Air Pollution and Respiratory Difficulties

P408 - Main Body Word Count: 1974

December 17, 2025

Contents

Declaration of AI use

In creating this work, I have used AI tools for the following tasks:

1. Literature search and ideation: not used
2. Spelling and grammar: Overleaf has in-built spell-check and AI (TexGPT). TexGPT highlights parts of the text in the editor, though I have not used it.
3. Improvements to writing: not used
4. Coding: I used Claude AI for research on the stargazer package.
5. Plotting, formatting etc: I used Claude AI and ChatGPT Edu for help with latex formatting. In particular, for altering font size in tables, as well as controlling the number of significant figures in the output.

1 Introduction

This report analyses a dataset from a survey on air pollution in a South Asian city. The report begins with exploratory data analysis, which summarises the variables in the dataset and displays some high level associations between respiratory difficulties (henceforth, RD) and age, gender, physical activity, and air pollution. The next section concerns model selection, beginning with a Bernoulli generalised linear regression model (henceforth, GLM) of RD on all the regressors mentioned above and interactions between exposure to air pollution and the remaining covariates. I then present the results of model selection, finding that all criteria (likelihood ratio tests, Wald tests, and Information Criteria) support dropping all the main effects of the regressors (with the exception of activity), while keeping all the interactions. Invoking the hierarchy principle, I expose the fragility of the model to non-substantive changes in the data, such as recoding the air pollution binary variable. Given the main aim of the model is interpretive, I conclude that the preferred model must include all baselines. I then present model diagnostics based on plots, s.a., residual plots, QQ plots, and leverage/ Cook's distance plots. While some potential outliers are identified, there is no apparent reason to reject them. The following section then provides an interpretation of the findings, including an evaluation of the odds ratios for each variable and the respective confidence intervals. The final section predicts the probability of developing RD for two new data points, one within-sample and one out-of-sample. The predictions are paired with the respective confidence interval and a brief discussion of out of sample reliability.

2 Exploratory Data Analysis

The dataset comprises 285 observations of one binary variable (Respiratory Difficulty) and four regressors (Activity, Gender, Air Pollution, and Age). The first three regressors are uncontroversially categorical (binary), while the last is treated as continuous. In fact, age takes only four values (25, 35, 45, 55) representing 10-year age bands, rather than the age of the participants. This could substantiate a categorical interpretation of age as a factor, which is shown in the appendix (Table 12). Since part of the model's objective is to predict the probability of RD for an out-of-sample individual (85 year old), age will be treated as continuous for the remainder of the analysis.

The following figures illustrate the distribution of the data. First, Figure 1 shows the marginal distribution of each variable and a histogram of age.

Variable	Level	Count	Percent
Respiratory disease	No	154	54
Respiratory disease	Yes	131	46
Air pollution	No	123	43.20
Air pollution	Yes	162	56.80
Activity	No	38	13.30
Activity	Yes	247	86.70
Gender	Female	102	35.80
Gender	Male	183	64.20
Age	25	124	43.50
Age	35	76	26.70
Age	45	38	13.30
Age	55	47	16.50

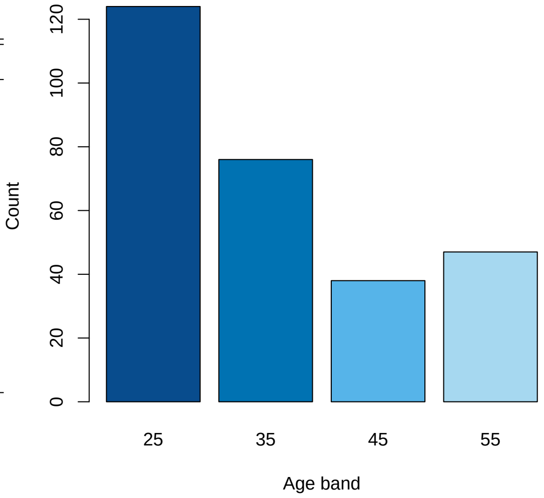

Figure 1: Counts and proportion breakdown of each category, with visual representation for age

Table 1 then breaks down the dataset by the values of the regressors. Of 32 possible combinations of the regressors, only 24 are represented in the data, and the proportion of each subset experiencing RD varies widely. The proportion of subjects experiencing RD tends to be higher in those exposed to air pollution.

Air Pollution	Activity	Gender	Age	Resp. diff. proportion	Count
No	No	Male	25	0.75	4
No	No	Male	35	0.00	1
No	No	Male	45	1.00	1
No	Yes	Female	25	0.11	18
No	Yes	Female	35	0.17	12
No	Yes	Female	45	0.25	4
No	Yes	Female	55	0.33	3
No	Yes	Male	25	0.26	34
No	Yes	Male	35	0.16	19
No	Yes	Male	45	0.00	8
No	Yes	Male	55	0.05	19
Yes	No	Female	25	1.00	6
Yes	No	Female	35	1.00	2
Yes	No	Male	25	0.25	16
Yes	No	Male	35	0.20	5
Yes	No	Male	45	0.00	3
Yes	Yes	Female	25	0.83	18
Yes	Yes	Female	35	0.94	16
Yes	Yes	Female	45	1.00	5
Yes	Yes	Female	55	0.94	18
Yes	Yes	Male	25	0.46	28
Yes	Yes	Male	35	0.57	21
Yes	Yes	Male	45	0.65	17
Yes	Yes	Male	55	1.00	7

Table 1: Counts and proportion affected by respiratory difficulty for each combination of the regressors

The following barplots display the conditional probability of being affected by RD given the values of the regressors.

$$\mathbf{P}(\texttt{respd} = 1 \,|\, X = x) \quad \forall X \in \{\texttt{gender,} \quad \texttt{activity,} \quad \texttt{airpollution,} \quad \texttt{age}\}$$
$$\mathbf{P}(\texttt{respd} = 0 \,|\, X = x) \quad \forall X \in \{\texttt{gender,} \quad \texttt{activity,} \quad \texttt{airpollution,} \quad \texttt{age}\}$$

The visual shows that air pollution and gender are strongly associated with RD, with females much more likely to suffer from RD, and similarly for subjects exposed to air pollution. The plot displays limited dependence on activity and a weak positive association between age and RD.

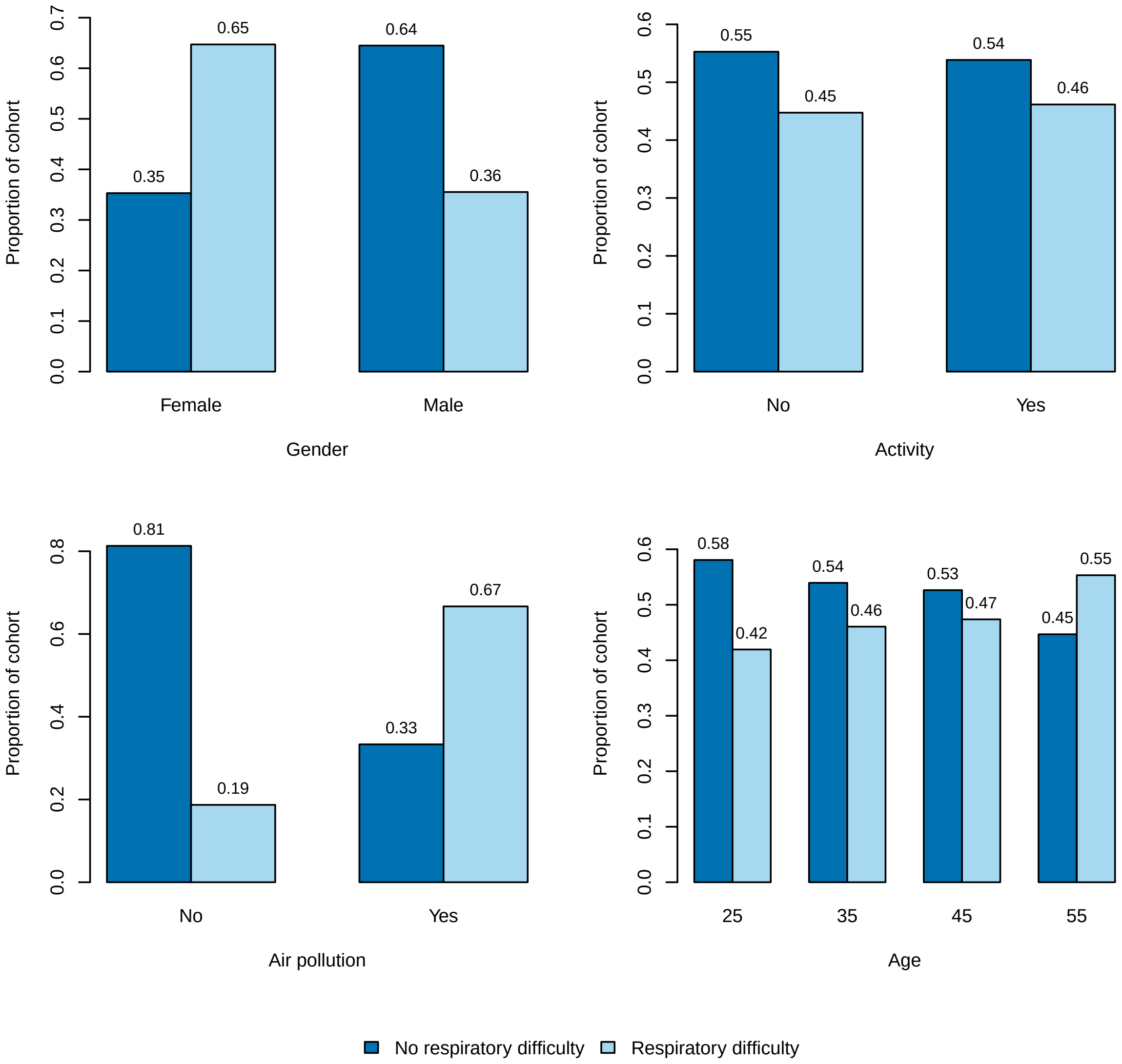

Figure 2: Incidence of respiratory difficulty by regressor levels

Since air pollution has the strongest intuitive link with RD, consider further conditioning the distributions on air pollution, so plotting:

$$\mathbf{P}(\texttt{respd} = 1 \,|\, X = x, \texttt{air pollution} = i) \quad \forall X \in \{\texttt{gender, activity, age}\}, i \in \{0,1\}$$

$$\mathbf{P}(\texttt{respd} = 0 \,|\, X = x, \texttt{air pollution} = i) \quad \forall X \in \{\texttt{gender, activity, age}\}, i \in \{0,1\}$$

This illuminates the structure of the dataset, showing that RD appears almost conditionally independent of gender when air pollution is absent, and displaying a strong association with activity both when present and when absent, as well as a strong positive association with age when present. This is in line with our intuition, since being physically active should improve health outcomes in general, but would also expose subjects more to pollutants, leading to diverging effects on RD depending on pollution. Similarly, we expect older people to be more vulnerable to the effects of air pollution.

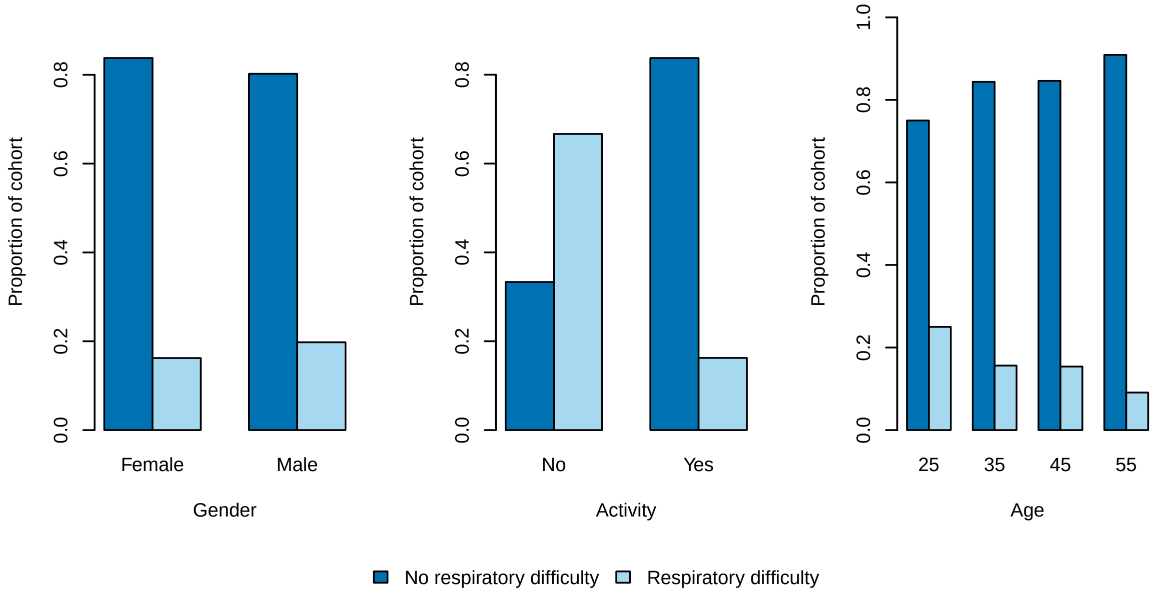

Figure 3: Incidence of respiratory difficulty in subjects not exposed to air pollution by regressor levels

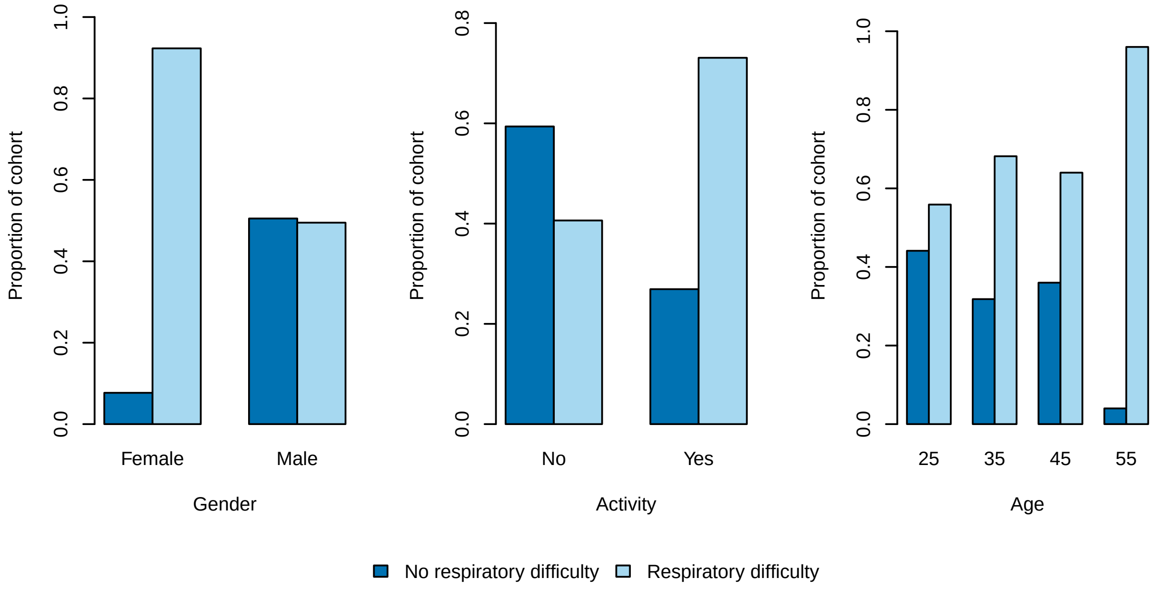

Figure 4: Incidence of respiratory difficulty in subjects exposed to air pollution by regressor levels

These plots are strongly informative, since they imply that interactions between air pollution and gender, activity, and age must be included to effectively capture the structure of the data.

3 Modelling

I begin with a Bernoulli GLM, regressing RD on all the regressors and including the interactions between air pollution and age, gender, and activity. This gives the linear predictor below, where the fitted value is $\mu = \sigma(\eta)$[1]

$$\begin{aligned}\eta = x^T\beta = \beta_0 &+ \beta_1\,\texttt{activity} + \beta_2\,\texttt{airpollution} + \beta_3\,\texttt{gender} + \beta_4\,\texttt{age}\\ &+ \beta_5\,\texttt{activity}\times\texttt{airpollution} + \beta_6\,\texttt{gender}\times\texttt{airpollution}\\ &+ \beta_7\,\texttt{age}\times\texttt{airpollution}\end{aligned}$$

This is also the **chosen model**, as justified below. Table 2 then provides the coefficients and standard errors, as well as the p-values for the Wald statistic. All the interaction terms are highly significant under a Wald test, while all the main effects are insignificant.[2]

	Dependent variable:	
	Respiratory difficulty	p-val (Wald)
Air pollution (Yes)	-1.55 (1.50)	p = 0.31
Age	-0.04 (0.02)	p = 0.16
Male	0.11 (0.55)	p = 0.85
Active	-2.17^{*} (0.93)	p = 0.02
Air pollution × Age	0.08^{*} (0.03)	p = 0.02
Air pollution × Male	-2.56^{***} (0.76)	p = 0.001
Air pollution × Active	3.17^{**} (1.04)	p = 0.003
Constant	1.65 (1.24)	p = 0.19
Observations	285	
Log Likelihood	-133.06	
Akaike Inf. Crit.	282.12	
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 2: Logistic Regression of Respiratory difficulty

Since the Wald test and the likelihood ratio test (LRT)[3] are asymptotically equivalent, we expect to reject the hypotheses that (1) the model is useless and (2) interactions should be

[1] The Bernoulli GLM takes the following form (Dobson & Barnett, 2018 [1]): given the canonical link function g, we have:

$$\begin{aligned} y = \texttt{respd} &\sim \mathit{Bernoulli}(\mu)\\ \eta = g(\mu(\texttt{respd})) &= \mathit{log}\left(\frac{\mu(\texttt{respd})}{1-\mu(\texttt{respd})}\right)\\ \Rightarrow \mu = g^{-1}(\eta) = \sigma(\eta) &= \sigma(x^T\beta) = \frac{e^{x^T\beta}}{1+e^{x^T\beta}}\end{aligned}$$

where the linear predictor η is defined as above.

[2] The Wald statistic is defined as follows:

$$w_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \overset{a}{\sim} N[0,1]$$

where $\hat{\beta}_i$ is the estimate obtained by iteratively reweighted least squares (IRLS) and $SE(\hat{\beta}_i)$ is the respective standard error. This is given by the respective entry of the inverse of the observed information, which for the Bernoulli model with canonical link above coincides with the Fisher information $J_{ii}^{-1} = I_{ii}^{-1}$.

[3] The LRT is derived as follows. Given the regressors (including the interactions) $\mathbf{x}$, we can partition the set and the coefficients into:

$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_{(1)}\\ \mathbf{x}_{(2)}\end{pmatrix}; \quad \beta = \begin{pmatrix}\beta_{(1)}\\ \beta_{(2)}\end{pmatrix},$$

where β_2 is a k-dimensional vector. We then define the maximum likelihood estimator (MLE) of the restricted dropped. Table 3 provides the results of the LRT, with the likelihood ratio statistic (LRS) and p-value given for each model restriction. The "drop" column indicates which regressor/interaction was removed in the test. The first test ("null") checks whether the model is useless, meaning it compares the model to the intercept-only model. The p-values in Table 3 provide the probability of obtaining such values of the LRT under the null.

Full Model	Restricted Model	Drop	Likelihood Ratio Statistic	P values
glm1	glm0	All (null test)	127.11	< 0.001
glm1	glm2	Age x Air Pollution	6.79	0.01
glm1	glm3	Activity x Air Pollution	9.69	0.002
glm1	glm4	Gender x Air Pollution	12.95	< 0.001
glm1	glm5	Air Pollution, Age, Gender	2.49	0.48
glm1	glm6	Age	2.25	0.13
glm6	glm7	Gender	0	1.00
glm7	glm8	Activity	7.09	0.01

Table 3: Results of nested Likelihood Ratio Tests

The LRTs confirm what was expected from the Wald tests, so we reject (1) that the model is useless ("null" test) and (2) any models that drop interactions at the 1% significance level. Similarly, the LRT agrees with the Wald test that none of the main effects except activity are significant. Moreover, the joint likelihood ratio test for the main effects strongly supports dropping them, with a p-value of 0.48. The last three tests also consider dropping age, gender, and activity sequentially, again showing that we cannot reject the null that the main effects are 0 with the sole exception of activity (as in the Wald test). This leads to the reduced model below, with estimates and Wald p-values in Table 4.

$$\begin{aligned}\eta = x^T\beta = \beta_0 + \beta_1\,\texttt{activity} + \beta_2\,\texttt{airpollution} \times \texttt{activity}\\ + \beta_3\,\texttt{airpollution} \times \texttt{gender} + \beta_4\,\texttt{airpollution} \times \texttt{age}\end{aligned}$$

	Dependent variable:	
	Respiratory Difficulty	p-val (Wald)
Active	-2.04^{**} (0.64)	p = 0.002
Air pollution × Age	0.04^{*} (0.02)	p = 0.03
Air pollution × Male	-2.53^{***} (0.51)	p $<$ 0.0001
Air pollution × Active	3.00^{***} (0.74)	p = 0.0001
Constant	0.40 (0.59)	p = 0.51
Observations	285	
Log Likelihood	-134.31	
Akaike Inf. Crit.	278.61	
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 4: Logistic Regression of Respiratory Difficulty

This model (glm5 in Table 5) is not only the most parsimonious, but it also outperforms the initial model and all the other nested models that were tested under both information criteria, with little change in the value of the R^2 KL. This would provide strong evidence for selecting model 5.

model under the null $beta_2 = 0$ as $\widetilde{\beta}_{(1)}$, with maximised log-likelihood given by $\ell\left(\widetilde{\beta}_{(1)}; y\right)$. Similarly, the maximised log-likelihood of the unrestricted model is $\ell\left(\widehat{\beta}_{(1)}; y\right)$, giving the LRT statistic:

$$\Lambda(y) = 2\left(\ell(\widehat{\beta}; Y) - \ell\left(\widetilde{\beta}_{(1)}; y\right)\right) \xrightarrow{d} \chi^2_k$$

The distribution converges to χ^2_k under the null, so we reject the null for large values of the statistic relative to the degrees of freedom k.

Model	Description	Residual Deviance	AIC	BIC	R^2 Kullback Leibler
glm1	Full model	266.12	282.12	311.34	0.32
glm2	Drop Age x Air Pollution	272.92	286.92	312.48	0.31
glm3	Drop Activity x Air Pollution	275.82	289.82	315.38	0.30
glm4	Drop Gender x Air Pollution	279.08	293.08	318.64	0.29
glm5	Drop Air Pollution, Age, Gender	268.61	278.61	296.88	0.32
glm6	Drop Age	268.37	282.37	307.94	0.32
glm7	Drop Age, Gender	268.37	280.37	302.29	0.32
glm8	Drop Activity, Age, Gender	275.46	285.46	303.72	0.30

Table 5: Information Criteria, Deviances and R^2 KL

However, dropping the main effects represents a violation of the principle of marginality[4] which holds that we should not drop the main effect of a variable when including the interaction term. This is crucial here, given that:

1. The main purpose of the model is interpretation. Since the coefficient on the interaction $x_1 \times x_2$ indicates the change in the effect of x_1 on the log-odds when x_2 is present, interpretation is difficult when there is no main effect of x_1 (and vice versa)

2. The data is binary, so arbitrary encoding decisions can have serious effects on the significance of the coefficients.

The below exemplifies the second pitfall. The toy model here redefines[5] `airpollution` = "Yes" as 0 and "No" as 1, resulting in:

	Dependent variable:	
	Respiratory difficulty	p-val (Wald)
Air pollution (No)	1.55 (1.50)	p = 0.31
Age	0.05* (0.02)	p = 0.04
Male	−2.45*** (0.52)	p = 0.0000
Active	1.00* (0.48)	p = 0.04
Air pollution × Age	−0.08* (0.03)	p = 0.02
Air pollution × Male	2.56*** (0.76)	p = 0.001
Air pollution × Active	−3.17** (1.04)	p = 0.003
Constant	0.11 (0.84)	p = 0.90
Observations	285	
Log Likelihood	−133.06	
Akaike Inf. Crit.	282.12	
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 6: Toy model – Logistic Regression of Respiratory difficulty

The LRT of the nested model without the main effects for age, gender, and air pollution now yields LRS 49.3, with p-value < 0.001, so we would **not** prefer the reduced model after this minor non-substantive re-encoding. This corroborates interpretive fragility, since reencoding `airpollution` now provides a model where `gender` alone is significant. Given the goal of interpretation, we must opt to keep the main effects despite statistical insignificance. The remaining parts of the analysis for the reduced model can be found in the appendix, and are pointed to in the relevant sections.

[4] Also known as the hierarchy principle

[5] The binary variables are encoded alphabetically, by the R default. This means that, in the context of regression, gender assigns 0 to F and 1 to M, while activity and air pollution assign 0 to No and 1 to Yes.

4 Model Diagnostics

This section assesses the quality of model fit using residual plots to determine if any large outliers may be skewing the analysis, and if patterns in the residuals display obvious signs of misfit.

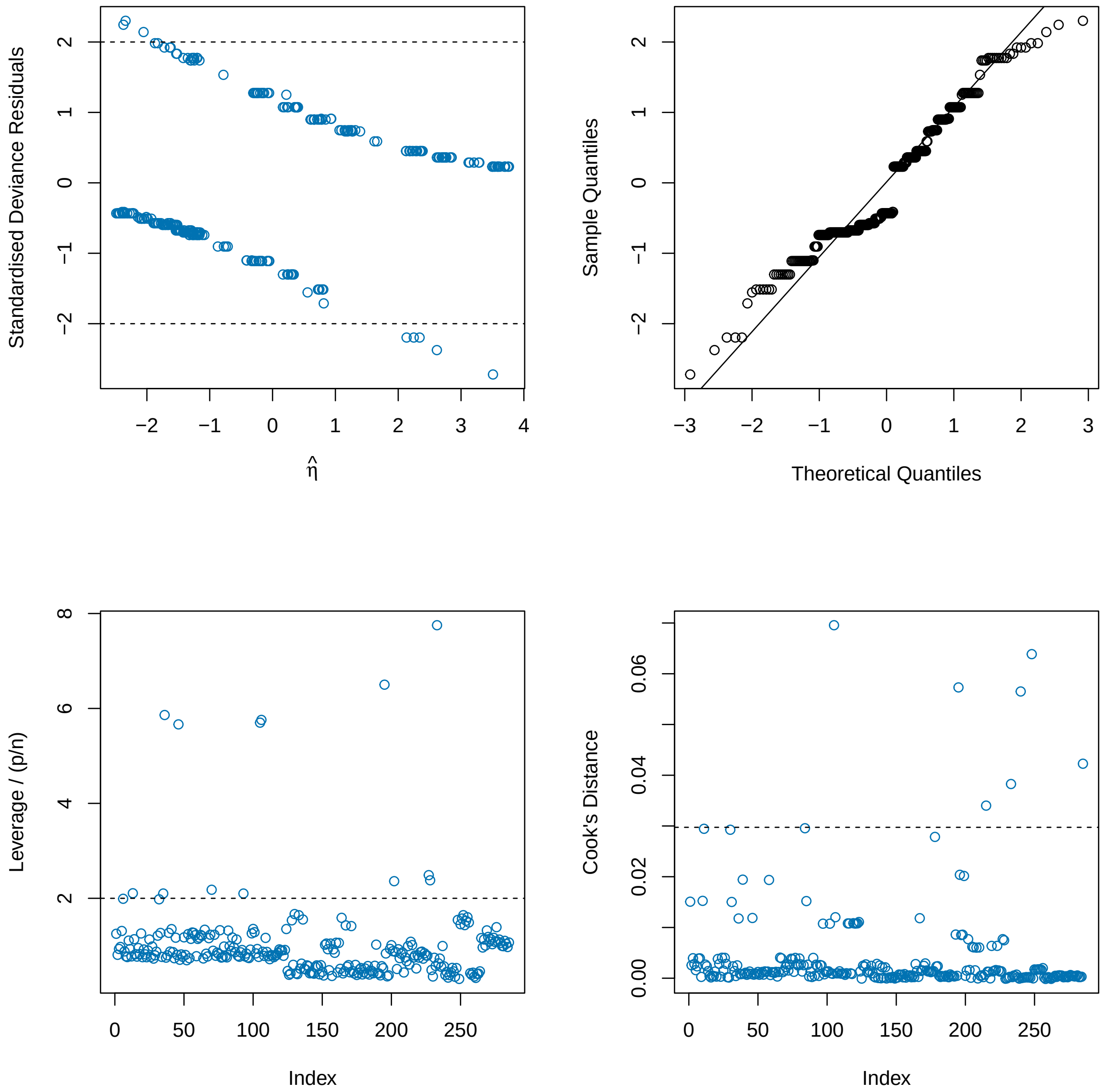

Figure 5: In order Left to Right, Top to Bottom: residuals vs fitted linear predictor, residual QQ-plot, leverage plot, Cook's distance plot

The first plot displays the values of the standardised deviance residuals against the fitted values of the linear predictor η. A plot indexing against the fitted response μ is in the appendix, but the linear predictor is preferable for the present analysis (Faraway, 2016 [2]). For a Bernoulli model, residual normality is not expected as for Poisson or Binomial with large counts. However, the standardisation means the residuals have approximately unit variance, allowing us to identify residuals outside of $(-2, 2)$ as potential outliers. The obvious pattern in the plot is due to the binary nature of dependent variable, where the higher cluster will be residuals of positive deviations (`respd` $= 1$) and vice versa for the lower cluster. Eight observations exceed the ± 2

threshold: not particularly alarming for 285 observations.[6]

Again, since the model is Bernoulli we do not expect the residuals to be approximately normal. However, we expect the QQ plot to be approximately centred around 0, have symmetric tails, and display no strong curvature. This is satisfied by the plot, with the separation in the middle due to the binary nature of `respd`.

The third plot displays leverage divided by the ratio of parameters to sample size (here 8/285).[7] Large observations will have a value greater than $2p/n$, displayed by the values above the dashed line in the third plot.

Finally, the fourth plot displays Cook's Distance (CD) - a measure of the influence of observations on the estimates.[8] Heuristically, potential outliers exceed $8/(n-2p)$, where the denominator comes from the degrees of freedom of the numerator after two estimations of the coefficient vector, and the 8 in the numerator is a conservative empirical threshold. [9]

We then isolate these observations, with scaled leverage and Cook's Distance above the respective thresholds (2, 0.030).

Air Pollution	Age	Respiratory Difficulty	Activity	Gender	Leverage	Cook's D
No	25	No	No	Male	5.775	0.070
No	35	No	No	Male	6.445	0.057
No	45	Yes	No	Male	7.870	0.038

Table 7: Observations with High Leverage and High Cook's Distance

Notably, all three are inactive males not exposed to air pollution - a combination with very few observations (Table 1). This low count naturally produces higher leverage, and there is no apparent theoretical justification to exclude these observations from the dataset. Accordingly, we proceed conservatively and leave the sample intact.

[6] In fact, if normality did hold, given a cutoff of ±2 we would expect just under 5% to exceed this threshold, which would be approximately 14 values.

[7] The leverage of observation i indicates the entry h_{ii} of the model's hat matrix H:

$$H = W^{1/2} X \left(X^T W X \right)^{-1} X^T W^{1/2}$$

where W is the weight matrix from the IRLS algorithm. The trace of the matrix is p, by the cyclic property, so the average leverage value is p/n. Thus, the heuristic $2p/n$ indicates a value with leverage twice the mean. For the scaled leverage, the threshold is naturally 2.

[8] CD is defined as

$$C_i = \frac{\left(\widehat{\beta} - \widehat{\beta}_{-i}\right)^T \left(X^T W X\right) \left(\widehat{\beta} - \widehat{\beta}_{-i}\right)}{p\hat{\phi}} = \frac{\left(\widehat{\beta} - \widehat{\beta}_{-i}\right)^T \left(X^T W X\right) \left(\widehat{\beta} - \widehat{\beta}_{-i}\right)}{8}$$

where $\hat{\phi}$ would estimate the scaling parameter, which is 1 for the Bernoulli GLM. This represents the average change in the estimated coefficient vector when removing one observation, scaled by the covariance.

[9] The threshold is based partly on parallelism with linear models, where Cook's Distance follows an F(p, n - p) distribution which has 5% tail bound ≤ 8 for any p, n $>$ 3. Notice, however, that the distribution does not apply to GLMs.

5 Interpretation (Odds Ratios)

In order to comprehensively interpret the model, we must introduce several concepts. First, we define the odds for the model as

$$\text{odds} = \frac{\mu}{1-\mu} = \frac{e^{x^T\beta}}{1+e^{x^T\beta}} \times \left(\frac{1}{1+e^{x^T\beta}}\right)^{-1} = e^{x^T\beta}$$

and thus the log-odds as

$$\text{log-odds} = log\left(\frac{\mu}{1-\mu}\right) = x^T\beta$$

Then, the effect of a unit change in any of the regressors on the log odds will be given by the respective coefficient. If interactions are present, the effect is the sum of the main and interaction coefficients. For instance, with age we will have:

$$\log(\text{odds}(\texttt{age} = x+1)) - \log(\text{odds}(\texttt{age} = x)) = \hat{\beta}_4 + \hat{\beta}_7\texttt{airpollution}$$

which simplifies to $\hat{\beta}_4$ when air pollution is absent and $\hat{\beta}_4 + \hat{\beta}_7$ when present. This provides a blueprint to interpret the coefficients of the model. The coefficients on the main effects represent the additive effect on the log-odds of the dependent variable from a unit increase in the regressor (when the interaction is absent). Similarly, the coefficients on the interaction terms represent the extra additive effect on the log-odds of a unit increase in the regressor when air pollution is present. We can provide 95% confidence intervals for the coefficients, and thus the effect on the log-odds, as follows:

Variable	Estimate	95 % CI
Intercept	1.65	(-0.78, 4.09)
Air pollution (Yes)	-1.55	(-4.48, 1.39)
Age	-0.04	(-0.08, 0.01)
Male	0.11	(-0.96, 1.18)
Active	-2.17	(-3.99, -0.35)
Air pollution × Age	0.08	(0.02, 0.15)
Air pollution × Male	-2.56	(-4.04, -1.07)
Air pollution × Active	3.17	(1.13, 5.22)

Table 8: Logistic Regression Coefficients with 95% Confidence Intervals

Notice that, for the baseline effects on age, gender, and air pollution, the effect is **not** statistically significant, so we cannot offer a strong directional interpretation. We can, however, provide a directional interpretation for activity, with the log-odds of RD between -3.99 and -0.35 lower for active individuals not exposed to air pollution. Conversely, individuals exposed to air pollution who are active will have log-odds of RD between 1.13 and 5.22 higher than active individuals who were not exposed to air pollution (interaction). Similar interpretations follow for the other coefficients.

While the additive effect on the log-odds is the most immediate interpretation, the effect on the odds is more intuitive. Following from the example of `age` above, we can find the effect on the Odds, i.e., the Odds Ratio (OR)

$$\begin{aligned} OR = \frac{\text{odds}(\texttt{age}+1)}{\text{odds}(\texttt{age})} &= e^{(\log(\text{odds}(\texttt{age}=x+1) - \log(\text{odds}(\texttt{age}=x))} \\ &= e^{(\hat{\beta}_4 + \hat{\beta}_7\texttt{airpollution})} \end{aligned}$$

Since it affects the ratio, the exponentiated effect is **multiplicative**, rather than additive as in the effect on the log-odds.

We now compute confidence intervals and estimates for the ORs. The coefficients are correlated (the covariance matrix is not diagonal), so we must compute the standard errors for the change in the log-odds for every combination of the interactions using the delta method. The calculation is illustrated in the below for age (10-year increase) and further down for air pollution, since gender and activity are structurally similar to age, *mutatis mutandis*, while air pollution is more involved. After the derivation, two tables provide an exhaustive summary of all the multiplicative effects on the odds.[10]

Let the coefficient vector be $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)^\top$, where β_4 is the age main effect and β_7 is the age×airpollution interaction (in line with the Modelling section above). Write $\widehat{V} = \widehat{\mathrm{Var}}(\widehat{\beta})$ for the estimated covariance matrix.

First, consider a 10-year increase in age when air pollution is absent.

$$\text{The effect on the log-odds is:} \quad \Delta_{\text{age}|\text{airpollution}=0} = 10 \times \beta_4,$$

so $\widehat{\mathrm{SE}}$ is the square root of the corresponding diagonal entry, scaled by the age increase:

$$\widehat{\mathrm{SE}}\big(\Delta_{\text{age}|\text{airpollution}=0}\big) = 10 \times \sqrt{\widehat{V}_{44}}.$$

Then, consider the effect of a 10-year increase in age when air pollution is present.

$$\text{The effect on the log-odds is:} \quad \Delta_{\text{age}|\text{airpollution}=1} = 10 \times (\beta_4 + \beta_7).$$

By the delta method by the symmetry of the covariance matrix,

$$\widehat{\mathrm{SE}}\big(\Delta_{\text{age}|\text{airpollution}=1}\big) = \widehat{\mathrm{Var}}\big(\Delta_{\text{age}|\text{airpollution}=1}\big) = 10 \times \sqrt{\widehat{V}_{44} + \widehat{V}_{77} + 2\widehat{V}_{47}},$$

where $\widehat{V}_{jk}$ denotes the (j,k)-entry of $\widehat{V}$.

For any estimated linear effect $\widehat{\Delta}$ with estimated standard error $\widehat{\mathrm{SE}}(\widehat{\Delta})$, a (Wald) (1-$\alpha$) CI for the log-odds change is

$$\widehat{\Delta} \ \pm \ z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\widehat{\Delta}).$$

and thus we can derive the CI for the OR as

$$\mathrm{CI}_{\mathrm{OR}} = \exp\Big(\widehat{\Delta} \pm z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\widehat{\Delta})\Big).$$

The same derivation applies to gender and activity,without the 10x factor. The decision to use 10-year increments for age comes from the fact that age is meant to represent 10-year bands, and also for ease of interpretation since one year has a relatively small effect.

Variable	Air Pollution	Odds Ratio	95 % CI
Age (10-year increase)	Present	1.58	(1.04, 2.40)
	Absent	0.70	(0.43, 1.14)
Gender (Male vs Female)	Present	0.09	(0.03, 0.24)
	Absent	1.11	(0.38, 3.25)
Activity (Active vs Inactive)	Present	2.72	(1.07, 6.93)
	Absent	0.11	(0.02, 0.70)

Table 9: Odds Ratios with 95% CI: age, gender, activity

Table 9 summarises the results: the odds of experiencing RD are between 1.04x and 2.40x higher, with 95% confidence, when increasing the age of a subject exposed to air pollution by 10 years. Conversely, if the subject is not exposed to air pollution, age is ambiguous (insignificant main effect), with the odds multiplicative factor between 0.43x and 1.14x for a 10-year increase.

[10] refer to the R-code in the Appendix for the implementation

Similarly, the odds of RD for men are much lower than women when exposed to air pollution, at just 3-24%, whereas we find no significant effect when pollution is absent (0.38-3.25, very wide interval). More strikingly, when air pollution is present, the odds of RD are much higher among active subjects, between 1.07 and 6.93, whereas they are much lower when air pollution is absent, at just 2-70% of the inactive population.[11]

Now, we can exemplify the effect of air pollution when all interactions are present, with gender = 1 (male), activity = 1 (active) and age = 35.

$$\begin{aligned}\Delta_{\text{airpollution}|g=1,a=1,\text{age}=35} &= \beta_2 + \beta_5 \cdot (\text{activity} = 1) + \beta_6 \cdot (\text{gender} = 1) + \beta_7 \cdot (\text{age} = 35) \\ &= \beta_2 + \beta_5 + \beta_6 + 35\,\beta_7.\end{aligned}$$

Let $c = (0, 0, 1, 0, 0, 1, 1, 35)^\top$. Then,

$$\widehat{\text{Var}}\big(\Delta_{\text{air}|g=1,a=1,\text{age}=35}\big) = c^\top \widehat{V}\, c, \quad \widehat{\text{SE}}\big(\Delta_{\text{air}|g=1,a=1,\text{age}=35}\big) = \sqrt{c^\top \widehat{V}\, c}.$$

Expanding in covariance notation:

$$\begin{aligned}\widehat{\text{Var}}(\Delta) = \widehat{V}_{22} + \widehat{V}_{55} + \widehat{V}_{66} + 35^2\widehat{V}_{77} \\ + 2\big(\widehat{V}_{25} + \widehat{V}_{26} + 35\widehat{V}_{27} + \widehat{V}_{56} + 35\widehat{V}_{57} + 35\widehat{V}_{67}\big).\end{aligned}$$

The standard error and confidence intervals for the odds ratio follow as above. Table 10 breaks down all the effects of air pollution on the odds by the values of the other regressors. Some particularly interesting effects are exposure to air pollution for 25-year-old inactive men, where the odds of developing RD are **lower** (0.02x to 0.87x) than when air pollution is absent, and active females of all ages, where air pollution substantially increases the odds of RD, ranging from 9.74x-154.25x for 25-year-old women, to 61.10x-3205.61x for 55-year-old women.

Gender	Activity	Age	Odds Ratio	95 % OR Confidence Interval
M	Active	25	3.01	(1.17, 7.76)
M	Active	35	6.77	(3.16, 14.52)
M	Active	45	15.24	(5.38, 43.20)
M	Active	55	34.33	(7.28, 161.87)
F	Active	25	38.75	(9.74, 154.25)
F	Active	35	87.27	(23.14, 329.03)
F	Active	45	196.52	(41.31, 934.80)
F	Active	55	442.55	(61.10, 3205.61)
M	Inactive	25	0.13	(0.02, 0.87)
M	Inactive	35	0.28	(0.04, 1.94)
M	Inactive	45	0.64	(0.08, 5.32)
M	Inactive	55	1.44	(0.12, 16.98)
F	Inactive	25	1.62	(0.16, 16.54)
F	Inactive	35	3.65	(0.35, 38.36)
F	Inactive	45	8.22	(0.64, 105.01)
F	Inactive	55	18.50	(1.04, 328.00)

Table 10: Odds Ratios with 95% CI: air pollution

Of course, we should refrain from causal interpretations of the results, since it seems highly unlikely that exposure to air pollution should cause a decline in RD in any demographic.

[11] This is in line with the interpretation provided in the EDA section.

6 Prediction

Finally, we estimate the probability of developing RD for a 45-year-old woman who was active during an extreme air pollution event, and an 85-year-old inactive man who was exposed to air pollution. First, notice that the first datapoint is an in-sample prediction, since from Table 1 we see only five individuals that fall into this description, all of whom developed RD. Conversely, the 85-year-old man is an out-of-sample prediction, since we do not have any observations above the age of 55. Therefore, we expect the predictions in the second case to be less accurate (wider confidence interval) and less reliable, given extreme extrapolation.

Activity	Age	Gender	Air Pollution	Predicted Probability	95 % CI
Yes	45	Female	Yes	0.960	(0.889, 0.986)
No	85	Male	Yes	0.826	(0.293, 0.982)

Table 11: Predicted Probabilities of Respiratory Disease - Full Model

As expected, the predicted probability for the first observation is high, in line with the prevalence of RD in that demographic, with a relatively tight confidence interval. Again as expected, the second interval is very wide, offering limited guidance for the probability of RD. The predicted value is high at 0.826, in line with the marginal effect of 0.04 (-0.04 + 0.08, see Table 8) from an additional year of age during an extreme air pollution event. However, this out-of-sample prediction must be treated cautiously, since it relies on substantial extrapolation beyond the observed age range.

7 Appendix

The table below summarises the result of the full GLM with all main effects and interactions with air pollution, but treating age as a categorical variable. This choice is more reflective of the data, since age represents 10-year bands rather than the actual age of the participants. This makes the continuous interpretation more fragile, since the coefficient on age indicates the effect of an additional year of age on the log-odds of RD, which requires some level of extrapolation. Moreover, an unwanted effect of the continuous treatment is that the fitted values for, say, 24 and 26 will be different, when they could presumably be the same, since may be in the same 10-year band. That is also the reason the paper displays the effect of a 10-year increase in age, rather than a 1-year increase. While more reflective of the data, the model below is also more restrictive, since it does not allow extrapolation for ages beyond the upper limit of 55. Also, notice that the model below includes 4 more coefficients, which leads to higher AIC (288.6 vs 282.1).

	Dependent variable:
	Respiratory difficulty
Air pollution (Yes)	0.44 (1.21)
Age 35	−0.50 (0.60)
Age 45	−0.69 (0.88)
Age 55	−1.02 (0.82)
Male	0.10 (0.55)
Active	−2.17* (0.94)
Air pollution × Age 35	0.87 (0.76)
Air pollution × Age 45	1.23 (1.03)
Air pollution × Age 55	3.22* (1.37)
Air pollution × Male	−2.46** (0.76)
Air pollution × Active	3.15** (1.05)
Constant	0.81 (1.06)
Observations	285
Log Likelihood	−132.32
Akaike Inf. Crit.	288.63
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 12: Logistic Regression of Respiratory difficulty, age as factor

The plot below displays the standardised residuals against the predicted responses for the chosen model (glm1 or the "full" model). We do not see any differences in terms of outlying observations from the plot, while the two clusters appear non-monotonic due to the sigmoid transformation.

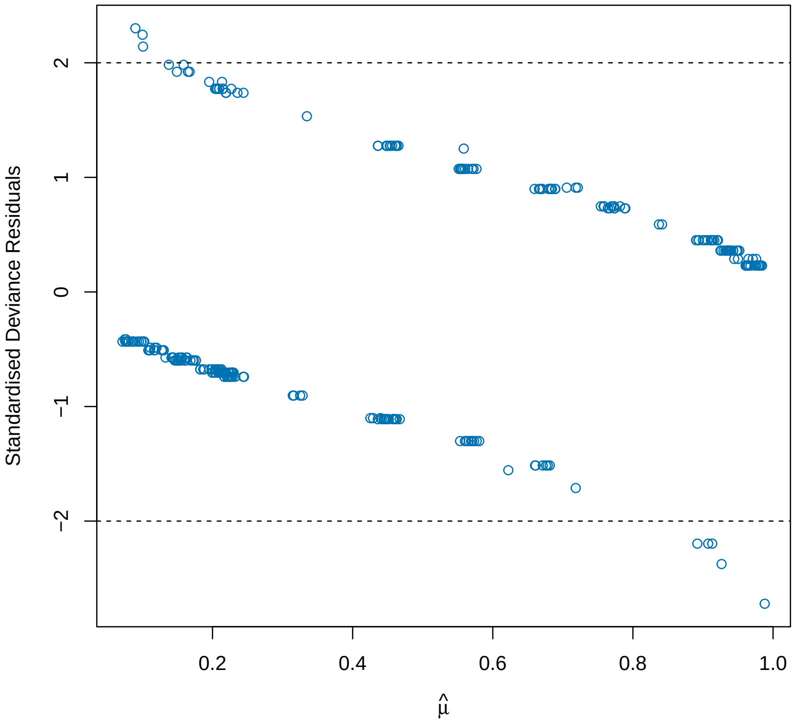

Figure 6: Plot of standardised deviance residuals vs predicted responses

The following plots show the model diagnostics plots for the reduced model (glm5). Interestingly, we see fewer outlying observations in the standardised residual plot. Notice that this may be partly due to fewer fitted values for a more constrained model. The plot uses jitter to display the concentration of values, but of course the total number of fitted values will at most be 18, since we only have 18 possible combinations of the regressors in then model and thus 36 possible values of the residuals. Therefore, fewer outliers will be expected. The comments in the model diagnostics section otherwise extend to the remaining plots. The QQ plot is somewhat more anomalous but still merely reflects the limited number of residual values.

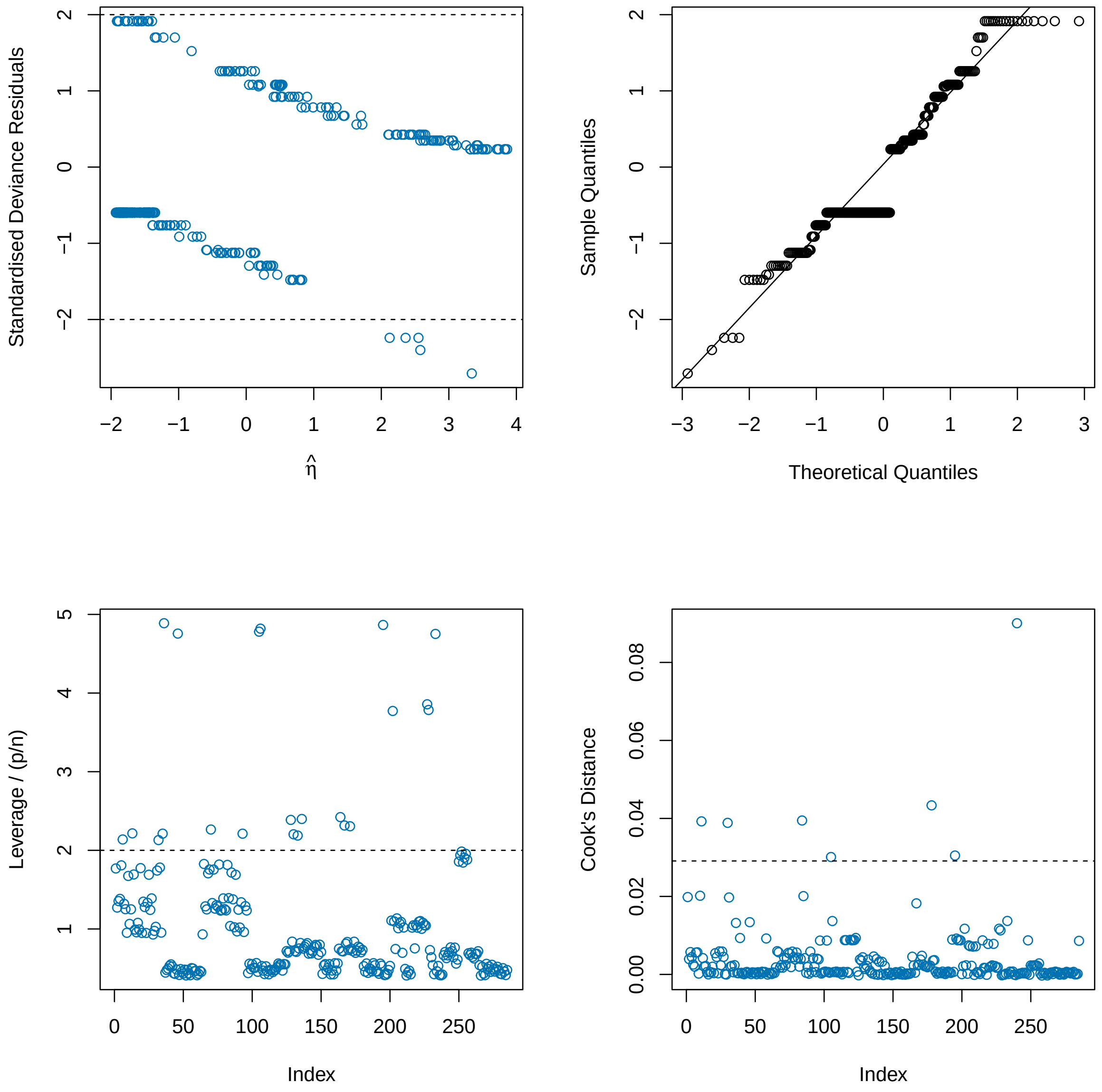

Figure 7: In order Left to Right, Top to Bottom: residuals vs fitted linear regressor, residual QQ-plot, leverage plot, Cook's distance plot

Finally, the table below illustrates the odds ratio and respective confidence intervals for the reduced model. These are broadly in line with the full model, but notice that this model has much lower variance for high values of age. This is because the standard error of age is larger in the full model (0.03 vs 0.02), which has a minor effect when age is small but increases exponentially with age.

Gender	Activity	Age	Odds Ratio	95% OR Confidence Interval
M	Active	25	4.42	(1.39, 14.10)
M	Active	35	6.63	(2.22, 19.84)
M	Active	45	9.96	(3.17, 31.25)
M	Active	55	14.95	(4.11, 54.44)
F	Active	25	55.59	(19.38, 159.46)
F	Active	35	83.46	(28.42, 245.11)
F	Active	45	125.30	(37.35, 420.34)
F	Active	55	188.12	(45.36, 780.16)
M	Inactive	25	0.22	(0.07, 0.68)
M	Inactive	35	0.33	(0.08, 1.29)
M	Inactive	45	0.49	(0.10, 2.54)
M	Inactive	55	0.74	(0.11, 5.16)
F	Inactive	25	2.76	(1.13, 6.73)
F	Inactive	35	4.15	(1.19, 14.44)
F	Inactive	45	6.23	(1.25, 30.96)
F	Inactive	55	9.35	(1.32, 66.38)

Table 13: Odds Ratios with 95% CI: air pollution, under the reduced model (glm5)

Below is the complete R code used for this analysis.

```r
data <- read.csv("airpollution.csv", stringsAsFactors = TRUE)
packages <- c("ggplot2", "xtable", "stargazer", "rsq", "emmeans")
install.packages(packages)
library(xtable)
library(ggplot2)
library(stargazer)
library(rsq)
library(emmeans)

#### DEFACTORED BECAUSE OF WHAT SAID IN Qe ABOUT OOS AGE PREDICT
# Convert age to factors, since they represent age ranges
# data$age <- as.factor(data$age)
attach(data) #attach lets us call the columns directly, making code more concise
summary(data)


###############################################################################
### A - Exploratory data analysis
###############################################################################

# summary table with counts and proportions for each variable
make_summary <- function(var, name) {
  tbl <- table(var)
  data.frame(
    Variable = name,
    Level = names(tbl),
    Count = as.numeric(tbl),
    Percent = round(100 * as.numeric(prop.table(tbl)), 1)
  )
}

summary_table <- rbind(
  make_summary(respd, "Respiratory Difficulty"),
  make_summary(airpollution, "Air pollution"),
  make_summary(activity, "Activity"),
  make_summary(gender, "Gender"),
  make_summary(age, "Age")
)


```

```r
stargazer::stargazer(summary_table,
                     type = "latex",
                     summary = FALSE,
                     rownames = FALSE,
                     digits = 2,
                     font.size = "scriptsize",
                     out = "tables/summary_table.tex")

# proportion with respd == "Yes"
prop_tbl <- aggregate(I(respd == "Yes") ~ airpollution + activity + gender + age,
                      data = data, FUN = mean)
names(prop_tbl)[5] <- "PropYes"

# count for each combination
count_tbl <- aggregate(respd ~ airpollution + activity + gender + age,
                       data = data, FUN = length)
names(count_tbl)[5] <- "Count"

prop_tbl <- merge(prop_tbl, count_tbl,
                  by = c("airpollution", "activity", "gender", "age"))

prop_tbl <- prop_tbl[order(prop_tbl$airpollution, prop_tbl$activity, prop_tbl$gender, prop_tbl$age), ]
prop_tbl$PropYes <- sprintf("%.2f", prop_tbl$PropYes)

names(prop_tbl) <- c("Air Pollution", "Activity", "Gender", "Age",
                     "Proportion", "Count")

stargazer::stargazer(prop_tbl,
                     type = "latex",
                     summary = FALSE,
                     rownames = FALSE,
                     digits = 2,
                     font.size = "scriptsize",
                     out = "tables/combo_prop_table.tex")


# PLOTS
plot_names <- list("Air pollution", "Age", "Respiratory Difficulty", "Activity", "Gender")
names(plot_names) <- names(data)

blue_palette <- c("#084C8D", "#0072B2", "#56B4E9", "#A6D8F0")

# Bar plot for age
pdf("figs/age_barplot_cex1.pdf", width = 6, height = 6)
par(mfrow = c(1,1), cex = 1.1)
barplot(table(age),
        col = blue_palette,
        xlab = "Age band",
        ylab = "Count")
dev.off()

# Plot proportions with respiratory Difficulty for binary categories
prop_plot <- function(x, y, text = FALSE) {
  x_label <- plot_names[[deparse(substitute(x))]]
  y_label <- plot_names[[deparse(substitute(y))]]
  table <- table(x, y)
  print(table)
  prop <- prop.table(table, 2)
  print(prop)
  bp <- barplot(prop,
          col = blue_palette[c(2, 4)],
          xlab = y_label,
          ylab = "Proportion of cohort",
          beside = TRUE,
          ylim = c(0, max(prop) + 0.1))
  if(text == TRUE) {text(x = bp,
       y = prop,
       labels = signif(prop, 2),
       pos = 3,
       cex = 0.9)}
}

pdf("figs/prop_plots_initial.pdf", width = 7, height = 7)
layout(matrix(c(1, 2, 3, 4, 5, 5), nrow = 3, byrow = TRUE),
       heights = c(1, 1, 0.3))  # Bottom row is shorter

par(mar = c(4, 4, 2, 2))
prop_plot(respd, gender, text = TRUE)
prop_plot(respd, activity, text = TRUE)
prop_plot(respd, airpollution, text = TRUE)
prop_plot(respd, age, text = TRUE)

par(mar = c(0, 0, 0, 0))
plot.new()
legend("center",
       legend = c("No respiratory difficulty", "Respiratory difficulty"),
       fill = blue_palette[c(2, 4)],
       bty = "n",
       horiz = TRUE)

dev.off()

# Now breaking the data down by airpollution = yes or no, showing that gender
# is irrelevant when there is no air pollution, but very relevant when there is
# whereas activity seems to have opposite effects with/ without air pollution

data_pollutionyes <- subset(data, subset = airpollution == "Yes")
data_pollutionno <- subset(data, subset =airpollution == "No")

plot_names <- list(
  respd                         = "Respiratory difficulty",
  gender                        = "Gender",
  activity                      = "Activity",
  age                           = "Age",
  "data_pollutionno$respd"      = "Respiratory difficulty",
  "data_pollutionno$gender"     = "Gender",
  "data_pollutionno$activity"   = "Activity",
  "data_pollutionno$age"        = "Age",
  "data_pollutionyes$respd"     = "Respiratory difficulty",
  "data_pollutionyes$gender"    = "Gender",
  "data_pollutionyes$activity"  = "Activity",
  "data_pollutionyes$age"       = "Age"
)

pdf("figs/barplots_noairpollution.pdf", width = 7, height = 4)
par(mar = c(5.1, 4.1, 4.1, 2.1))
layout(matrix(c(1, 2, 3, 4, 4, 4), nrow = 2, byrow = TRUE),
       heights = c(1, 0.1))  # Bottom row is shorter

prop_plot(data_pollutionno$respd, data_pollutionno$gender)
prop_plot(data_pollutionno$respd, data_pollutionno$activity)
prop_plot(data_pollutionno$respd, data_pollutionno$age)

```

```r
par(mar = c(0, 0, 0, 0))
plot.new()

legend("center",
       legend = c("No respiratory difficulty", "Respiratory difficulty"),
       fill = blue_palette[c(2, 4)],
       bty = "n",
       horiz = TRUE)
dev.off()

pdf("figs/barplots_yesairpollution.pdf", width = 7, height = 4)
par(mar = c(5.1, 4.1, 4.1, 2.1))
layout(matrix(c(1, 2, 3, 4, 4, 4), nrow = 2, byrow = TRUE),
       heights = c(1, 0.1))  # Bottom row is shorter

prop_plot(data_pollutionyes$respd, data_pollutionyes$gender)
prop_plot(data_pollutionyes$respd, data_pollutionyes$activity)
prop_plot(data_pollutionyes$respd, data_pollutionyes$age)

par(mar = c(0, 0, 0, 0))
plot.new()

legend("center",
       legend = c("No respiratory difficulty", "Respiratory difficulty"),
       fill = blue_palette[c(2, 4)],
       bty = "n",
       horiz = TRUE)

dev.off()

##############################################################################
### B - MODELLING
##############################################################################

# Baseline model: all interactions between airpollution and other explanatory variables

resp.glm1 <- glm(respd ~ airpollution * (age + gender + activity),
                 data = data,
                 family = binomial)

summary(resp.glm1)

anova(resp.glm1) #initial anova suggests all interactions are significant
# thus by the hierarchy principle we should keep all the lower order terms too
# by the wald we should get rid of all baselines except activity. Gender is especially insignificant
# by the LR gender is extremely significant and activity insignificant.
# this bc anova sequential and goes gender first, then activity and interactions. see glm1.1 below

LRT <- function(baseline, restricted) {
  LRT <- restricted$deviance - baseline$deviance
  p_val <- 1 - pchisq(LRT, restricted$df.residual - baseline$df.residual)
  out <- list(
    LRT = LRT,
    p_value = p_val
  )
  return(out)
}

# likelihood ratio for model useless:

```

```r
resp.glm0 <- glm(respd ~ 1, data = data, family = binomial)

L1 <- LRT(resp.glm1, resp.glm0) # approximately 0 --> the model is not useless

# Consider dropping any of the interactions
# All effects are significant, do not drop any. Agrees with Wald

resp.glm2 <- glm(respd ~ airpollution * (gender + activity) + age,
                 data = data,
                 family = binomial)

L2 <- LRT(resp.glm1, resp.glm2)
resp.glm3 <- glm(respd ~ airpollution * (gender + age) + activity,
                data = data,
                family = binomial)
L3 <- LRT(resp.glm1, resp.glm3)

resp.glm4 <- glm(respd ~ airpollution * (activity + age) + gender,
                 data = data,
                 family = binomial)
L4 <- LRT(resp.glm1, resp.glm4)

# Dropping the baseline effects except activity. This is suggested by Wald
data$agexpollution <- data$age*(data$airpollution == "Yes")
data$genderxpollution <- as.integer(data$gender == "Male" & data$airpollution == "Yes")
data$activityxpollution <- as.integer(data$activity == "Yes" & data$airpollution == "Yes")

resp.glm5 <- glm(respd ~ activity + agexpollution + genderxpollution + activityxpollution,
                 data = data,
                 family = binomial)
L5 <- LRT(resp.glm1, resp.glm5) # p-value is 0.478, suggesting we should drop all baselines!

# drop age baseline
resp.glm6 <- glm(respd ~ airpollution*(gender + activity) + agexpollution,
                 data = data,
                 family = binomial)

L6 <- LRT(resp.glm1, resp.glm6)

# drop gender baseline
resp.glm7 <- glm(respd ~ airpollution*(activity) + agexpollution + genderxpollution,
                 data = data,
                 family = binomial)
summary(resp.glm7)

L7 <- LRT(resp.glm6, resp.glm7)

# drop activity main effect (against Wald)
resp.glm8 <- glm(respd ~ airpollution + activityxpollution + agexpollution + genderxpollution,
                 data = data,
                 family = binomial)
summary(resp.glm8)
L8 <- LRT(resp.glm7, resp.glm8) # reject

# finally, drop airpollution main effect
# since all the coefficients are statistically significant, we should choose final model resp.glm5
# for final verification, see a plot of the AICs and a table of the R^2 KL for all the models

# why we should keep the baseline interaction - the hierarchy principle

data$airpollutionprime <- ifelse((data$airpollution == "Yes"), 0, 1)
resp.glmairpolltoy <- glm(respd ~ airpollutionprime * (age + gender + activity),
                          data = data,
                          family = binomial)

resp.glmairpolltoy_nointeractions <- glm(respd ~ airpollutionprime + airpollutionprime:gender +
                                           airpollutionprime:activity + airpollutionprime:age,
                                         data = data,
                                         family = binomial)
summary(resp.glmairpolltoy_nointeractions)
LRT(resp.glmairpolltoy, resp.glmairpolltoy_nointeractions)

chosen_model <- resp.glm1
reduced_model <- resp.glm5

########### TABLES ###############
stargazer(resp.glm1,
          type = "latex",
          title = "Logistic Regression of Respiratory difficulty",
          dep.var.labels = "Respiratory difficulty",
          covariate.labels = c(
            "Air pollution (Yes)",
            "Age",
            "Male",
            "Active",
            "Air pollution    Age",
            "Air pollution    Male",
            "Air pollution    Active"
          ),
          digits = 2,
          star.cutoffs = c(0.05, 0.01, 0.001),
          no.space = TRUE,
          float = TRUE,
          font.size = "scriptsize",
          report = "vc*sp",
          single.row = TRUE,
          out = "tables/resp.glm1.tex")

stargazer(resp.glm5,
          type = "latex",
          title = "Logistic Regression of Respiratory Difficulty",
          dep.var.labels = "Respiratory Difficulty",
          covariate.labels = c(
            "Active",
            "Air pollution    Age",
            "Air pollution    Male",
            "Air pollution    Active"
          ),
          digits = 2,
          star.cutoffs = c(0.05, 0.01, 0.001),
          no.space = TRUE,
          float = TRUE,
          single.row = TRUE,
          font.size = "scriptsize",
          report = "vc*sp",
```

```r
          out = "tables/resp.glm5.tex")

likelihood_ratios <- round(c(L1$LRT, L2$LRT, L3$LRT, L4$LRT, L5$LRT, L6$LRT, L7$LRT, L8$LRT), 2)
p_values <- round(c(L1$p_value, L2$p_value, L3$p_value, L4$p_value, L5$p_value, L6$p_value, L7$p_value, L8$p_value), 3)
full_model <- c("glm1", "glm1", "glm1", "glm1", "glm1", "glm1", "glm6", "glm7")
restricted_model <- c("glm0","glm2","glm3","glm4","glm5","glm6","glm7","glm8")
drop_vector <- c(
  "All (null test)",
  "Age x Air Pollution",
  "Activity x Air Pollution",
  "Gender x Air Pollution",
  "Air Pollution, Age, Gender",
  "Age",
  "Gender",
  "Activity"
)

likelihood_ratio_table <- data.frame(
  "Full Model" = full_model,
  "Restricted Model" = restricted_model,
  "Drop:" = drop_vector,
  "Likelihood Ratio Statistic" = likelihood_ratios,
  "P-values" = p_values)

stargazer(likelihood_ratio_table,
          type = "latex",
          summary = FALSE,
          rownames = FALSE,
          digits = 2,
          font.size = "scriptsize",
          out = "tables/likelihood_ratio_table.tex")

models <- list(resp.glm1, resp.glm2, resp.glm3, resp.glm4, resp.glm5, resp.glm6, resp.glm7, resp.glm8)
model_names <- c("glm1", "glm2", "glm3", "glm4", "glm5", "glm6", "glm7", "glm8")
aics <- sapply(models, AIC)
bics <- sapply(models, BIC)
deviances <- sapply(models, deviance)
rsq_vals <- round(sapply(models, rsq.kl), 3)

model_desc <- c(
  "Full model",
  "Drop Age x Air Pollution",
  "Drop Activity x Air Pollution",
  "Drop Gender x Air Pollution",
  "Drop Air Pollution, Age, Gender",
  "Drop Age",
  "Drop Age, Gender",
  "Drop Activity, Age, Gender"
)

aic_bic_etc_table <- data.frame(
  "Model" = model_names,
  "Description" = model_desc,
  "Residual Deviance" = deviances,
  "AIC" = aics,
  "BIC" = bics,
  "R^2 Kullback-Leibler" = rsq_vals
)

```

```
stargazer(aic_bic_etc_table,
          type = "latex",
          font.size = "scriptsize",
          summary = FALSE,
          rownames =FALSE,
          digits = 2,
          out = "tables/aic_bic_etc_table.tex")

stargazer(resp.glmairpolltoy,
          type = "latex",
          title = "Toy model - Logistic Regression of Respiratory difficulty",
          dep.var.labels = "Respiratory difficulty",
          covariate.labels = c(
            "Air pollution (No)",
            "Age",
            "Male",
            "Active",
            "Air pollution    Age",
            "Air pollution    Male",
            "Air pollution    Active"
          ),
          digits = 2,
          star.cutoffs = c(0.05, 0.01, 0.001),
          no.space = TRUE,
          float = TRUE,
          font.size = "scriptsize",
          report = "vc*sp",
          single.row = TRUE,
          out = "tables/toymodel.tex")



################################################################################
### C - Model diagnostics
################################################################################

# Residual plots

set.seed(30)

pdf("figs/residual_plots.pdf", width = 9.5, height = 10)
par(mfrow = c(2, 2), margin(5, 5, 3, 3), cex = 1)

plot(jitter(predict(chosen_model, type = 'link'), 10),
                       jitter(rstandard(chosen_model), 0), col = blue_palette[2],
                       xlab = expression(hat(eta)),
                       ylab = "Standardised Deviance Residuals")
abline(a = -2, b = 0, lty = 2)
abline(a = 2, b = 0, lty = 2)

# QQ plot

qqnorm(rstandard(chosen_model), main = "")
qqline(rstandard(chosen_model))

# Leverage

p <- chosen_model$df.null - chosen_model$df.residual + 1
n <- nrow(data)

plot(jitter(influence(chosen_model)$hat/(p/n), 200), ylab = "Leverage / (p/n)",
     col = blue_palette[2])
```

```r
abline(2, 0, lty = 2)

# Cook's distance

plot(jitter(cooks.distance(chosen_model), 100), ylab = "Cook's Distance",
     col = blue_palette[2])
abline(8/(n - 2 * p), 0, lty = 2)
dev.off()

leverage <- influence(chosen_model)$hat
cooks_d <- cooks.distance(chosen_model)

# Thresholds
leverage_threshold <- 2 * (p / n)
cooks_threshold <- 8 / (n - 2 * p)

# Problematic observations
high_leverage <- which(leverage > leverage_threshold)
high_cooks <- which(cooks_d > cooks_threshold)

# Summary table of problematic observations
problematic <- intersect(high_leverage, high_cooks)
diag_table <- data.frame(
  data[problematic, 1:5],
  Leverage = round(leverage[problematic], 4) / (p/n),
  Cooks_D = round(cooks_d[problematic], 4)
)
print(diag_table)

colnames(diag_table) <- c("Air Pollution","Age", "Respiratory Difficulty", "Activity","Gender", "Leverage", "Cook's D")

stargazer(diag_table,
          summary = FALSE,
          rownames = FALSE,
          type = "latex",
          title = "Observations with High Leverage and High Cook's Distance",
          font.size = "scriptsize",
          float = TRUE,
          out = "tables/diag_table.tex")

##############################################################################
### D - INTERPRETATION
##############################################################################

beta <- coef(chosen_model)
se <- sqrt(diag(vcov(chosen_model)))
z <- qnorm(0.975)

# Calculate CIs
ci_lower <- beta - z * se
ci_upper <- beta + z * se

# Build table
ci_table <- data.frame(
  Variable = c(
    "Intercept",
    "Air pollution (Yes)",
    "Age",
    "Male",
    "Active",
    "Air pollution    Age",
    "Air pollution    Male",
    "Air pollution    Active"
  ),
  Estimate = sprintf("%.2f", beta),
  CI = sprintf("(%.2f, %.2f)", ci_lower, ci_upper)
)

colnames(ci_table) <- c("Variable", "Estimate", "95\\% CI")

stargazer(ci_table,
          summary = FALSE,
          rownames = FALSE,
          type = "latex",
          title = "Logistic Regression Coefficients with 95\\% Confidence Intervals",
          font.size = "scriptsize",
          float = TRUE,
          out = "tables/ci_table.tex")



odds_ratio_and_ci <- function(variable) {
  beta <- coef(chosen_model)
  V <- vcov(chosen_model)
  interaction <- paste("airpollutionYes", variable, sep = ":")

  # Air pollution present
  logor_airpollution1 <- beta[variable] + beta[interaction]
  variance_1 <- V[variable, variable] + V[interaction, interaction] + 2 * V[variable, interaction]
  se_1 <- sqrt(variance_1)
  logor_airpollution1_ci <- logor_airpollution1 + 1.96 * se_1 * c(-1, 1)

  # Air pollution absent
  logor_airpollution0 <- beta[variable]
  variance_0 <- V[variable, variable]
  se_0 <- sqrt(variance_0)
  logor_airpollution0_ci <- logor_airpollution0 + 1.96 * se_0 * c(-1, 1)

  # For age: multiply log-odds by 10
  if (variable == "age") {
    logor_airpollution1 <- logor_airpollution1 * 10
    logor_airpollution1_ci <- logor_airpollution1_ci * 10
    logor_airpollution0 <- logor_airpollution0 * 10
    logor_airpollution0_ci <- logor_airpollution0_ci * 10
  }

  data.frame(
    Air_Pollution = c("Present", "Absent"),
    Odds_Ratio = sprintf("%.2f", exp(c(logor_airpollution1, logor_airpollution0))),
    CI = c(
      sprintf("(%.2f, %.2f)", exp(logor_airpollution1_ci[1]), exp(logor_airpollution1_ci[2])),
      sprintf("(%.2f, %.2f)", exp(logor_airpollution0_ci[1]), exp(logor_airpollution0_ci[2]))
    )
  )
}

variable_names <- names(beta)[3:5]
variable_nicenames <- c("Age", "Gender", "Activity")

# calculate using the delta method
```

```r
odds_ratio_and_ci_airpollution <- function(age, gender, activity) {
  beta <- coef(chosen_model)
  V <- vcov(chosen_model)
  ap <- "airpollutionYes"
  apag <- "airpollutionYes:age"
  apge <- "airpollutionYes:genderMale"
  apac <- "airpollutionYes:activityYes"
  var <- V[ap, ap] +
    sum(c(V[apag, apag], V[apge, apge], V[apac, apac]) * c(age, gender, activity)**2) +
    2 * sum(c(V[ap, apag], V[ap, apge], V[ap, apac]) * c(age, gender, activity)) +
    2 * age * sum(c(V[apag, apge], V[apag, apac]) * c(gender, activity)) +
    2 * gender * activity * V[apge, apac]
  se <- sqrt(var)
  log_odds_ratio <- sum(beta[c(ap, apag, apge, apac)] * c(1, age, gender, activity))
  log_odds_ci <- log_odds_ratio + 1.96 * c(-1, 1) * se
  odds_ratio <- exp(log_odds_ratio)
  odds_ci <- exp(log_odds_ci)
  out <- c(
    "Odds Ratio" = round(odds_ratio, 2),
    "Odds Ratio CI" = round(odds_ci, 2)
  )
  return(out)
}

######### TABLES #############

# OTHER ODDS

combined_table <- rbind(
  cbind(Variable = "Age (10-year increase)", odds_ratio_and_ci("age")),
  cbind(Variable = "Gender (Male vs Female)", odds_ratio_and_ci("genderMale")),
  cbind(Variable = "Activity (Active vs Inactive)", odds_ratio_and_ci("activityYes"))
)
combined_table$Variable[c(2,4,6)] <- ""  # Remove duplicate labels

colnames(combined_table) <- c("Variable", "Air Pollution", "Odds Ratio", "95 % CI")

stargazer(combined_table,
          summary = FALSE,
          rownames = FALSE,
          type = "latex",
          font.size = "scriptsize",
          digits = 2,
          out = "tables/combined_odds_ratios_table.tex")

# AIRPOLLUTION ODDS

combos <- expand.grid(
  age = c(25, 35, 45, 55),
  gender = c(1, 0),
  activity = c(1, 0)
)

results <- t(mapply(
  gender = combos$gender,
  activity = combos$activity,
  age = combos$age,
  odds_ratio_and_ci_airpollution
```

```r
))

odds_table <- cbind(combos, results)
odds_table$gender <- ifelse(odds_table$gender == 1, "\texttt{M}", "\texttt{F}")
odds_table$activity <- ifelse(odds_table$activity == 1, "Active", "Inactive")
odds_table$CI <- sprintf("(%.2f, %.2f)", odds_table[,5], odds_table[,6])

odds_table <- odds_table[, c("gender", "activity", "age", "Odds Ratio", "CI")]
colnames(odds_table) <- c("Gender", "Activity", "Age", "Odds Ratio", "95\\% OR Confidence Interval")

stargazer(odds_table,
          type = "latex",
          font.size = "scriptsize",
          summary = FALSE,
          rownames =FALSE,
          digits = 2,
          out = "tables/odds_table.tex")


##############################################################################
### E - Prediction
##############################################################################
# new prediction data frames
predict_1 <- data.frame(
  activity = "Yes",
  age = 45,
  gender = "Female",
  airpollution = "Yes"
)

predict_2 <- data.frame(
  activity = "No",
  age = 85,
  gender = "Male",
  airpollution = "Yes"
)

# linear regression first (for conf. int’s)
pred1_link <- predict(chosen_model, newdata = predict_1, type = "link", se.fit = TRUE)
pred2_link <- predict(chosen_model, newdata = predict_2, type = "link", se.fit = TRUE)

get_prob_ci <- function(pred_link) {
  fit <- pred_link$fit
  se <- pred_link$se.fit
  ci_link <- fit + 1.96 * se * c(-1, 1)
  prob <- plogis(fit)
  ci_prob <- plogis(ci_link)
  return(c(prob, ci_prob))
}

prob1 <- get_prob_ci(pred1_link)
prob2 <- get_prob_ci(pred2_link)

pred_table <- data.frame(
  Activity = c("Yes", "No"),
  Age = c(45, 85),
  Gender = c("Female", "Male"),
  Air_Pollution = c("Yes", "Yes"),
  Probability = sprintf("%.3f", c(prob1[1], prob2[1])),
  CI = c(
    sprintf("(%.3f, %.3f)", prob1[2], prob1[3]),
    sprintf("(%.3f, %.3f)", prob2[2], prob2[3])
  )
)

colnames(pred_table) <- c("Activity", "Age", "Gender", "Air Pollution",
                          "Predicted Probability", "95 % CI")

stargazer(pred_table,
          summary = FALSE,
          rownames = FALSE,
          type = "latex",
          title = "Predicted Probabilities of Respiratory Difficulty - Full Model",
          font.size = "scriptsize",
          out = "tables/predicted_probs.tex")


# new prediction data frames - reduced model
predict_1_rm <- data.frame(
  activity = "Yes",
  agexpollution = 45,
  genderxpollution = 0,
  activityxpollution = 1
)

predict_2_rm <- data.frame(
  activity = "No",
  agexpollution = 85,
  genderxpollution = 1,
  activityxpollution = 0
)

# linear regression first (for confidence intervals)
pred1_link_rm <- predict(reduced_model, newdata = predict_1_rm, type = "link", se.fit = TRUE)
pred2_link_rm <- predict(reduced_model, newdata = predict_2_rm, type = "link", se.fit = TRUE)

get_prob_ci(pred1_link_rm)  # Female, 45, active, pollution
get_prob_ci(pred2_link_rm)  # Male, 85, inactive, pollution

# notice that while we can get confidence intervals for the *probabilities*, we cannot get
# prediction intervals for the probabilities, since the realisation will always jsut be 0 or 1 for resp
# so prediction intervals are meaningless here

##############################################################################
### APPENDIX CODE
##############################################################################

### THE FACTOR MODEL ###

data$agef <- as.factor(data$age)

respd.glmf <- glm(respd ~ airpollution *(agef + gender + activity),
                  data = data,
                  family = binomial)
respd.glmf.1 <- glm(respd ~ airpollution *(agef + activity + gender),
                    data = data,
                    family = binomial)
respd.glmf.2 <- glm(respd ~ airpollution *(activity + agef + gender),
                    data = data,
                    family = binomial)
summary(respd.glmf)
anova(respd.glmf)
anova(respd.glmf.1)
anova(respd.glmf.2)

stargazer(respd.glmf,
          type = "latex",
          title = "Logistic Regression of Respiratory difficulty",
          dep.var.labels = "Respiratory difficulty",
          covariate.labels = c(
            "Air pollution (Yes)",
            "Age 35",
            "Age 45",
            "Age 55",
            "Male",
            "Active",
            "Air pollution    Age 35",
            "Air pollution    Age 45",
            "Air pollution    Age 55",
            "Air pollution    Male",
            "Air pollution    Active"
          ),
          digits = 2,
          star.cutoffs = c(0.05, 0.01, 0.001),
          no.space = TRUE,
          float = TRUE,
          font.size = "scriptsize",
          single.row = TRUE,
          out = "tables/resp.glmf.tex")

# in all cases in the factor model we see that the baseline age effect is not significant
# this provides further evidence for dropping
# activity is significant when added first, gender is always highly significant in the factor model


### extra plot - residuals vs fitted values

pdf("figs/residual_vs_mu.pdf")
plot(jitter(predict(chosen_model, type = 'response'), 10),
     jitter(rstandard(chosen_model), 0), col = blue_palette[2],
     xlab = expression(hat(mu)),
     ylab = "Standardised Deviance Residuals")
abline(a = -2, b = 0, lty = 2)
abline(a = 2, b = 0, lty = 2)
dev.off()


######### FOR THE REDUCED MODEL ##############

##### DIAGNOSTICS #######


# residual plots

set.seed(30)

pdf("figs/residual_plots_reducedmodel.pdf", width = 9.5, height = 10)
par(mfrow = c(2, 2), margin(5, 5, 3, 3), cex = 1)

```

```
plot(jitter(predict(reduced_model, type = ’link’), 10),
     jitter(rstandard(reduced_model), 0), col = blue_palette[2],
     xlab = expression(hat(eta)),
     ylab = "Standardised Deviance Residuals")
abline(a = -2, b = 0, lty = 2)
abline(a = 2, b = 0, lty = 2)

# qq plot

qqnorm(rstandard(reduced_model), main = "")
qqline(rstandard(reduced_model))

# leverage

p <- reduced_model$df.null - reduced_model$df.residual + 1
n <- nrow(data)

plot(jitter(influence(reduced_model)$hat/(p/n), 200), ylab = "Leverage / (p/n)",
     col = blue_palette[2])
abline(2, 0, lty = 2)

# cook’s distance

plot(jitter(cooks.distance(reduced_model), 100), ylab = "Cook’s Distance",
     col = blue_palette[2])
abline(8/(n - 2 * p), 0, lty = 2)
dev.off()

# ACTIVITY
# odds ratio depends on air pollution

beta <- coef(reduced_model)
V <- vcov(reduced_model)

se_sum <- function(b1_name, b2_name) {
  b_sum <- beta[b1_name] + beta[b2_name]
  var_sum <- V[b1_name, b1_name] + V[b2_name, b2_name] + 2 * V[b1_name, b2_name]
  return(se_sum <- sqrt(var_sum))
}

OR_activity_1 <- exp (beta[activity])
or_act_ap0 <- exp(reduced_model$coefficients[2])

CI_or_act_ap1 <- exp(summary(reduced_model)$coefficients[2,1] + summary(reduced_model)$coefficients[5,1]
                     + 1.96 * se_sum("activityYes", "activityxpollution") *
                       c(-1, 1))

CI_or_act_ap0 <- exp(summary(reduced_model)$coefficients[2,1]
                     + 1.96 * summary(reduced_model)$coefficients[2,2] *
                       c(-1, 1))

# Genderxairpollution
# odds ratio again depends on air pollution, but there is no term for airpo = 0

or_gender_ap1 <- exp(reduced_model$coefficients[4])
ci_or_gender_ap1 <- exp(summary(reduced_model)$coefficients[4,1]
                        + 1.96 * summary(reduced_model)$coefficients[4,2] *
                          c(-1, 1))

# age --> measured for 10 years
```

```
or_age_ap1 <- exp(reduced_model$coefficients[3] * 10)
ci_or_age_ap1 <- exp(summary(reduced_model)$coefficients[3,1] * 10
                     + 1.96 * 10 *summary(reduced_model)$coefficients[3,2] *
                       c(-1, 1))

# airpollution
# start with age = 25
act1_gender1 <- function(age) {
  or_ap_a1g1 <- exp(reduced_model$coefficients[3]*age +
                      reduced_model$coefficients[4] +
                      reduced_model$coefficients[5])

  se_sum_three <- sqrt(se_sum("genderxpollution", "activityxpollution") +
                         age ** 2 * V["agexpollution", "agexpollution"] +
                         2 * age * (V["agexpollution", "genderxpollution"] +
                                      V["agexpollution", "activityxpollution"])
  )

  CI_or_ap_a1g1 <- exp(reduced_model$coefficients[3]* age +
                         reduced_model$coefficients[4] +
                         reduced_model$coefficients[5] +
                         1.96 *
                         se_sum_three * c(-1, 1))
  output <- c(odds_ratio = or_ap_a1g1,
              conf_int = CI_or_ap_a1g1)
  return(output)
}

act1g1confints <- lapply(c(25,35,45,55), act1_gender1)

act1_gender0 <- function(age) {
  or_ap_a1g1 <- exp(reduced_model$coefficients[3]*age +
                      reduced_model$coefficients[5])

  se_sum_three <- sqrt(V["activityxpollution", "activityxpollution"]+
                         age ** 2 * V["agexpollution", "agexpollution"] +
                         2 * age * V["agexpollution", "activityxpollution"]
  )

  CI_or_ap_a1g1 <- exp(reduced_model$coefficients[3]* age +
                         reduced_model$coefficients[5] +
                         1.96 *
                         se_sum_three * c(-1, 1))
  output <- c(odds_ratio = or_ap_a1g1,
              conf_int = CI_or_ap_a1g1)
  return(output)
}

act1g0confints <- lapply(c(25,35,45,55), act1_gender0)

act0_gender1 <- function(age) {
  or_ap_a1g1 <- exp(reduced_model$coefficients[3]*age +
                      reduced_model$coefficients[4])

  se_sum_three <- sqrt(V["genderxpollution", "genderxpollution"]+
                         age ** 2 * V["agexpollution", "agexpollution"] +
                         2 * age * V["agexpollution", "genderxpollution"]
  )

  CI_or_ap_a1g1 <- exp(reduced_model$coefficients[3]* age +
                         reduced_model$coefficients[4] +
                         1.96 *
                         se_sum_three * c(-1, 1))
  output <- c(odds_ratio = or_ap_a1g1,
              conf_int = CI_or_ap_a1g1)
  return(output)
}

act0g1confints <- lapply(c(25,35,45,55), act0_gender1)

act0_gender0 <- function(age) {
  or_ap_a1g1 <- exp(reduced_model$coefficients[3]*age)

  se_sum_three <- sqrt(age ** 2 * V["agexpollution", "agexpollution"])

  CI_or_ap_a1g1 <- exp(reduced_model$coefficients[3]* age +
                         1.96 *
                         se_sum_three * c(-1, 1))
  output <- c(odds_ratio = or_ap_a1g1,
              conf_int = CI_or_ap_a1g1)
  return(output)
}

act0g0confints <- lapply(c(25,35,45,55), act0_gender0)
```

References

[1] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton, FL, 4 edition, 2018.

[2] Julian J. Faraway. *Extending the Linear Model with R*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition, 2016.