggests that estimating the intercept, e.g., estimating residuals for risk scores around 0, might be a recurrent issue with polynomial models.

Table 3: Modelling of signed residuals for the Medical Expenditure case for Black and White races. Preferred model is in bold, parentheses show p-values.

degree	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE
1	$-1.1\cdot10^{3}$	$3.7\cdot10^{2}$						$2.8\cdot10^{6}$
	$(9.1\cdot10^{-4})$	$(6.75\cdot10^{-9})$						
	$-1.5\cdot10^{3}$	$3.6\cdot10^{1}$						$1.2\cdot10^{7}$
	$(3.4\cdot10^{-2})$	$(3.3\cdot10^{-3})$						
2	$2.8\cdot10^{2}$	$-4.8\cdot10^{1}$	$8.3\cdot10^{-1}$					$2.4\cdot10^{6}$
	$(5.5\cdot10^{-1})$	$(3.0\cdot10^{-2})$	$(1.2\cdot10^{-4})$					
	$8.0\cdot10^{2}$	$-9.8\cdot10^{1}$	1.3					$1.1\cdot10^{7}$
	$(4.3\cdot10^{-1})$	$(3.4\cdot10^{-2})$	$(3.3\cdot10^{-3})$					
3	$-1.3\cdot10^{3}$	$1.4\cdot10^{2}$	-3.7	$3.0\cdot10^{-2}$				$2.1\cdot10^{6}$
	$(3.3\cdot10^{-2})$	$(9.2\cdot10^{-3})$	$(2.2\cdot10^{-3})$	$(1.8\cdot10^{-4})$				
	$-2.0\cdot10^{3}$	$2.2\cdot10^{2}$	-6.6	$5.2\cdot10^{-2}$				$9.8\cdot10^{6}$
	$(1.3\cdot10^{-1})$	$(4.6\cdot10^{-2})$	$(1.1\cdot10^{-2})$	$(2.1\cdot10^{-3})$				
4	$6.4\cdot10^{2}$	$-2.3\cdot10^{2}$	$1.3\cdot10^{1}$	$-2.2\cdot10^{-1}$	$1.2\cdot10^{-3}$			$1.8\cdot10^{6}$
	$(3.6\cdot10^{-1})$	$(1.7\cdot10^{-2})$	$(1.4\cdot10^{-3})$	$(2.0\cdot10^{-4})$	$(2.5\cdot10^{-5})$			
	$1.0\cdot10^{3}$	$-3.4\cdot10^{2}$	$1.8\cdot10^{1}$	$-3.3\cdot10^{-1}$	$1.9\cdot10^{-3}$			$9.1\cdot10^{6}$
	$(5.3\cdot10^{-1})$	$(1.2\cdot10^{-1})$	$(3.8\cdot10^{-2})$	$(1.2\cdot10^{-2})$	$(3.6\cdot10^{-3})$			
5	$\mathbf{-1.7}\cdot10^{3}$	$\mathbf{4.1}\cdot10^{2}$	$\mathbf{-3.1}\cdot10^{1}$	$\mathbf{9.2}\cdot10^{-2}$	$\mathbf{-1.1}\cdot10^{-2}$	$\mathbf{5.0}\cdot10^{-5}$		$\mathbf{1.4}\cdot10^{6}$
	$(\mathbf{3.5}\cdot10^{-2})$	$(\mathbf{9.1}\cdot10^{-3})$	$(\mathbf{1.2}\cdot10^{-3})$	$(\mathbf{1.4}\cdot10^{-4})$	$(\mathbf{1.7}\cdot10^{-5})$	$(\mathbf{2.0}\cdot10^{-6})$		
	$-3.0\cdot10^{3}$	$7.8\cdot10^{2}$	$-5.8\cdot10^{1}$	1.7	$-2.0\cdot10^{-2}$	$8.8\cdot10^{-5}$		$8.0\cdot10^{6}$
	$(1.1\cdot10^{-1})$	$(3.6\cdot10^{-2})$	$(1.1\cdot10^{-2})$	$(3.3\cdot10^{-3})$	$(1.1\cdot10^{-3})$	$(3.7\cdot10^{-4})$		
6	$7.2\cdot10^{2}$	$-4.9\cdot10^{2}$	$5.5\cdot10^{1}$	-2.5	$5.1\cdot10^{-2}$	$-4.9\cdot10^{-4}$	$2.0\cdot10^{-6}$	$1.1\cdot10^{6}$
	$(3.9\cdot10^{-1})$	$(2.9\cdot10^{-2})$	$(4.1\cdot10^{-3})$	$(6.3\cdot10^{-4})$	$(9.2\cdot10^{-5})$	$(1.3\cdot10^{-5})$	$(2.0\cdot10^{-6})$	
	$\mathbf{2.4}\cdot10^{3}$	$\mathbf{-1.3}\cdot10^{3}$	$\mathbf{1.4}\cdot10^{2}$	$\mathbf{-6.1}$	$\mathbf{1.2}\cdot10^{-1}$	$\mathbf{-1.2}\cdot10^{-3}$	$\mathbf{4.0}\cdot10^{-6}$	$\mathbf{6.5}\cdot10^{6}$
	$(\mathbf{2.4}\cdot10^{-1})$	$(\mathbf{2.0}\cdot10^{-2})$	$(\mathbf{2.9}\cdot10^{-3})$	$(\mathbf{5.2}\cdot10^{-4})$	$(\mathbf{1.1}\cdot10^{-4})$	$(\mathbf{2.4}\cdot10^{-5})$	$(\mathbf{6.0}\cdot10^{-6})$	



(a) Degree 1 (b) Degree 5 (c) Degree 6

Fig. 5: Signed residuals vs risk score, for the Medical Expenditure case.

Figure 5 shows the distribution of signed residuals, the baseline model, and the preferred models. For visual clarity, 3 outliers with high risk scores are removed from these graphs[1] but were not removed when fitting the polynomial models. These outliers skew the models, especially the polynomial models.

For both races, we observe a cone-shaped pattern of heteroscedasticity, where the range of residuals widens as risk scores increase. However, with high risk scores, residuals tend to be negative for the Black race (i.e., risk scores are under-estimated), and positive for the White race (i.e., risk scores are over-estimated). Furthermore, for all risk scores, the range of residuals is larger for the Black

[1] (1) Risk score = 99, residuals ≈ 1 400, race = Black. (2) Risk score = 100, residuals ≈ 1 700, race = White, (3) Risk score = 100, residuals ≈ 3 000, race = Black.

race. These discrepancies create fairness issues. Such cone-shaped patterns of heteroscedasticity might be better modelled using absolute residuals, which we investigate next.

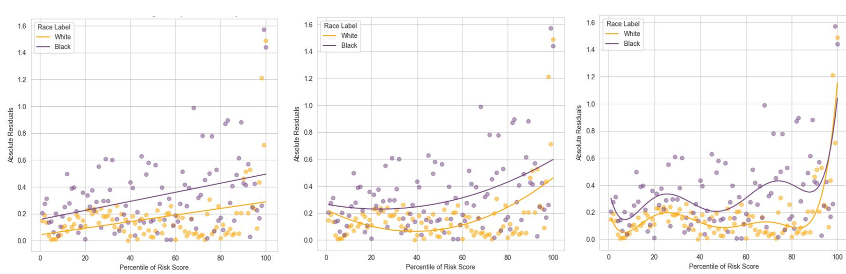
4.2 Absolute Residuals

Chronic Illness Case: Table 4 shows the regression parameters (with their p-values) and MSE for polynomial models of degrees 1 to 6.

For the White race models with degree 2 and 6 both have all their coefficient statistically significant. However, degree 6 has lower MSE (57.6% lower than degree 2). For the Black race, only the model with degree 1 has all its parameters β statistically significant. However, this model has the highest MSE. But other models offer only a slightly lower MSE (e.g., 10% lower at most with degree 6).

Table 4: Modelling of absolute residuals for the Chronic Illness case for Black and White races. Preferred model is in bold, parentheses show p-values.

degree	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE
1	$4.3\cdot10^{-2}$	$2.5\cdot10^{-3}$						$3.9\cdot10^{-2}$
	$(2.8\cdot10^{-1})$	$(4.9\cdot10^{-4})$						
	$\mathbf{1.6}\cdot10^{-1}$	$\mathbf{3.4}\cdot10^{-3}$						
	$(\mathbf{3.2}\cdot10^{-3})$	$(\mathbf{2.8}\cdot10^{-4})$						$\mathbf{6.7}\cdot10^{-2}$
2	$2.3\cdot10^{-1}$	$-8.3\cdot10^{-3}$	$1.1\cdot10^{-4}$					$3.3\cdot10^{-2}$
	$(8.8\cdot10^{-5})$	$(1.4\cdot10^{-3})$	$(2.8\cdot10^{-5})$					
	$2.7\cdot10^{-1}$	$-3.0\cdot10^{-3}$	$6.3\cdot10^{-5}$					$6.5\cdot10^{-2}$
	$(9.8\cdot10^{-4})$	$(4.0\cdot10^{-1})$	$(6.9\cdot10^{-2})$					
3	$-5.7\cdot10^{-2}$	$2.4\cdot10^{-2}$	$-7.0\cdot10^{-4}$	$5.0\cdot10^{-6}$				$2.2\cdot10^{-2}$
	$(3.6\cdot10^{-1})$	(1.2e-5)	$(9.8\cdot10^{-8})$	$(1.2\cdot10^{-9})$				
	$1.5\cdot10^{-1}$	$1.0\cdot10^{-2}$	$-2.7\cdot10^{-4}$	$2.0\cdot10^{-6}$				$6.4\cdot10^{-2}$
	$(1.6\cdot10^{-1})$	$(2.5\cdot10^{-1})$	$(2.0\cdot10^{-1})$	$(1.1\cdot10^{-1})$				
4	$1.4\cdot10^{-1}$	$-1.3\cdot10^{-2}$	$9.6\cdot10^{-4}$	$-2.0\cdot10^{-5}$	$1.3\cdot10^{-7}$			$1.9\cdot10^{-2}$
	$(5.4\cdot10^{-2})$	$(1.9\cdot10^{-1})$	$(1.7\cdot10^{-2})$	$(9.0\cdot10^{-4})$	$(3.2\cdot10^{-5})$			
	$2.0\cdot10^{-1}$	$9.5\cdot10^{-4}$	$1.5\cdot10^{-4}$	$-4.2\cdot10^{-6}$	$3.1\cdot10^{-8}$			$6.5\cdot10^{-2}$
	$(1.4\cdot10^{-1})$	$(9.6\cdot10^{-1})$	$(8.4\cdot10^{-1})$	$(7.0\cdot10^{-1})$	$(5.6\cdot10^{-1})$			
5	$-2.7\cdot10^{-2}$	$3.4\cdot10^{-2}$	$-2.2\cdot10^{-3}$	$6.3\cdot10^{-5}$	$-8.0\cdot10^{-7}$	$3.7\cdot10^{-9}$		$1.7\cdot10^{-2}$
	$(7.6\cdot10^{-1})$	$(4.9\cdot10^{-2})$	$(3.2\cdot10^{-2})$	$(1.5\cdot10^{-2})$	$(4.8\cdot10^{-3})$	$(1.1\cdot10^{-3})$		
	$6.5\cdot10^{-2}$	$3.8\cdot10^{-2}$	$-2.4\cdot10^{-3}$	$6.2\cdot10^{-5}$	$-7.1\cdot10^{-7}$	$2.9\cdot10^{-9}$		$6.4\cdot10^{-2}$
	$(6.9\cdot10^{-1})$	$(2.4\cdot10^{-1})$	$(2.3\cdot10^{-1})$	$(2.1\cdot10^{-1})$	$(1.9\cdot10^{-1})$	$(1.7\cdot10^{-1})$		
6	$\mathbf{2.1}\cdot10^{-1}$	$\mathbf{-5.5}\cdot10^{-2}$	$\mathbf{6.3}\cdot10^{-3}$	$\mathbf{-2.7}\cdot10^{-4}$	$\mathbf{5.3}\cdot10^{-6}$	$\mathbf{-5.0}\cdot10^{-8}$	$\mathbf{1.8}\cdot10^{-10}$	$\mathbf{1.4}\cdot10^{-2}$
	$(\mathbf{3.0}\cdot10^{-2})$	$(\mathbf{3.2}\cdot10^{-2})$	$(\mathbf{4.1}\cdot10^{-3})$	$(\mathbf{9.3}\cdot10^{-4})$	$(\mathbf{2.7}\cdot10^{-4})$	$(\mathbf{8.7}\cdot10^{-5})$	$(\mathbf{2.7}\cdot10^{-5})$	
	$3.7\cdot10^{-1}$	$-7.8\cdot10^{-2}$	$8.8\cdot10^{-3}$	$-3.7\cdot10^{-4}$	$7.4\cdot10^{-6}$	$-6.7\cdot10^{-8}$	$2.3\cdot10^{-10}$	$6.0\cdot10^{-2}$
	$(5.7\cdot10^{-2})$	$(1.4\cdot10^{-1})$	$(4.8\cdot10^{-2})$	$(2.2\cdot10^{-2})$	$(1.3\cdot10^{-2})$	$(8.1\cdot10^{-3})$	$(5.6\cdot10^{-3})$	



(a) Degree 1 (b) Degree 2 (c) Degree 6

Fig. 6: Absolute residuals vs risk score, for the Chronic Illness case.

Figure 6 shows the distribution of signed residuals, the baseline model, and the preferred models. There are outliers around the maximum risk score, which skews the models. For the Black race, the absolute residuals have a wider range that consistently exceed those of the White race, and increases with the risk scores. In contrast, the White race shows less variability in the residuals.

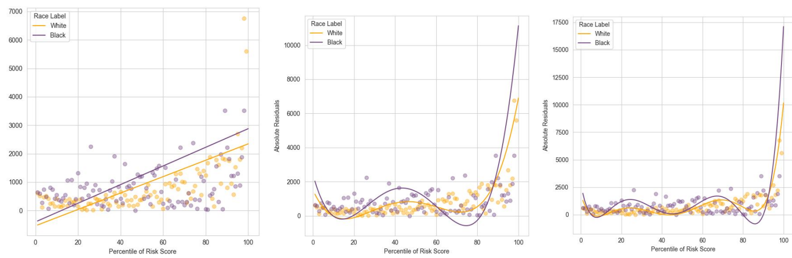
Medical Expenditure case: Table 5 shows the regression parameters (with their p-values) and MSE for polynomial models of degrees 1 to 6.

The parameters β for degrees 2, 4, and 6 are all statistically significant for the White race, with degree 6's MSE being considerably lower than degree 1's MSE (44.4% lower). For the Black race, only degree 2 has all its parameters β statistically significant. For degree 6, only the intercept β_0 is not statistically significant. The MSE of degree 6 is much lower than the MSE of degree 2 (51.6% lower), thus it is also a well-performing model.

Figure 7 shows the distribution of signed residuals, the baseline model, and the preferred models. For the Black race, the spread of residuals is notably wider than for the White race. For both races, residuals increase with the risk scores, and this increase is more pronounced around the maximum risk scores. There are outliers for high risk scores, which significantly skew the models, especially for polynomial models, and for the Black race.

Table 5: Modelling of absolute residuals for the Medical Expenditure case for Black and White races. Preferred model is in bold, parentheses show p-values.

degree	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE
1	$-5.4\cdot10^{2}$ $(1.2\cdot10^{-1})$	$2.8\cdot10^{1}$ $(3.0\cdot10^{-6})$						$2.8\cdot10^{6}$
	$-4.0\cdot10^{2}$ $(5.4\cdot10^{-1})$	$3.3\cdot10^{1}$ $(4.0\cdot10^{-3})$						$1.0\cdot10^{7}$
2	$1.1\cdot10^{3}$ $(2.2\cdot10^{-2})$	$-6.7\cdot10^{1}$ $(1.2\cdot10^{-5})$	$9.5\cdot10^{-1}$ $(2.3\cdot10^{-3})$					$2.4\cdot10^{6}$
	2.0$\cdot10^{3}$ (**4.0**$\cdot10^{-2}$)	**-1.1**$\cdot10^{2}$ (**1.6**$\cdot10^{-2}$)	**1.4** (**1.2**$\cdot10^{-3}$)					**9.3**$\cdot10^{6}$
3	$-4.7\cdot10^{2}$ $(4.3\cdot10^{-1})$	$1.1\cdot10^{2}$ $(3.3\cdot10^{-3})$	-3.5 $(2.7\cdot10^{-2})$	$2.9\cdot10^{-2}$ $(1.9\cdot10^{-4})$				$2.1\cdot10^{6}$
	$-1.4\cdot10^{3}$ $(2.5\cdot10^{-1})$	$2.8\cdot10^{2}$ $(6.7\cdot10^{-3})$	-8.1 $(6.9\cdot10^{-4})$	$6.2\cdot10^{-2}$ $(6.9\cdot10^{-5})$				$7.9\cdot10^{6}$
4	**1.5**$\cdot10^{3}$ (**3.0**$\cdot10^{-2}$)	**-2.6**$\cdot10^{2}$ (**7.4**$\cdot10^{-4}$)	**1.3**$\cdot10^{1}$ (**6.1**$\cdot10^{-3}$)	**-2.3**$\cdot10^{-1}$ (**1.1**$\cdot10^{-4}$)	**1.3**$\cdot10^{-3}$ (**1.3**$\cdot10^{-5}$)			**1.7**$\cdot10^{6}$
	$2.5\cdot10^{3}$ $(7.7\cdot10^{-2})$	$-4.4\cdot10^{2}$ $(1.9\cdot10^{-2})$	$2.4\cdot10^{1}$ $(2.0\cdot10^{-3})$	$-4.3\cdot10^{-1}$ $(2.1\cdot10^{-4})$	$2.4\cdot10^{-3}$ $(2.4\cdot10^{-5})$			$6.7\cdot10^{6}$
5	$-6.6\cdot10^{2}$ $(3.9\cdot10^{-1})$	$3.4\cdot10^{2}$ $(2.9\cdot10^{-3})$	$-2.8\cdot10^{1}$ $(2.6\cdot10^{-2})$	$8.5\cdot10^{-1}$ $(3.3\cdot10^{-4})$	$-1.1\cdot10^{-2}$ $(4.2\cdot10^{-5})$	$4.7\cdot10^{-5}$ $(5.0\cdot10^{-6})$		$1.4\cdot10^{6}$
	$-1.6\cdot10^{3}$ $(2.9\cdot10^{-1})$	$6.9\cdot10^{2}$ $(2.6\cdot10^{-2})$	$-5.4\cdot10^{1}$ $(4.6\cdot10^{-3})$	1.6 $(7.7\cdot10^{-4})$	$-2.0\cdot10^{-2}$ $(1.2\cdot10^{-4})$	$8.9\cdot10^{-5}$ $(1.8\cdot10^{-5})$		$5.5\cdot10^{6}$
6	$1.9\cdot10^{3}$ $(2.5\cdot10^{-2})$	$-6.1\cdot10^{2}$ $(6.3\cdot10^{-3})$	$6.3\cdot10^{1}$ $(8.9\cdot10^{-4})$	-2.7 $(1.2\cdot10^{-4})$	$5.5\cdot10^{-2}$ $(1.7\cdot10^{-5})$	$-5.3\cdot10^{-4}$ $(2.3\cdot10^{-6})$	$2.0\cdot10^{-6}$ $(3.2\cdot10^{-7})$	$1.0\cdot10^{6}$
	$2.8\cdot10^{3}$ $(9.5\cdot10^{-2})$	$-1.0\cdot10^{3}$ $(2.9\cdot10^{-2})$	$1.1\cdot10^{2}$ $(5.2\cdot10^{-3})$	-4.8 $(1.0\cdot10^{-3})$	$9.7\cdot10^{-2}$ $(2.1\cdot10^{-4})$	$-9.3\cdot10^{-4}$ $(4.0\cdot10^{-5})$	$3.2\cdot10^{-6}$ $(8.0\cdot10^{-6})$	$4.5\cdot10^{6}$



(a) Degree 1 (b) Degree 4 (c) Degree 6

Fig. 7: Absolute residuals vs risk score, for the Medical Expenditure case.

4.3 Preferred Model

Table 6 recaps the evaluation of the preferred models for all use cases. We can observe that the models' performance is worse for the Black race compared to the White race, except for the Chronic Illness case with signed residuals. Thus modelling the residuals is generally more uncertain for the Black race.

Table 6: Preferred models for each use case and Black and White races.

		Degree	MSE	Deviance	ΔD	$p_\beta < 0.05$
Chronic Illness	Signed Residuals	4	$4.1 \cdot 10^{-2}$	$-4.2 \cdot 10^{1}$	$-5.4 \cdot 10^{1}$	Yes
		4	$1.7 \cdot 10^{-1}$	$9.9 \cdot 10^{1}$	$-1.1 \cdot 10^{1}$	All but β_0
	Absolute Residuals	6	$1.4 \cdot 10^{-2}$	$-1.5 \cdot 10^{2}$	$-1.1 \cdot 10^{2}$	Yes
		1	$6.7 \cdot 10^{-2}$	$1.1 \cdot 10^{1}$	0	Yes
Total Expenditure	Signed Residuals	5	$1.4 \cdot 10^{6}$	$1.7 \cdot 10^{3}$	$-7.3 \cdot 10^{1}$	Yes
		6	$6.5 \cdot 10^{6}$	$1.8 \cdot 10^{3}$	$-6.4 \cdot 10^{1}$	All but β_0
	Absolute Residuals	4	$1.7 \cdot 10^{6}$	$1.7 \cdot 10^{3}$	$-5.5 \cdot 10^{1}$	Yes
		2	$9.3 \cdot 10^{6}$	$1.9 \cdot 10^{3}$	$-1.1 \cdot 10^{1}$	Yes

Signed vs absolute residuals: For the Chronic Illness case, with the Black race, the linear model using absolute residuals is statistically more reliable than the one using signed residuals (deviance, p-values of β parameters). However, its MSE is much higher. This is consistent with our assumption that the Back race's cone-shaped pattern of heteroscedasticity, horizontally centered around the zero line (Fig. 4), can be modelled with simple linear models and absolute residuals.

For the Medical Expenditure case, cone-shaped pattern of heteroscedasticity were also observed (Fig. 5). For these, the preferred models using absolute residuals are less complex than those using signed residuals (e.g., degree 2 vs 6 for the Black race). The deviance is similar with signed or absolute residuals, but MSE is lower with signed residuals. These cone-shaped patterns were not well-centered around the horizontal zero line, which explains why using signed residuals is preferable.

We conclude that using signed residuals is preferable to absolute residuals if the pattern of heteroscedasticity is not symmetrical around the horizontal zero line. However, with signed residuals the intercept parameter β_0 might be less reliable (e.g., not statistically significant), e.g., impeding the modelling of residuals and heteroscedasticity for risk scores around zero.

Linear vs polynomial models: A simple linear model is preferable only for one case out of eight. For the remaining cases, compared to simple linear models, MSE is significantly lower with polynomial models, and the reduction of deviance ΔD is noticeable (especially for the White race). The MSE consistently decreased as model complexity increased from degree 1 to 6. We conclude that using polynomial regression enhances the ability to capture the distribution of residuals and the heteroscedasticity.

4.4 Discussion

What is the advantage of polynomial regression in complex cases of heteroscedasticity? Polynomial models offers an important advantage over simple linear models in capturing curvature within residuals. This curvature allows for more flexibility when modelling heteroscedasticity where the distribution of residuals is not linear or constant, an ability that simple linear regression, limited to straight lines, does not have. Our finding shows that relying solely on simple linear regression for modelling heteroscedasticity may be insufficient. This approach requires to tune the model complexity (i.e., the polynomial degrees). Beware that even if non-significant parameters are found for a given degree, and significant parameters may yet be found for a higher degree higher (e.g., Table 4). However, the model complexity may not be increased indefinitely as this may lead to overfitting.

We do not claim that polynomial models are the best approach to modelling heteroscedasticity, and other forms parametric regressions must be investigated in future work. Non-parametric models may also be more accurate for modelling residuals [15, 9]. However, parametric models offer interpretable means for identifying heteroscedasticity, i.e., through interpreting their parameters.

How to interpret polynomial models to determine fairness issues due to heteroscedasticity? Heteroscedasticity can be identified when at least one parameters $\beta_1, ..., \beta_n$ is non-zero and statistically significant, wether the model uses absolute or signed residuals.

For social groups defined by a categorical sensitive feature (e.g., race or gender), a polynomial model must be fitted to each social group[2]. If all social groups are impacted by heteroscedasticity, it remains challenging the assess whether one group is more impacted than the other. To address this issue, the β parameters could be compared, e.g., fairness issues arise if β_n is higher for a protected group compared to another group. However, this approach is difficult to apply if the heteroscedasticity models have different complexity for each group, or if parameters β_n are not significant for each groups.

[2] Note that fairness issues also arise if parameter β_0 is different for each social group.

For social groups defined by a numeric sensitive feature (e.g., age or income), a single polynomial model can be fitted using the sensitive feature as $\mathbf{x}$ in equations (3-4). Once heteroscedasticity is identified, it remains challenging to identify for which range of the sensitive feature is heteroscedasticity present, or of higher magnitude.

Hence approaches to explore in future work include (1) studying the difference between residuals models for each groups, and (2) identifying the range of values $\mathbf{x}$ for which residuals is higher, and the magnitude of this heteroscedastic increase in residuals (e.g., fairness issues arise if within the same range of $\mathbf{x}$ values, residuals are higher for a specific social group), and (3) multivariate models that account for feature interactions, and intersectionality [17].

What is the difference in effectiveness between using signed residuals and absolute residuals for detecting heteroscedasticity? This is complex to determine, and the impact of outliers may be different.

Using absolute residuals can reduce the cumulative prediction errors of polynomial models, and give a more accurate indication of heteroscedasticity (e.g., with lower MSEs and β parameters' p-values). This especially applies to patterns of heteroscedasticity that are symmetrical around the horizontal zero line, e.g., when using signed residuals.

However, the models using signed residuals performed systematically better in Medical Expenditure case, and offered much lower MSE for the Black race in the Chronic Illness case. Signed residuals preserve information about to over- or under-estimation. The identification of systematic biases in AI models can depend on this information. Signed residuals also offer crucial information to assess the practical impacts of heteroscedasticity, as over- or under-estimations can have opposite consequences.

5 Conclusion

This study highlights the advantages and limitations of polynomial regression models in detecting and modelling heteroscedasticity, compared to simple linear methods. We demonstrate the polynomial model's ability to capture patterns with curvature. This is a distinct advantage compared to linear models which are limited to straight lines. This study also underscores the importance of using multiple evaluation methods, such as β parameters' p-values, deviance, mean squared error (MSE), and data visualization, to support model selection. This study also underlines the importance of considering diverse modelling approaches tailored to each protected group, as there is no one-size-fits-all approach. While the advantages of polynomial regression are demonstrated, challenges remain with comparing the magnitude of heteroscedasticity across social groups. Hence future work must explore approaches to compare heteroscedasticity across social group, e.g., with parametric and non-parametric approaches.

References

1. Agarwal, A., Dudík, M., Wu, Z.S.: Fair regression: Quantitative definitions and reduction-based algorithms. In: International Conference on Machine Learning. pp. 120–129. PMLR (2019)
2. Breusch, T.S., Pagan, A.R.: A simple test for heteroscedasticity and random coefficient variation. Econometrica **47**(5), 1287 (September 1979). https://doi.org/10.2307/1911963
3. Cook, R.D., Weisberg, S.: and university of minnesota school of statistics department of applied statistics, "diagnostics for heteroscedasticity in regression. Technical Report 405, " (May 1982)
4. Glejser, H.: A new test for heteroskedasticity. Journal of the American Statistical Association **64**(325), 316–323 (March 1969). https://doi.org/10.1080/01621459.1969.10500976
5. Goldfeld, S.M., Quandt, R.E.: Some tests for homoscedasticity. Journal of the American Statistical Association **60**(310), 539–547 (June 1965). https://doi.org/10.1080/01621459.1965.10480811
6. Hsiao, C.W., Chan, Y.C., Lee, M.Y., Lu, H.P.: Heteroscedasticity and precise estimation model approach for complex financial time-series data: An example of taiwan stock index futures before and during covid-19. Mathematics **9**(21), 2719 (October 2021). https://doi.org/10.3390/math9212719
7. J. L. A. Mattu Lauren Kirchner, S.: How we analyzed the compas recidivism algorithm (2020), https://www.propublica.org/Article/How-We-Analyzed-the-Compas-Recidivism-Algorithm, proPublica, Feb. 29
8. Klein, A.G.: 1. C. Gerhard, R. D. Büchner, S. Diestel, and K. Schermelleh-Engel, "The detection of heteroscedasticity in regression models for psychological data, " Psychological Test and Assessment Modeling (December 2016)
9. Le, Q.V., Smola, A.J., Canu, S.: Heteroscedastic gaussian process regression. In: Proceedings of the 22nd international conference on Machine learning. pp. 489–496 (2005)
10. MacKinnon, J.G., White, H.: Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties (1983), http://hdl.handle.net/10419/189084
11. McCullagh, P., Nelder, J.A.: Generalized linear models (2019)
12. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6472), 447–453 (October 2019), https://gitlab.com/labsysmed/dissecting-bias
13. Rosopa, P.J., University, C., Schaffer, M.M., Walmart, A.N.S., University, W.K.: Managing heteroscedasticity in general linear models (2013)
14. Skeem, J.: University of california (June 2016), https://ssrn.com/abstract=2687339, berkeley, C. T. Lowenkamp, and Administrative Office, U. S. Courts, "Risk, Race, & Recidivism: Predictive bias and Disparate impact, " journal-article
15. Sluijterman, L., Cator, E., Heskes, T.: Optimal training of mean variance estimation neural networks. Neurocomputing **597**, 127929 (2024). https://doi.org/https://doi.org/10.1016/j.neucom.2024.127929, https://www.sciencedirect.com/science/article/pii/S0925231224007008
16. Su, L., Zhao, Y., Yan, T.: Two-stage method based on local polynomial fitting for a linear heteroscedastic regression model and its application in economics. Discrete Dynamics in Nature and Society **2012**, 1–17 (January 2012). https://doi.org/10.1155/2012/696927

17. Wang, A., Ramaswamy, V.V., Russakovsky, O.: Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 336–349 (2022)