



MSc Health Data Science Project Title

Manasvi Saite, Kaushal Mayur¹

¹ School of Mathematical and Statistical Sciences, University of Galway

Background and Problem

Healthcare analytics often requires models suited to non-standard outcomes (e.g., counts or ordered categories), where standard linear regression is inappropriate. In this poster we study **self-reported general health** (HealthGen) from NHANES, recorded on an ordered scale from “Excellent” to “Poor”, and use an ordinal regression framework to quantify how age, BMI, sex and smoking shift health ratings.

Objectives of Project

- Fit a proportional-odds (cumulative logit) model for HealthGen.
- Estimate and visualise predicted probabilities of each health category across age.
- Compare predicted health profiles for current smokers vs non-smokers.

Data Sources and Datasets

NHANES Overview: Nationally representative survey of ~10,000 US residents annually. Combines interviews, physical exams, and lab tests from civilian noninstitutionalized population. [web:297]

Dataset: NHANES 2013-2014 (R package **NHANES**). ~9,700 participants. [web:303]

HealthGen (Outcome): Self-reported general health rating (ages ≥ 12). 5 ordered categories: - **Excellent** (best) \rightarrow **Very good** \rightarrow **Good** \rightarrow **Fair** \rightarrow **Poor** (worst) - Typical US distribution: 15% Excellent, 41% Very good, 49% Good, 16% Fair, 4% Poor [web:317]

Key Predictors (supervisor-specified model): - **Age:** Years (continuous, adults only) - **Gender:** Male/Female - **BMI:** kg/m² (body mass index from measured height/weight) - **SmokeNow:** Current smoker (Yes/No)

Analytic Sample: Complete cases only (n=5,736). Filtered for non-missing Age, Gender, BMI, SmokeNow, HealthGen.

Early Results / Descriptive Statistics of Datasets

The modelling dataset includes **2,036** complete cases with mean age **50 \pm 17** years and mean BMI **29.1 \pm 6.7** kg/m².

Current smoking prevalence is **48.6%** and females comprise **42.1%** of the analytic sample.

Some basic summaries of the dataset are below:

Table 1: Table of Nhanes data

Measure	Value
N participants	2,036
Age (mean \pm SD)	50 \pm 17
BMI (mean \pm SD)	29.1 \pm 6.7
Current smokers (%)	48.6%
Females (%)	42.1%

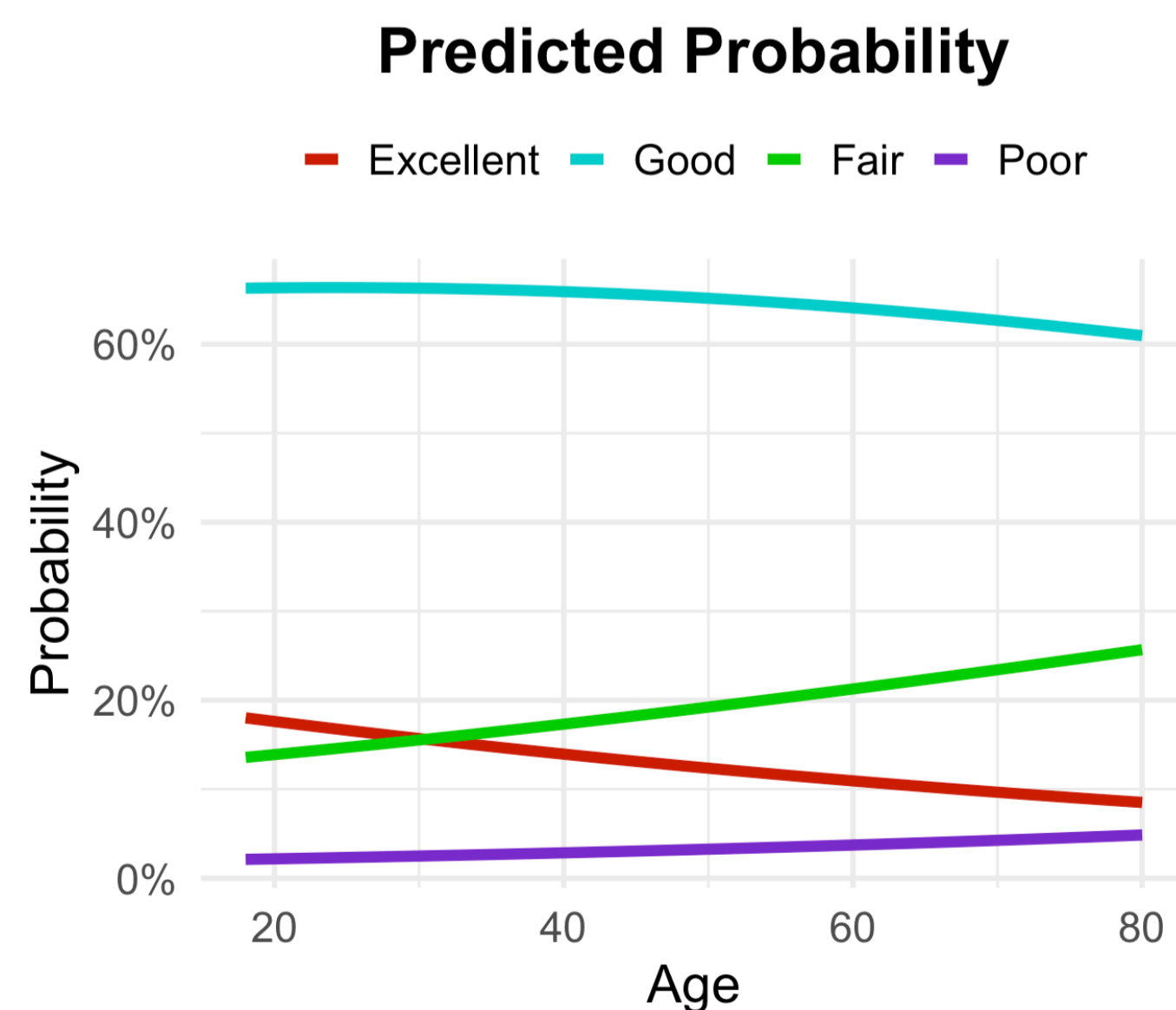


Figure 1: Predicted health probabilities by age (typical covariates).

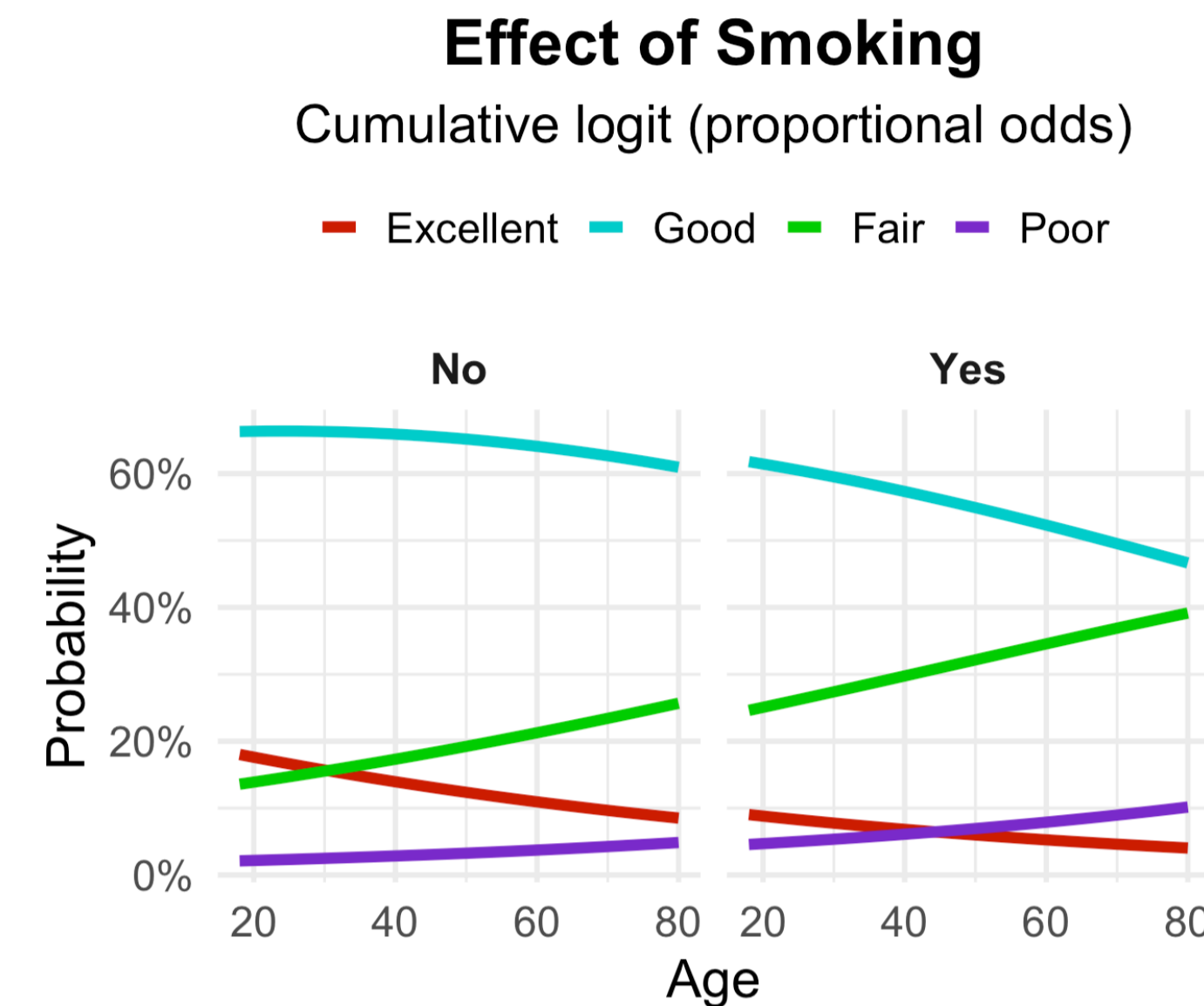


Figure 2: Smoking effect on predicted health probabilities.
Interactive: Smoking effect on predicted general health

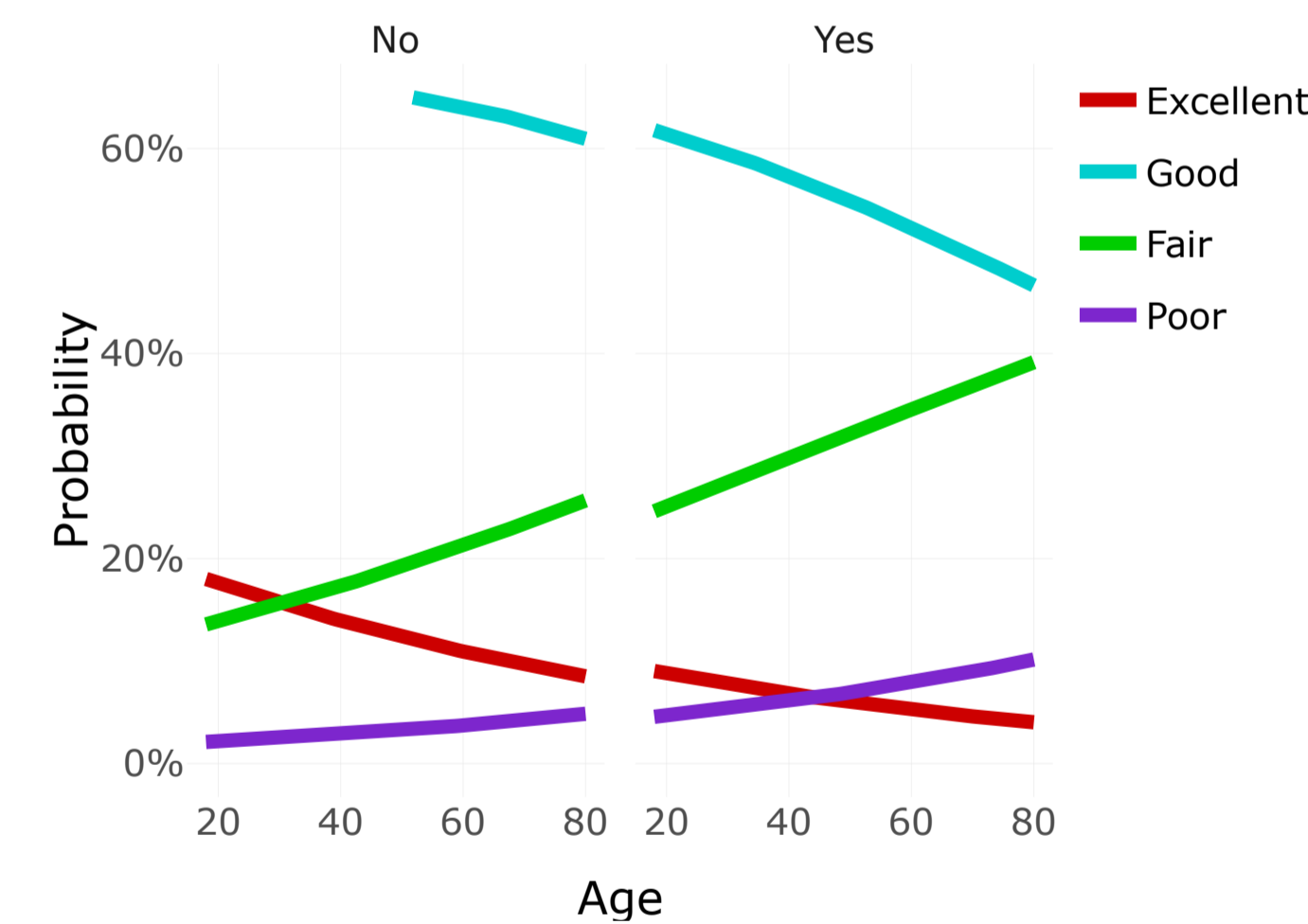


Figure 3: Interactive plot (hover to see probabilities) showing predicted HealthGen probabilities by age and current smoking status.

Next Project Steps

We plan to conduct further analysis:

-Compare model variants using information criteria (AIC/BIC) and diagnostic plots.

-Consider additional predictors available in NHANES (e.g., activity, alcohol, comorbidity measures).

-Report uncertainty (confidence intervals) for key effects and validate predictions on a holdout split.

GitHub

The code and datasets for this project can be viewed at our GitHub repository here:
https://github.com/manasvisaite1504/Count_data_health_Industry-project

References

NHANES data package (R).

VGAM cumulative logit modelling.