Robust Inference for ASR Subgroup Disparities via Negative Binomial Bootstrapping

Anonymous ACL submission

Abstract

Standard statistical methods for ASR bias evaluation, from simple WER comparisons to mixed-effects Poisson regression, routinely violate distributional assumptions or depend on speaker metadata unavailable in privacy-sensitive settings. We demonstrate that error counts such as those used to measure ASR model performance exhibit significant overdispersion, making Poisson regression poorly calibrated for inference, and propose Negative Binomial regression (NB2) as a principled alternative. We further show that by combining NB2 with bootstrapping, Weighted Effect Coding, and Bonferroni–Holm corrections, we produce better-calibrated confidence intervals and p-values for all group effects, including the reference category. Simulations across multiple misspecification scenarios confirm that the bootstrapping method showed the best robustness among the evaluated methods, and we validate the methodology on an imbalanced Welsh ASR dataset.

1 Introduction

Word error rates (WER) are commonly used in automatic speech recognition (ASR) literature, and the absolute difference between WERs for different subgroups (e.g., gender, dialect, age, etc.) are often taken at face value without any formal statistical inference tests. In addition, this approach does not quantify the uncertainty or effects of speaker attributes robustly, especially when multiple attributes are available.

Methods such as paired bootstrap tests and permutation tests are designed to compare two models (Dror et al., 2018), and are not well suited to this setting. These methods are unable to account for multivariable effects, including not only the speaker attributes of interest, but also effects caused by other factors such as noise, duration, etc. When using smaller evaluation datasets, this is especially problematic because we are often dealing with small and imbalanced sample sizes across groups where the effect of these additional factors can distort the results significantly. As a consequence, it is difficult to determine whether the performance differences are reproducible and reliable, or whether the observed differences arise from sampling variability.

There have been attempts (Liu et al., 2022; Jahan et al., 2023, 2025) at improving the status quo using Poisson regression. While it does handle the problems outlined above, as we will show in this paper, Poisson regression faces significant obstacles when applied to ASR data, rendering its results unreliable. In addition, these methods often rely upon speaker metadata to handle overdispersion, which might not always be available.

In this paper, we show that ASR data often contain significant overdispersion, and we propose negative binomial regression (NB2) as an improvement upon the current Poisson regression methodology. We also propose a method combining NB2, bootstrapping, weighted effect coding (WEC) and Bonferroni–Holm corrections, which in some scenarios offers benefits over traditional regression methods. Notably, it is more robust against misspecification, allows for family-wise error rate (FWER) correction, and provides confidence intervals and p-values for all effects, including the reference. We analyse the coverage and robustness of these methods through simulations, before discussing the strengths, weaknesses, and limitations of these methods. We also include a case study on a smaller imbalanced Welsh ASR dataset to show the methodology in action.

1.1 Contributions

Through this paper:

- We show through simulations the miscalibration of commonly used regression methods

when applied to realistic ASR data.

- We propose improvements to the status quo that ensure that statistical hypothesis tests for bias analysis are more robustly calibrated and more reliable.
- We demonstrate that bootstrapping yields reliable p-values and confidence intervals for all variables, including the reference variable. This enables post-hoc analyses with family-wise error corrections to determine statistically significant categorical effects.
- We show that the dispersion ratio is a sensitive diagnostic for data quality issues in ASR output, including hallucination loops and other model generation artefacts, and demonstrate how it can be used as a first-pass validation step before inference.
- We apply our improved methodology on an existing small imbalanced Welsh dataset.

2 Background and Related Work

2.1 Bias in ASR systems

Bias, or more specifically systematic performance differences in ASR systems caused by speaker attributes such as gender, accent, and age, has been well documented (Ngueajio and Washington, 2022).

Tatman (2017) demonstrates that both accent and gender significantly affect per-speaker WER, using linear mixed-effects models with speaker and year as random effects. Koenecke et al. (2020) provided large-scale evidence of racial disparities in ASR, evaluating five commercial systems across a matched corpus of utterances from white and African American Vernacular English (AAVE) speakers. They found that error rates for black speakers are approximately twice those for white speakers across all five systems, and further demonstrate that this gap is consistent across identical phrases, pointing to acoustic rather than language model deficiencies as the primary cause. Fückner et al. (2023) demonstrated similar performance disparities in Dutch ASR systems across speaker subgroups, illustrating that bias is not limited to English-language systems.

Together, these works establish that demographic bias in ASR is a well-documented and practically significant problem, while also illustrating a broader pattern: effect sizes are often large enough to be convincing on their own, regardless of which statistical tests are used.

2.2 Statistical practices in NLP / ASR evaluation

While prior work has established the existence of bias in ASR systems, the statistical methodologies used to quantify these effects are often underdeveloped. Dror et al. (2018) and von Werra et al. (2022) both highlight the broader lack of rigorous statistical testing in NLP literature.

Resampling-based methods such as bootstrap tests and permutation tests have seen some use (Dror et al., 2018), and while there has been work done to improve upon these methods (e.g. Liu and Peng (2020)), they are primarily designed for pairwise model comparison and do not naturally extend to multivariable settings where multiple interacting factors (e.g., dialect, gender, noise conditions) must be analysed jointly.

This limitation is evident in ASR bias studies specifically. Analyses such as the study by Koenecke et al. (2020) compare results between geographic subgroups without formal statistical testing, leaving open the question of whether differences between closely performing groups are statistically reliable.

A related issue is that distributional assumptions in ASR evaluation are rarely explicitly validated. Linear mixed-effects regression as used by Tatman (2017) assumes normally distributed residuals for valid statistic inference, an assumption frequently violated in ASR data where WER distributions exhibit zero-inflation, heavy right tails, and clustering at rational fractions due to short utterances. While speaker-level aggregation can partially mitigate these issues, this is not guaranteed with small sample sizes.

In both these cases the scale of disparities is often large enough that formal inference is unlikely to change headline conclusions.

2.3 Statistical Modelling Approaches for ASR Evaluation

To address these limitations, several studies have proposed more principled statistical approaches for ASR evaluation. Tatman (2017) introduced linear mixed-effects models to account for speaker and temporal variability, representing an early step toward rigorous inference. The most notable subsequent advance is the mixed-effects Poisson regression framework by Liu et al. (2022).

Compared to the linear mixed-effect models used by Tatman (2017), Poisson is a more principled fit for modelling ASR error counts, as it naturally handles the discrete, non-negative nature of the data (though, as we show in Section 4.1.1, it is not without limitations). Liu et al. (2022) recognise that overdispersion might be present and attempt to address it through speaker-level random effects; however, this approach has two significant weaknesses that limit its general applicability. First, speaker-level random effects can absorb some excess variance but do not guarantee that overdispersion is fully eliminated, and the residual overdispersion goes unaddressed. Overdispersion means that the variance in the data greater than expected. In the case of Poisson, the variance is modelled independently of the mean, hence if the data is heterogenous overdispersion can occur. Second, speaker IDs are frequently withheld in real-world datasets for privacy reasons, making the method inapplicable in precisely the settings where fairness evaluation is most needed. While the work by Liu et al. (2022) represents a meaningful step forward in statistically rigorous ASR fairness evaluation and is being applied by others (Jahan et al., 2023, 2025), these limitations motivate the improvements we propose in this paper.

Overall, existing approaches either rely on descriptive statistics, assume distributional properties that are often violated in ASR data, or require metadata that is not always available. Consequently, there is a need for a statistically robust and flexible methodology that can handle overdispersion, imbalance, and uncertainty in a principled manner.

3 Methodology

3.1 Baseline: Poisson Regression

The most commonly used metrics for evaluating ASR systems are error-rate based, such as character error rate (CER) and word error rate (WER). While these metrics are expressed as ratios, they are derived from underlying counts of recognition errors relative to the number of reference tokens.

To get reliable hypothesis testing, we need a statistical model of the metric. Count models such as Poisson regression are commonly used to model rates when the underlying data consist of event counts observed with varying exposure. In the context of ASR evaluation, the number of recognition errors can be treated as a count variable, while the number of words in the reference transcript represents the exposure.

We therefore model the number of errors in utterance i as a Poisson random variable:

$$\text{errors}_i \sim \text{Poisson}(\lambda_i) \tag{1}$$

where $\lambda_i > 0$ is the expected number of errors in utterance i.

Because longer utterances naturally accumulate more errors, we cannot model the raw count λ_i directly without accounting for the utterance length (words_i, i.e. number of words in reference i). Instead, our goal is to model the expected error rate ($\frac{\lambda_i}{\text{words}_i}$) using a standard log-linear relationship based on the speaker attributes K:

$$\log(\frac{\lambda_i}{\text{words}_i}) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} \tag{2}$$

where β_0 is the intercept, β_k is the fixed-effect coefficient for attribute k, and x_{ik} is an indicator variable representing the presence of attribute k for utterance i.

Using the logarithmic property $\log(\frac{A}{B}) = \log(A) - \log(B)$, we can isolate $\log(\lambda_i)$ on the left side of the equation:

$$\log(\lambda_i) - \log(\text{words}_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} \tag{3}$$

Adding $\log(\text{words}_i)$ to both sides yields the standard generalized linear model formulation:

$$\log(\lambda_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \log(\text{words}_i) \tag{4}$$

This formulation naturally models the expected error rate while accounting for varying utterance lengths and the non-negative, discrete nature of error counts. While this model serves as a foundational starting point for our methodology, we will extend it to our final model in the following section.

3.1.1 Weighted Effect Coding (WEC)

In highly imbalanced datasets, standard effect coding estimates deviations from the unweighted mean of group effects rather than the dataset-wide mean. However, ASR datasets, especially those for lower-resourced languages, often exhibit significant group imbalance (e.g. see Table 1). In addition, the regression model uses one of the accents as a reference and this is accent is omitted from the output of the model.

To ensure that the intercept represents the actual population mean, we employ Weighted Effect Coding (WEC) (te Grotenhuis et al., 2017). WEC corrects the imbalance in our dataset by weighting coefficients by group sample size. The effect of the omitted reference level for a factor with k can be reconstructed as follows:

$$\beta_{k[missing]} = -\frac{\sum_{i=1}^{k-1} n_i \beta_{ki}}{n_{\text{ref}}} \tag{5}$$

where $\beta_{k[missing]}$ is the reconstructed coefficient for the omitted reference level, n_i and β_{ki} are the sample size and estimated coefficient for the i-th category respectively, and n_{ref} is the sample size of the reference level.

3.1.2 Interpretation of Effects

The coefficients of the Poisson regression model are expressed on the logarithmic scale and are therefore not directly interpretable. However, exponentiating the coefficients converts them to multiplicative effects:

$$\exp(\beta) = \text{relative error rate effect.} \tag{6}$$

Under sum coding, coefficients represent deviations from the global mean. Exponentiating a coefficient therefore gives the multiplicative change in the expected error rate relative to the mean.

For WER and related error-rate metrics, the interpretation is:

$$\begin{cases} \exp(\beta) > 1.0, & \text{WER is higher than the mean} \\ \exp(\beta) < 1.0, & \text{WER is lower than the mean.} \end{cases} \tag{7}$$

The interpretation of whether higher is better or worse depends on the metric being analysed. For metrics where higher values indicate better performance (e.g., BLEU or accuracy), the interpretation of multiplicative effects is reversed.

3.2 Diagnosing Misspecification: The Dispersion Ratio

While Poisson regression assumes equality between the mean and variance of the count distribution, real-world ASR data may exhibit overdispersion. Errors within an utterance are not independent, and some utterances are systematically more difficult than others.

To quantify this deviation, we use the *dispersion ratio* $\hat{\phi}$, defined as the Pearson chi-squared statistic χ^2_P divided by the residual degrees of freedom df_{resid}:

$$\hat{\phi} = \frac{\chi^2_P}{\text{df}_{\text{resid}}} = \frac{1}{\text{df}_{\text{resid}}} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{8}$$

where y_i are the observed error counts, $\hat{\mu}_i$ are the model-predicted means, and $V(\hat{\mu}_i)$ is the assumed variance which is equal to $\hat{\mu}_i$ under Poisson. A value of $\hat{\phi} \approx 1$ indicates adequate model fit, while $\hat{\phi} \gg 1$ signals overdispersion.

It is important to note that overdispersion in ASR data is not always a fundamental property of the data, and it could be a symptom of data quality issues. For ASR data, this mostly manifests as inflated zero-counts and hallucination loops. Hallucination loops is the term given to a specific type of model artefact where the model enters a loop and repeat a set of tokens e.g. "dw i'n mynd i mynd i mynd i…". The overdispersion ratio is particularly sensitive to outliers caused by hallucination loops and other model artefacts, and can therefore be used as a "first-pass" test to validate the data integrity. If $\hat{\phi} \gg 1$, there are likely significant outliers in the data that should be investigated. In our testing, a handful of artefacts representing approximately 0.1% of the overall dataset was sufficient to increase the dispersion ratio from the range $[1, 2]$ to as high as $[3, 12]$.

3.3 Refined Model: Negative Binomial Regression

While Poisson regression provides a natural and computationally stable baseline for modelling ASR error counts, it imposes a strong assumption: that the mean and variance of the count distribution are equal. As discussed in Section 3.2, this assumption is frequently violated in practice.

The Negative Binomial (NB2) distribution extends the Poisson model by introducing a dispersion parameter α that explicitly models the excess variance. The variance function becomes:

$$\text{Var}(y_i) = \mu_i + \alpha \mu_i^2. \tag{9}$$

When $\alpha \to 0$, the Negative Binomial converges to the Poisson distribution, making Poisson a special case of the more general NB2 model.

The regression model is otherwise identical in structure to the Poisson formulation as seen in Equation 4, with coefficients interpreted identically to those described in Section 3.1.2. The dispersion parameter α is estimated jointly with the regression coefficients via maximum likelihood.

3.3.1 Convergence

Compared to Poisson regression, NB2 needs to converge and requires sufficient data in all groups. In our testing, failure to converge is rare in practice, especially if model artefacts such as hallucination loops are filtered out before applying the regression.

3.4 Confidence Intervals and Uncertainty

Both Poisson and NB2 regression produce coefficient estimates via maximum likelihood. By asymptotic theory, the MLE is approximately normally distributed around the true parameter value for large samples under correct model specification (McCullagh, 2019). This motivates Wald confidence intervals, which take the form $\hat{\beta} \pm z_{\alpha/2} \cdot \hat{SE}(\hat{\beta})$, where the standard error is derived from the curvature of the log-likelihood at its maximum via the Fisher information matrix. While computationally convenient, it inherits the model's distributional assumptions. Hence, if the model is misspecified, the standard errors are biased and coverage degrades. As we show in Section 4, this is a practical concern for ASR data.

Since both the implementations of Poisson and NB2 use the Wald derived confidence intervals by default, we will refer to these methods as Poisson+Wald and NB+Wald respectively for brevity.

3.5 Alternative Model: Bootstrapping

Wald intervals derived from both Poisson and NB2 regression rely on asymptotic normality of the MLE, an assumption that is violated under the misspecification patterns present in real ASR data (see Section 4.2). These violations may cause the confidence intervals to be miscalibrated.

In addition, while we are able to reconstruct the effect of the reference level, reconstructing a reliable p-value or confidence interval for this effect is challenging. It is often desirable to apply a family-wise error correction (FWER) when considering a group of hypotheses (e.g. each dialect in our data), but this is not straightforward when one level is missing.

To overcome these issues we propose an alternative model that combines NB2, bootstrapping, WEC, and Bonferroni–Holm corrections (Holm, 1979). By using stratified sampling with replacement, applying NB2 to each resample, and reconstructing the reference level effect using WEC, we are able to build an empirical distribution of the multiplicative effects for each parameter. By using the percentile method, we are then able to derive a confidence interval for each parameter with an accompanying p-value. The FWER can then be properly handled using Bonferroni–Holm corrections, and only the parameters that remain statistically significant following this correction are considered statistically significant. For brevity, we will refer to this method as NB+Boot for the rest of the paper.

Note that this model has some limitations, mainly that it becomes miscalibrated if there are not enough utterances, but the benefits can be summarised as: we get confidence intervals and p-values for all effects, including the reference; WEC handles imbalanced datasets well; we are able to trivially apply Bonferroni–Holm corrections to account for FWER; and the model is more robust to misspecification than NB+Wald, with some caveats.

4 Empirical Validation

4.1 Simulation Study

To validate the statistical calibration and robustness of the proposed methodologies, we conducted a controlled simulation study. The goal of this experiment is to assess whether the three inference procedures discussed in Section 3 are well calibrated. To achieve this, we evaluated three different inference procedures on two group-size scenarios and four test data generation scenarios with known true effects. We ran these simulations 1000 times to check how often the true effect was within the 95% confidence interval for each inference procedure. 1000 samples gives a standard error of:

$$SE = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{\sqrt{0.95 \cdot 0.05}}{\sqrt{1000}} \pm 0.69\% \tag{10}$$

where p is the probability of the true value being with the confidence interval (equal to $1 - \alpha$) and n is the number of samples.

Four test data generation methods were used:

- Clean NB: Data was sampled from a regular negative binomial distribution ($\alpha = 0.5, \mu = 2.44$).
- Zero-inflated: Since well-performing ASR models can often feature an inflation of perfect transcriptions, we use a function that sets a pre-determined rate of error counts to zero ($p_0 = 0.1$).

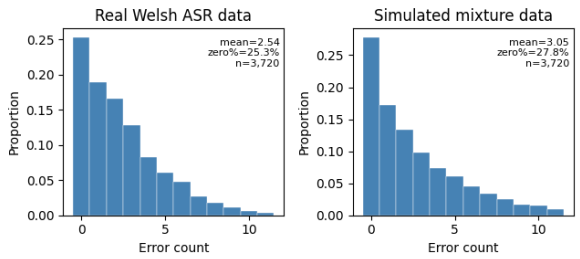

Figure 1: Error count from the generated data compared to the error count of a fine-tuned Welsh ASR model.

- Heavy right tail: ASR results often feature catastrophic errors, e.g. hallucination loops. To simulate this, the function has a random chance to multiply the error by ten ($p_c = 0.005$).
- Mixture: This is a combination of the zero-inflated function and the heavy right tail function.

The output of the mixture was tuned to align with the WERs of a Welsh fine-tuned Whisper V3 model taken from HuggingFace (Jones, 2025). As we will discuss in Section 4.2, due to residual misspecification, a perfect fit is not possible, but as can be seen in Figure 1, the error distribution from the model closely aligns with that of the generated data. The `statsmodels` library (Seabold and Perktold, 2010) was used for the Poisson and NB2 models.

The two group-size scenarios were:

- Balanced: Five groups, each with 600 samples
- Imbalanced: Five groups: 1,700, 1000, 600, 80, 15.

The true effects for the groups were $(0.0, 0.1, -0.1, 0.2, 0.0)$. Note that the first group is always chosen as the reference for the regression model.

4.1.1 Results

The full table showing the results from the simulations can be seen in Appendix A.

In all scenarios, regardless of group size and misspecification, Poisson+Wald showed significant miscalibration ($\mu = 0.67$, 0.55-0.77) compared to the nominal calibration of 0.95. This is not surprising given that the distribution of error counts do not neatly follow a Poisson distribution and there is no in-built mechanism to handle overdispersion. As such, Poisson+Wald is not appropriate in these scenarios due to systematic overdispersion.

NB+Wald showed that under correct model specification, it is very robust with all the deviations ($\mu = 0.95$, 0.94-0.96) being within the margin of error of the desired 0.95. This is expected. However, it is sensitive to model misspecification. Zero-inflation generally caused the model to become slightly more conservative ($\mu = 0.96$, 0.95-0.97), while heavy right tails caused it to become less conservative ($\mu = 0.90$, 0.88-0.94). The degradation in coverage caused by heavy right tails is the most concerning, with the model being very sensitive to outliers. The contamination rate ($p_c = 0.005$) and multiplicative factor (10) used for this simulation underrepresents the outliers that we have observed in real ASR data, which implies that the coverage could be worse in some realistic scenarios.

In contrast, for the balanced group scenario, NB+Boot remained more robust than NB+Wald under all misspecification scenarios ($\mu = 0.94$, 0.92-0.95). NB+Boot seems to struggle most with lower group sizes which causes a degradation of coverage in the $n = 80$ and $n = 15$ scenarios. These were groups that NB+Wald handled almost as well as the balanced groups.

To summarise: Given enough data in each group, NB+Boot is more robust to model misspecification than NB+Wald which inherits the distributional assumptions from the underlying model. NB+Wald remains robust and well-calibrated under correct model specification, hence if the data does not contain significant outliers and model artefacts, it should be reasonably well suited for ASR regression. Poisson+Wald struggles with overdispersion and the types of misspecification that are prevalent in realistic ASR data, and unless the overdispersion is handled properly (e.g. through speaker-level random effects), it is likely a poor fit for many ASR scenarios.

4.2 Empirical Misspecification Analysis

To evaluate the misspecification of the ASR data, we fitted a Poisson model and a NB2 model to the data, then compared the expected frequency of errors to the observed frequency of errors. The results for the NB2 model, including the difference between results before and after filtering out hallucination loops can be seen in Figure 2.

Removing the outliers significantly improved the fit and reduced the dispersion ratio from 3.693 to 1.051. The dispersion ratio for the Poisson model was reduced from 7.903 to 1.672. The overdispersion ratio of 1.672 for Poisson shows that there

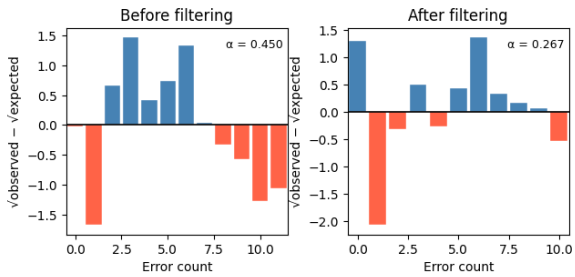

Figure 2: Rootogram showing observed vs. fitted NB2 error counts before (left) and after (right) filtering out hallucination loops (WER > 500%) on the fine-tuned Welsh ASR model. The residual misspecification after filtering (excess mass at counts 2-6 and deficit at 0 and 8+) motivates the use of NB+Boot over NB+Wald for inference in some cases.

Accent	Utterances
Gogledd Orllewin	1,793
De Ddwyrain	1,073
De Orllewin	626
Canolbarth	146
Gogledd Ddwyrain	82
Unknown	15

Table 1: Overview of the number of utterances in Lleisiau ARFOR by accent.

is still significant overdispersion in the data. The rootogram in Figure 2 also shows that despite the improved dispersion ratio for NB2, there is still residual misspecification.

5 Case Study: Welsh ASR Evaluation

5.1 Dataset: Lleisiau ARFOR

Lleisiau ARFOR (Cymen et al., 2024) is a transcribed spontaneous speech corpus featuring speakers from all five traditional Welsh dialects. It is released under the CC0 license. It is the only dataset for Welsh that contains this metadata.

Since speaker data is anonymised, it is not possible to determine with confidence how many speakers belong to each category. However, according to the dataset creators "Canolbarth" likely involves a single speaker while the "Unknown" category is likely just missing data. For this experiment we will exclude the unknown category as a consequence.

5.2 Initial NB2 Fit and Diagnosing Overdispersion

We initially fitted a NB2 distribution to the data with accent as a categorical group and duration as a tracer variable. The overdispersion ratio of this

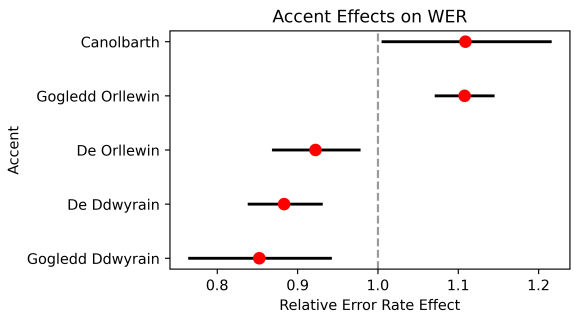

Figure 3: Forest plot showing the mean relative effect size of each accent with the 95% CI after removing the three outliers. Means marked in red are statistically significant.

initial fit was 3.694, significantly greater than 1, which indicated that it was highly probable that the test data contained model artefacts.

A manual inspection of the data found that there were three hallucination loops present in the data, which represented 0.08% of the overall dataset. These hallucination loops were the only utterances with WERs over 500%, hence we added a filter to remove any utterances above that threshold. Following the removal of those three outliers, the overdispersion ratio was reduced to 1.051.

We subsequently ran 25k bootstrap iterations. The reason for the high number was that several of the confidence intervals (e.g. Canolbarth) were close to 1 so we wanted to get as accurate a result as possible. We also ran the same experiment using the Poisson+Wald and NB2+Wald methods.

5.3 Results

A forest plot with the results from the NB+Boot method can be seen in Figure 3.

The multiplicative factors for the various accents are consistent across the three methods. Where they differ is in the width of the confidence intervals. Poisson+Wald returns narrower CIs than the other methods and as such returns a significant result for all accents. NB2+Wald had the widest confidence intervals, and as such only returned a significant result for the accent De Ddwyrain. The width of the CIs for NB+Boot fell between Poisson+Wald and NB+Wald but meant that all accents had significant results (though Canolbarth and De Orllewin were barely significant). After applying Holm correction to the p-values for the accents using the open-source *[redacted for blind-review]* library (*Author*, 2026), both Canolbarth and De Orllewin became non-significant as ex-

	Poisson		NB2+Wald		NB+Boot		
	Mean	**Sig.**	**Mean**	**Sig.**	**Mean**	**Sig.**	**Holm corrected**
Gogledd Orllewin	N/A				1.108	Yes	Yes
Canolbarth	1.107	Yes	1.106	No	1.107	Yes	No
De Ddwyrain	0.881	Yes	0.885	Yes	0.884	Yes	Yes
De Orllewin	0.915	Yes	0.930	No	0.922	Yes	No
Gogledd Ddwyrain	0.857	Yes	0.844	No	0.854	Yes	Yes
duration	0.935	Yes	0.927	Yes	0.930	Yes	*N/A*

Table 2: Table showing the multiplicative factors from each of the methods and whether the results are significant to a confidence level of $\alpha = 0.05$. Poisson and NB2+Wald both use Gogledd Orllewin as a reference, hence the accent does not have results for these methods.

pected. The only accent NB+Wald and NB+Boot disagree about is Gogledd Ddwyrain (not sig. and sig. respectively). This accent has the lowest multiplicative factor of all accents, however, it only has 82 utterances, so the result is quite uncertain, especially given that neither method is perfectly calibrated.

Duration yielded significant results across all methods with a multiplicative factor of 0.94, 0.93, and 0.93 respectively. This means that the longer the clips become, the more accurate the transcriptions. Since the model tested is built on the transformer architecture, this may reflect the transformer model benefiting from longer contextual windows.

5.4 Further Analysis into Systematic Bias in Welsh ASR

We had access to the inference outputs of 12 different Welsh ASR models, and we ran the same process on the data for all 12 models. The multiplicative factors for the different accents are generally stable across the different models with some variation. Since the models have been trained and tested on similar data, this could indicate the following: that the training data introduces systematic bias across models; that the test data for different accents differs in difficulty; or a mixture of both. This warrants further investigation, but falls outside the scope of this paper.

6 Limitations

6.1 Speaker-Level Random Effects

As mentioned earlier, the use of mixed-effect models and speaker-level random effects as used by Liu et al. (2022) might be able to absorb the excess variance (though it is not guaranteed). However, our method extends their research by enabling subgroup performance disparity analyses in settings where speaker IDs might not be available.

6.2 Other Regression Alternatives

We limited the scope to Poisson and NB2 regression. However, there are other regression alternatives that might be suitable in different contexts, notably mixed-effect models and zero-inflated NB2 models. We encourage researchers undertaking model-based subgroup performance disparity analyses to investigate the suitability of different models in their contexts.

6.3 Model Selection

We only included one model in our case study, which was also the only model used for calibrating the data generation methods. This is obviously a limitation of this study. However, we have found that the findings hold for other models of different architectures (Wav2vec2 (Baevski et al., 2020), Whisper (Radford et al., 2023), and Kaldi (Povey et al., 2011)) as well so we are confident that the findings can be replicated for models with similar performance. Better models might face issues with zero-inflation; this needs to be investigated.

7 Potential Risks

The main risks posed by this research is the wrongful confirmation of non-existent bias or refutation of real bias caused by miscalibrated models. We hope that contribute to reducing this risk.

8 Conclusion

We have shown that standard modelling assumptions are routinely violated in realistic ASR data, and that no single method is universally appropriate. Poisson regression is systematically miscalibrated under overdispersion; NB+Wald performs well when data is clean but degrades under heavy-tailed misspecification; and NB+Boot offers the best robustness when group sizes are sufficient. We additionally show that the dispersion ratio provides a simple but sensitive first-pass diagnostic for data

quality issues such as hallucination loops. Applied to Welsh ASR, our methodology reveals accent-based performance differences whose statistical significance varies substantially across inference methods. We encourage practitioners to treat model selection and assumption-checking as integral parts of any subgroup performance disparity evaluation pipeline, rather than afterthoughts.

9 Source Code Availability

The source code is available online at `https://anonymous.4open.science/r/bias-evaluation-9E6E/`.

10 Acknowledgements

Left empty for blind review

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Cymen, ARFOR, and the Language Technologies Unit, Bangor University. 2024. Lleisiau arfor. distributed via Huggingface.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Marcio Fückner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151. ISSN: 2771-5787.

Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

Maliha Jahan, Priyam Mazumdar, Thomas Thebaud, Mark Hasegawa-Johnson, Jesús Villalba, Najim Dehak, and Laureano Moro-Velazquez. 2025. Unveiling Performance Bias in ASR Systems: A Study on Gender, Age, Accent, and More. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Maliha Jahan, Laureano Moro-Velazquez, Thomas Thebaud, Najim Dehak, and Jesús Villalba. 2023. Model-Based Fairness Metric for Speaker Verification. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. Conference Name: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) ISBN: 9798350306897 Place: Taipei, Taiwan Publisher: IEEE.

Dewi Bryn Jones. 2025. evals-speech-recognition-cy-en-2511 (Inference Results for whisper_large_v3_ft_btb_cv_cvad_ca_cy_2511). distributed via Huggingface. Canolfan Bedwyr, Bangor University.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689. Publisher: Proceedings of the National Academy of Sciences.

Zhe Liu and Fuchun Peng. 2020. Statistical testing on ASR performance via blockwise bootstrap. In *Interspeech 2020*, pages 596–600. ISSN: 2958-1796.

Zhe Liu, Irina-Elena Veliche, and Fuchun Peng. 2022. Model-Based Approach for Measuring the Fairness in ASR. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6532–6536. ISSN: 2379-190X.

P. McCullagh. 2019. *Generalized Linear Models*, 2 edition. Routledge, New York.

Mikel K. Ngueajio and Gloria Washington. 2022. Hey ASR System! Why Aren't You More Inclusive? In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 421–440, Cham. Springer Nature Switzerland.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. Place: Hilton Waikoloa Village, Big Island, Hawaii, US.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *9th python in science conference*.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of*

the First ACL Workshop on Ethics in Natural Language Processing, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Manfred te Grotenhuis, Ben Pelzer, Rob Eisinga, Rense Nieuwenhuis, Alexander Schmidt-Catran, and Ruben Konig. 2017. When size matters: advantages of weighted effect coding in observational studies. *International Journal of Public Health*, 62(1):163–167.

Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M. Rush, Thomas Wolf, and Douwe Kiela. 2022. Evaluate & evaluation on the hub: Better best practices for data and model measurements. ArXiv: 2210.01970 [cs.LG].

A Full simulation results table

Method	Group size	Misspecification	G1	G2	G3	G4	Mean
Poisson+Wald	balanced	clean_nb	0.743	0.751	0.734	0.753	0.745
		heavy_tail	0.610	0.644	0.616	0.617	0.622
		mixture	0.578	0.602	0.560	0.590	0.583
		zero_inflated	0.719	0.716	0.693	0.717	0.711
	imbalanced	clean_nb	0.737	0.755	0.740	0.766	0.750
		heavy_tail	0.576	0.609	0.614	0.741	0.635
		mixture	0.546	0.606	0.574	0.651	0.594
		zero_inflated	0.720	0.735	0.690	0.702	0.712
NB2+Wald	balanced	clean_nb	0.955	0.950	0.943	0.953	0.950
		heavy_tail	0.880	0.898	0.883	0.899	0.890
		mixture	0.908	0.910	0.911	0.907	0.909
		zero_inflated	0.966	0.954	0.964	0.957	0.960
	imbalanced	clean_nb	0.951	0.948	0.960	0.949	0.952
		heavy_tail	0.890	0.898	0.895	0.937	0.905
		mixture	0.911	0.925	0.919	0.906	0.915
		zero_inflated	0.967	0.959	0.955	0.966	0.962
NB2+Boot	balanced	clean_nb	0.947	0.950	0.935	0.936	0.942
		heavy_tail	0.923	0.932	0.929	0.935	0.930
		mixture	0.935	0.934	0.944	0.929	0.936
		zero_inflated	0.950	0.944	0.949	0.942	0.946
	imbalanced	clean_nb	0.941	0.935	0.949	0.891	0.929
		heavy_tail	0.925	0.933	0.904	0.885	0.912
		mixture	0.946	0.943	0.921	0.845	0.914
		zero_inflated	0.954	0.944	0.928	0.895	0.930

Table 3: Table showing the probability that each of the groups falls within the 95% confidence interval. Group 0 (the largest group) is omitted since it is used as a reference.