Lund University
Automatic Control
MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION
Spring 2024

Project 2
Logistic Regression

Yifei Zhang[1], Jialu Xu[2],

[1] yi4840zh-s@lu.se
[2] ji6606xu-s@lu.se

1 Introduction

The project is a focused the logistic regression models in predicting the likelihood of municipalities having a high number of cars per capita.

The report begins with an analysis of car ownership percentages across different parts of Sweden, demonstrating significant regional variations.

Further sections of the report detail the use of logistic regression to estimate the log-odds of high car numbers for specific regions and evaluate the model's accuracy through various statistical tests and metrics.

2 Part 1. Introduction to logistic regression

2.1 1(a). High number of cars without regression:

The percentage of high cars in different parts of sweden are presented in Table 1.

When we change from Gtaland to Norrlandthe Highcars Percentage times 4.

Table 1: Description of high percentages cars in different parts.

Part	High Percentage
1 Gtaland	11.4
2 Svealand	20.8
3 Norrland	40.7

2.2 1(b). High number of cars with regression:

Fit Model:

Fit a logistic regression model, Model.1(b), with Part as explanatory variable.The Results are shown in Table 2 and 3, correspond to -estimates, their standard errors, their 95 % profile likelihood confidence intervals,and e, their 95 % confidence intervals.

Identify the odds and odds ratios:

We have observed the impact of the Part variable on the highcars variable. The -estimate is 0.83, indicating that when the Part variable increases by one unit, the log-odds of the dependent variable increase by 0.83 units. In this scenario, $e^{0.83}$ represents the Odds Ratio. We find that $e^{0.83} = 2.3$, signifying that the odds of the dependent variable increase to 2.3 times when the Part variable increases by one unit.

By examining Table 1, we ascertain that when Part increases by one unit, the highcars variable closely increases to 2.3 times the original odds.

Estimation:

Use Model.1(b) to estimate the log-odds of high number of cars, for Gtaland, Svealand and Norrland. The Results are shown in table 4.

Table 2: Model 1b: Estimates and Confidence Intervals

Parameter	Estimate	2.5% CI	97.5% CI
(Intercept)	-2.93	-3.78	-2.15
Part	0.83	0.46	1.22

Table 3: Model 1b: Odds Ratios and Confidence Intervals

Parameter	Odds Ratio	2.5% CI	97.5% CI
(Intercept)	0.05	0.02	0.12
Part	2.30	1.58	3.40

Table 4: Logistic Regression Results by Part

Part	**logit.fit**	**logit.se.fit**	**logit.lwr**	**logit.upr**	**p**	**p.lwr**	**p.upr**
Gtaland	-2.09	0.244	-2.57	-1.62	0.11	0.071	0.166
Svealand	-1.26	0.152	-1.56	-0.959	0.221	0.174	0.277
Norrland	-0.423	0.249	-0.911	0.066	0.396	0.287	0.516

Test:

After conducting the Wald test on model 1b, the results are presented in Table 5. According to the criteria of $Pr < 0.05$ and $z > 1.96$, we find a significant relationship between 'part' and 'highcars'.

2.3 1(c). Access to a bus stop:

Plot

The plot of the 0/1 variable highcars against Transit are presented in Figure 1. It seem reasonable to use the proportion living close to a bus stop as an explanatory variable.

Fit Model

Fit a logistic regression model, Model.1(b), with Part as explanatory variable.The Results are shown in Table 6 and 7, correspond to β-estimates, their standard errors, their 95 % profile likelihood confidence intervals,and e^{β}, their 95 % confidence intervals.

Estimation Plot

Add the estimated probability of a high number of cars, and its confidence interval, to the plot, are presented in Figure 2.

Test:

After conducting the Wald test on model 1b, the results are presented in Table 8. According to the criteria of $Pr < 0.05$ and $z > 1.96$, we find a significant relationship between 'Transit' and 'highcars'.

According to the odds ratio, when the independent variable (Transit) increases by 1 percent, the dependent variable (Highcars) decreases by a factor of 0.93 compared to its original value.

Table 5: Wald Test

Parameter	**Estimate**	**Std. Error**	**z value**	$\Pr(>\|z\|)$
(Intercept)	-2.93	0.41	-7.08	1.44×10^{-12}
Part	0.83	0.19	4.30	1.67×10^{-5}

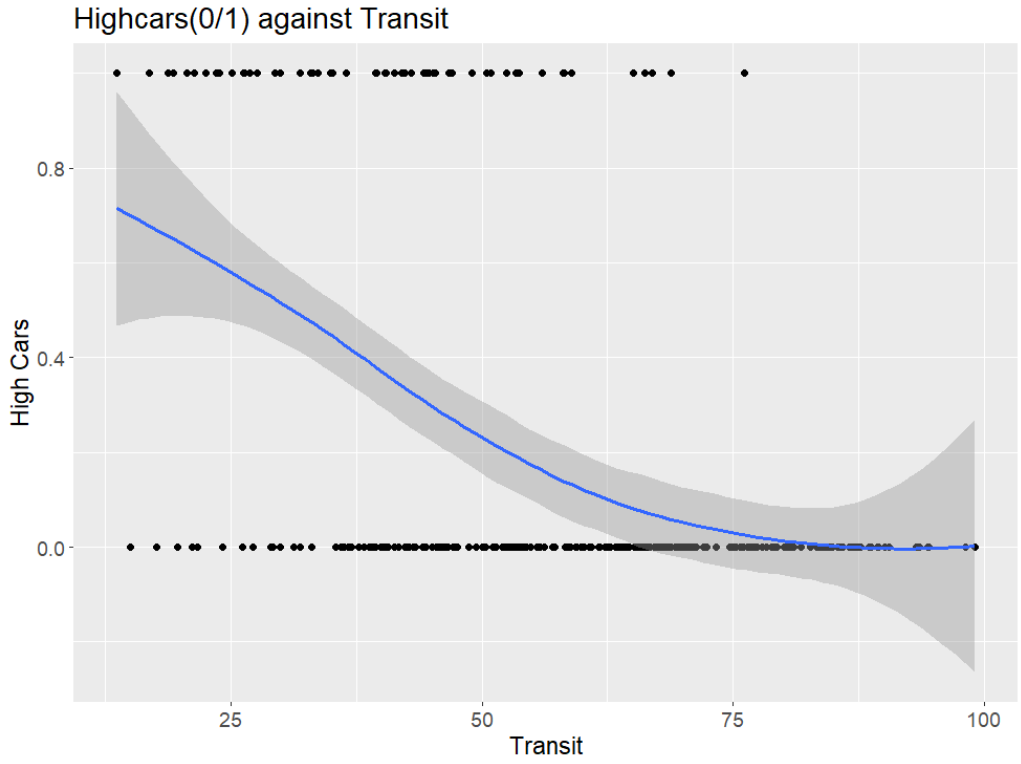

Figure 1: Highcars(0/1) against Transit.

1. Increase by 10 percentage units: Multiply the odds ratio (0.93) by itself ten times, or $(0.93)^{10} = 0.4840$.
2. Decrease by 1 percentage unit: Take the reciprocal of the odds ratio and multiply it once, or $\frac{1}{0.93} = 1.0752$.
3. Decrease by 10 percentage units: Take the reciprocal of the odds ratio and multiply it by itself ten times, or $\frac{1}{0.93}^{10} = 2.0662$.

2.4 1(d). Leverage:

Leverage Plot:

Calculate the leverage values for Model.1(c) and plot them against Transit. Add horizontal reference lines at the minimal value 1/n and at 2(p + 1)/n and make sure the y-axis includes zero. The plot is shown in Figure 3

Table 6: Model 1c: Estimates and Confidence Intervals

Variable	**Estimate**	**95% CI Lower**	**95% CI Upper**
(Intercept)	2.38	1.39	3.46
Transit	-0.07	-0.10	-0.05

Table 7: Model 1c: Odds Ratios and Confidence Intervals

Variable	Exp(Beta)	95% CI Lower	95% CI Upper
(Intercept)	10.83	4.01	31.67
Transit	0.93	0.91	0.95

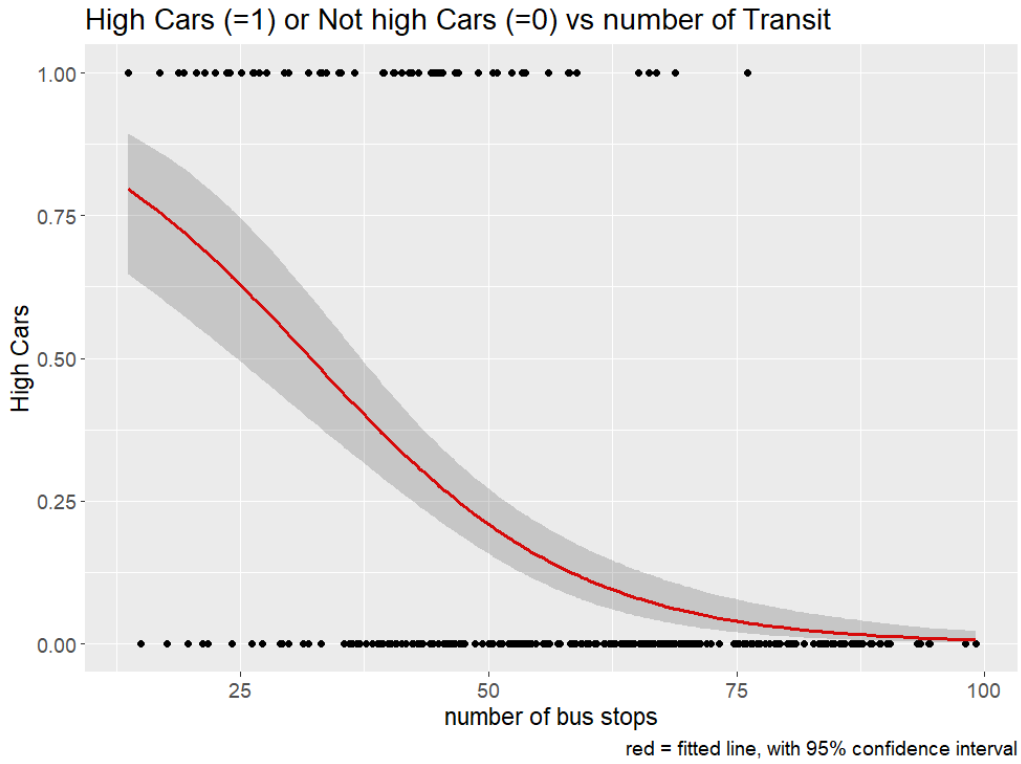

Figure 2: Highcars(0/1) against Transit.

2.5 1(e)

Calculate McFaddens adjusted pseudo R^2, AIC and BIC for Model.1(b) and Model.1(c). The results are shown in Table 9

1. McFadden's adjusted pseudo R^2 (R^2McF.adj): This metric is used to measure the goodness of fit of the model. A value closer to 1 indicates a better explanatory power of the model for the observed data.
2. AIC (Akaike Information Criterion): AIC is a comprehensive metric that combines the goodness of fit and complexity of the model. A smaller AIC value indicates that the model uses less information in fitting the data, suggesting a better model.
3. BIC (Bayesian Information Criterion): BIC is similar to AIC but is more effective for small sample sizes. Like AIC, a smaller BIC value indicates a better quality of the model.
4. Log-likelihood: Log-likelihood is typically used to assess the goodness of fit of statistical models, where a larger value indicates a better ability of the model to fit the data. By comparing the log-

Table 8: Wald test results for model 1c

Variable	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	2.38	0.52	4.54	5.58×10^{-6}
Transit	-0.07	0.01	-6.80	1.07×10^{-11}

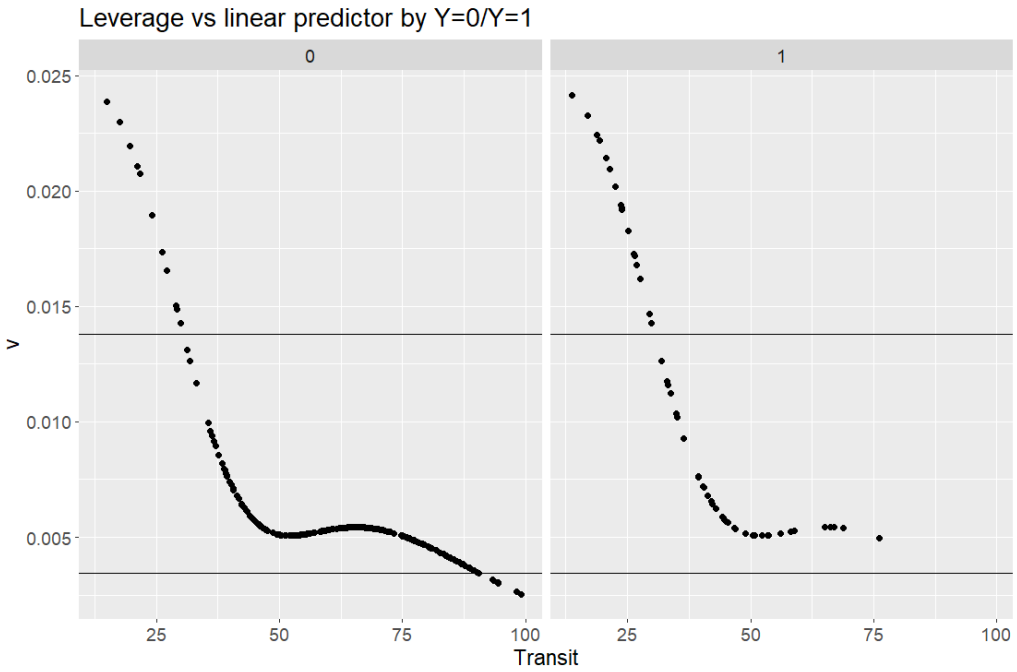

Figure 3: Leverage Plot Against Transit.

likelihood differences between two models, we can determine which model has a higher log-likelihood. In general, a model with a larger log-likelihood difference is considered better.

Considering these metrics collectively, we can judge which model is better by comparing their values. Typically, the model with higher McFadden's adjusted pseudo R^2, log-likelihood and smaller AIC, BIC values is preferred. By observe the results in table 9, we can observe that the AIC and BIC of Model 1(c) are smaller than those of Model 1(b), while the R^2 McF.adj and log-likelihood are higher. This suggests that Model 1(c) is superior to Model 1(b), indicating that access to a bus stop seems more important.

Table 9: Comparison of Model 1(b) and Model 1(c)

Model	**df**	**AIC**	**BIC**	**Log Likelihood**	**R^2McF**	**R^2McF.adj**
Model 1(b)	2	274.9130	282.2527	-135.4565	0.06656858	0.06312308
Model 1(c)	2	226.4977	233.8375	-111.2489	0.23338343	0.22993793

3 Part 2. Variable selection and influential observations

3.1 2(a). High number of cars without regression:

The row numbers for the municipalities that miss fertility rates are 263, 270, 273, 278 and 280. Ignoring these rows, we get a mean of **1.556** for the remaining rows.

3.2 2(b). Variable selection:

Start by fitting a full logistic regression model (Model.full) with all 11 continuous explanatory variables. The model we get (Model.full) is shown in Table 10.
And the VIF value for each cofficients is shown in Table 11. We note that the VIF values for both **Children** and **Persperhh** are greater than 5, and the VIF value for **Seniors** is nearly 5, both with possible multicollinearity problems.

After the stepwise selection, we get Model.AIC using AIC as criterion and Model.BIC using BIC as criterion. Model.AIC is using variables **Urban + log(Apartments) + Persperhh + log(Income) + Fertility + Transit**, while Model.BIC is using variables **Urban + log(Apartments) + Persperhh + log(Income)**. As the two models are nested, the VIF values are shown together in Table 12.
Since the VIF values for all varibales are less than 5, we don't think there is any worrying multicollinearity problem now.
In order to test the validity of the increased parameters of the AIC model with respect to the BIC model, we will perform the LR test, and we want to test $H_0 : \beta_{Fertility} = \beta_{Transit} = 0$. We can calculate **D_diff, df_diff, chi2_alpha and Pvalue** using **deviance** and **df.residual**. The result is shown in Table 13. The P-value is 0.049, which is smaller than 0.5, we should reject H_0. Faddens adjusted R_2, AIC and BIC are reported in Table 14. Combining various evaluation indexes, we believe that the Model.BIC achieves similar fitting effect and generalization ability (refer to Faddens adjusted R_2) as the Model.AIC, but uses fewer parameters and a simpler model, so we choose BIC as Model.2b.

Table 10: Summary of Model.full

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-33.34735	33.43096	-0.997	0.318522
log(Higheds)	-0.47361	1.21531	-0.390	0.696758
Children	-0.07683	0.31580	-0.243	0.807778
Seniors	-0.10683	0.13445	-0.795	0.426852
log(Income)	16.08558	6.48962	2.479	0.013188
log(GRP)	0.51651	0.75883	0.681	0.496088
Persperhh	-14.97101	4.59537	-3.258	0.001123
Fertility	-2.42583	1.51074	-1.606	0.108334
Urban	-0.04902	0.02651	-1.849	0.064477
Transit	-0.03023	0.01879	-1.609	0.107679
log(Apartments)	-4.68215	1.26623	-3.698	0.000218

Table 11: VIF value for Model.full

Parameter	VIF
log(Higheds)	1.992015
Children	6.336823
Seniors	5.070938
log(Income)	2.792095
log(GRP)	1.613995
Persperhh	5.659235
Fertility	1.576855
Urban	2.232396
Transit	1.848484
log(Apartments)	3.056299

Table 12: VIF value for Model.AIC and Model.BIC

Parameter	VIF of Model.AIC	VIF of Model.BIC
Urban	2.156499	1.735494
log(Apartments)	2.076635	1.933552
Persperhh	2.294965	2.186122
log(Income)	2.059561	2.007957
Fertility	1.278912	N/A
Transit	1.649123	N/A

Table 13: Result of Likelihood-ratio test

D_diff	df_diff	chi2_alpha	Pvalue
6.050757	2	5.991465	0.04853945

Table 14: AIC, BIC and Fadden's adjusted R_2 for both model

Model	AIC	BIC	Fadden's adjusted R_2
model_aic	170.4776	196.1668	0.4401829
model_bic	172.5284	190.8778	0.4262260

3.3 2(c). Influential observations:

We choose

$$2 * p/n$$

as horizontal line, as p is the number of independent variables in the model (including the intercept term), and n is the number of observations.
We selected municipalities that were above the threshold and ranked in the top 8 as having worryingly high leverage. The plot is shown in Figure 4.
These municipalities are:

- Hgsby
- Mrbylnga
- Tjrn
- rjng
- Lekeberg
- Gagnef
- Bjurholm
- Gllivare

We also calculate the cook's distance. Since none of the Cooks distance values are even close to 1, we only add the horizontal line at $4/n$ to the plot. The plot is shown in Figure 5. The top 8 municipalities with high Cook's distance are:

- Kiruna
- Gagnef
- Ydre
- Tjrn
- Orsa
- Jokkmokk
- Gllivare
- Mora

We extracted DFBETAS values for observations with high Cook's distances and determined for which parameters the estimates were affected. The result is shown in Table 15, while **True** means affected, **Flase** means not affected.

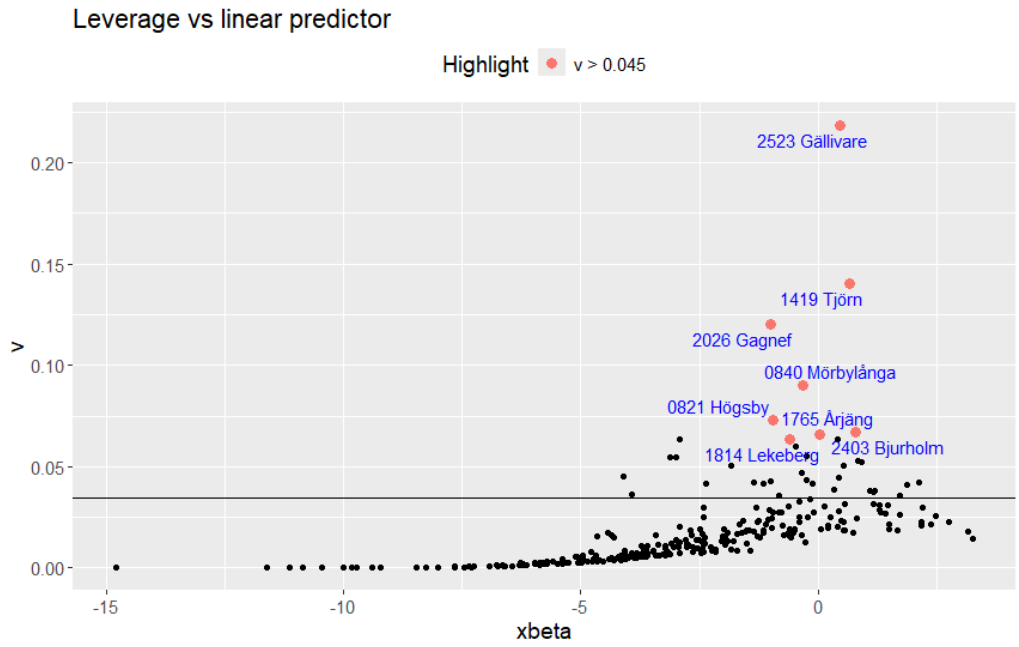

Figure 4: Leverage for Model.2(b) against the linear predictor

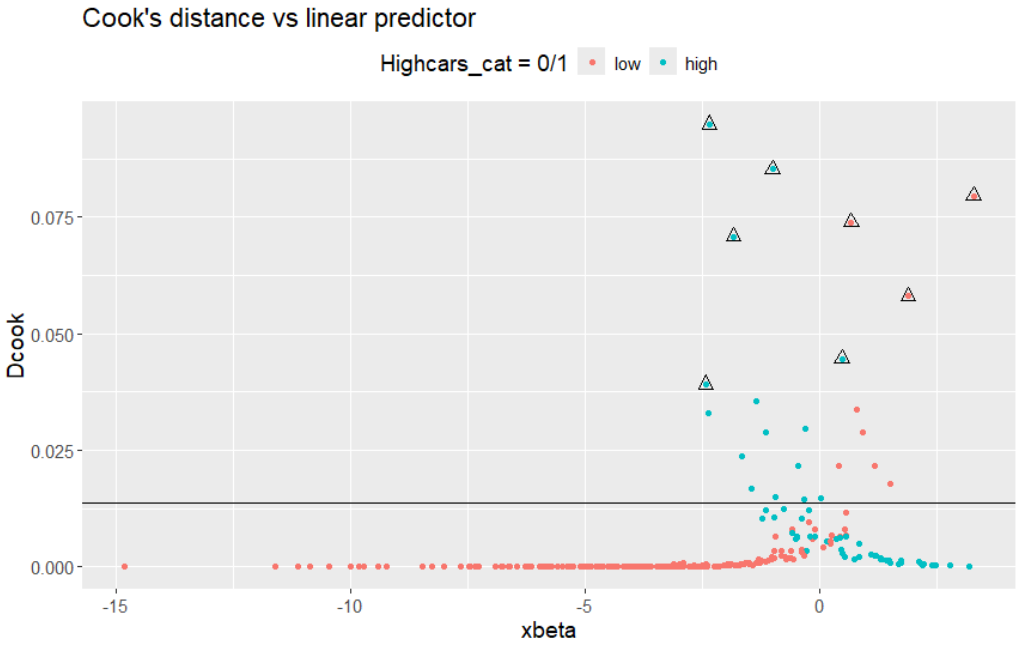

Figure 5: Cook's distance for Model.2(b) against the linear predictor

Table 15: Parameters affected by the observations with the high Cooks D

Municipalities	(Intercept)	Urban	log(Apartments)	Persperhh	log(Income)
Ydre	FALSE	TRUE	TRUE	TRUE	TRUE
Tjrn	TRUE	TRUE	TRUE	TRUE	TRUE
Gagnef	TRUE	TRUE	TRUE	TRUE	TRUE
Orsa	TRUE	TRUE	TRUE	FALSE	TRUE
Mora	TRUE	TRUE	FALSE	FALSE	TRUE
Jokkmokk	TRUE	TRUE	TRUE	TRUE	TRUE
Gllivare	TRUE	TRUE	FALSE	TRUE	TRUE
Kiruna	TRUE	FALSE	TRUE	FALSE	TRUE

3.4 2(d). Deviance residuals:

We examined the standardized deviance residuals for Model 2(b) against the linear predictor, incorporating color coding to distinguish between municipalities with a low or high number of cars (blue/orange). Additionally, we added suitable reference lines to aid visualization. Observations with high Cooks D values, as identified previously, were emphasized(red). Furthermore, we identified observations with large deviance residuals, denoted by $d_i > 3$(green). The plot is shown in Figure 6.
We also conducted a similar analysis by plotting the standardized deviance residuals against each of the predictor variables in the model, including suitable reference lines. The plot is shown in Figure 7.
These graphs show that the standardized deviation residuals are relatively homogeneous across most observations and have no significant systematic bias. The residuals for most of the data points are within plus or minus two standard deviations, indicating that the model is fitting appropriately for most of the data. However, the extreme residual values observed need to be further analyzed to determine if they may be indicative of model assumption violations, data anomalies, or other potential data quality issues.

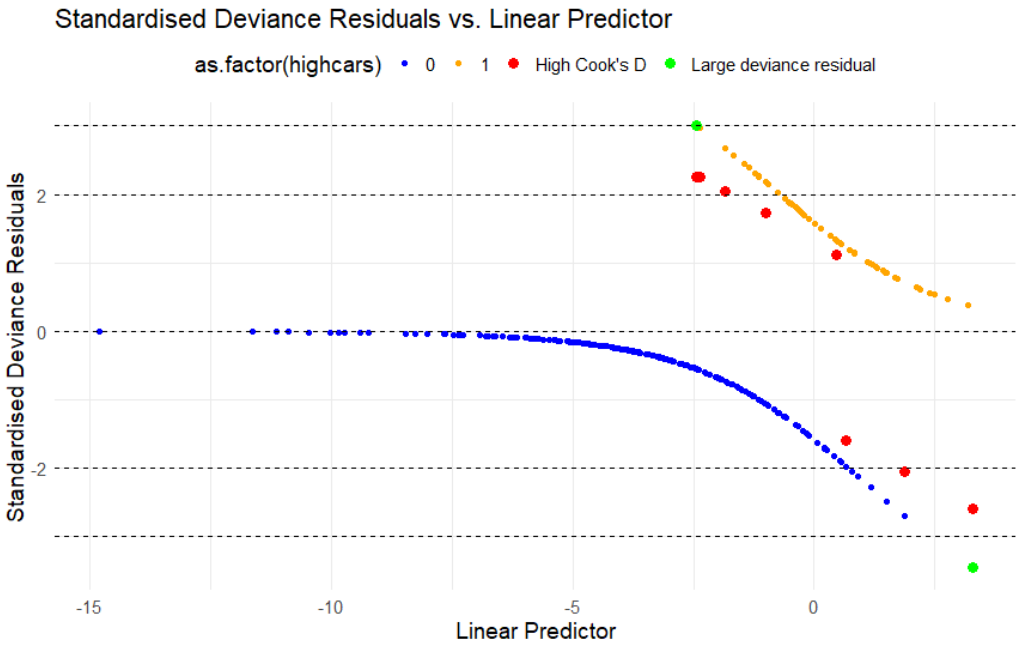

Figure 6: Standardised deviance residuals for Model.2(b) against the linear predictor

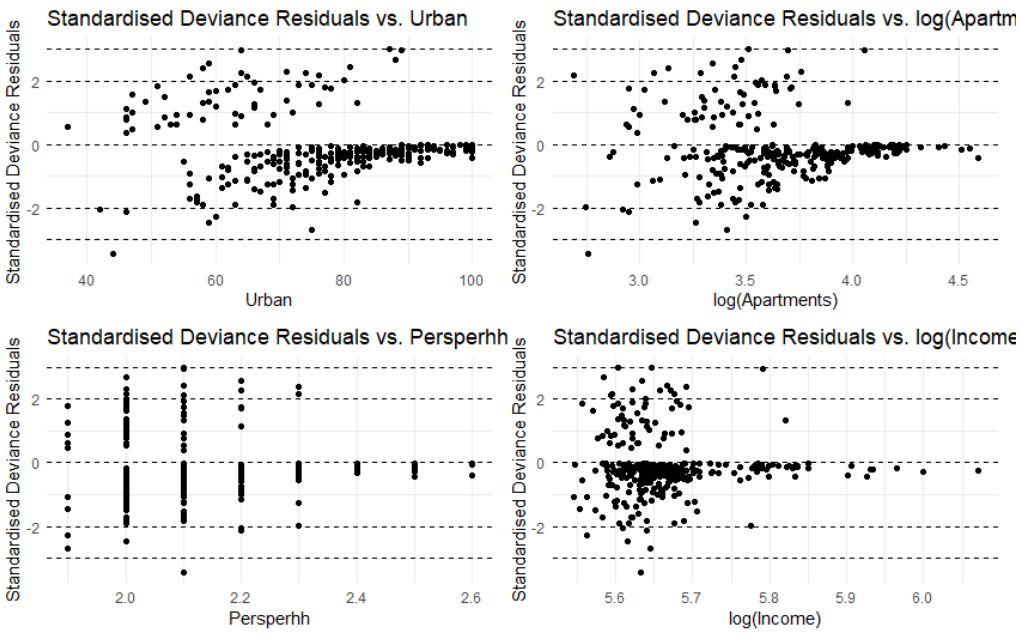

Figure 7: Standardised deviance residuals against each of the variables

4 Part 3. Goodness-of-fit

4.1 3(a). Confusion:

Present the resulting confusion matrices as well as a table, as Table 16, collecting the Accucracy, the P-value for Acc > NIR, Cohens , the P-value for McNemars test, Sensitivity and Specificity for all five models.

Table 16: Table 3(a) Confusion Matrix Comparative of Models

Model	Accuracy	Acc > NIR	Kappa	McNemar's P-Value	Sensitivity	Specificity
cm_null	0.8	0.5351	0	7.184×10^{-14}	0.0	1.0
cm_1b	0.8	0.5351	0	7.184×10^{-14}	0.0	1.0
cm_1c	0.8172	0.2571313	0.3081	0.0003551	0.31034	0.94397
cm_2b	0.8517	0.01432	0.508	0.22247	0.5517	0.9267
cm_aic	0.8724	0.000805	0.5767	0.188445	0.6034	0.9397
cm_full	0.8862	0.00006664	0.6224	0.1637	0.6379	0.9483

1. **Accuracy**: Accuracy measures the proportion of total true results (both true positives and true negatives) in the total population. It is a straightforward metric used to assess the overall effectiveness of a model. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}} \tag{1}$$

2. **P-value for Accuracy > No Information Rate (NIR)**: The P-value for accuracy greater than the no information rate tests the null hypothesis that the true accuracy is equal to the no information rate, against the alternative that it is greater.

   The no information rate is the accuracy that could be achieved by predicting the most frequent class. A low P-value (typically < 0.05) indicates strong evidence against the null hypothesis, thus the model's accuracy is statistically significantly better than the NIR.

3. **Cohens κ**: Cohen's κ is a statistic that measures inter-rater reliability for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as takes into account the agreement occurring by chance. Cohen's κ is defined as equation 2.

$$\kappa = \frac{Accuracy - P_c}{1 - P_c} \tag{2}$$

4. **P-value for McNemars Test**: McNemar's test is used to determine whether there are differences on paired proportions. This test is applied to a 2x2 contingency table with a binary classification task performed twice on the same samples. A significant P-value suggests that the performance of the two tests is statistically significantly different. The formula for the test statistic is shown as equation 3.

$$\chi^2 = \frac{(FP - FN)^2}{FP + FN} \tag{3}$$

5. **Sensitivity and Specificity**:Sensitivity (also called the true positive rate or recall) measures the proportion of actual positives that are correctly identified as such. Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified. These are crucial metrics for diagnostic tests and are defined as equation 4,where TP means True Positives, FN means False Negatives, TN means True Negatives, FP means False Positives.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

4.2 3(b). ROC-curves and AUC:

ROC Curve (Receiver Operating Characteristic Curve)

The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the True Positive Rate (TPR), also known as sensitivity or recall, against the False Positive Rate (FPR), also known as 1-specificity, at various threshold settings. The TPR is plotted on the y-axis and the FPR is on the x-axis.

The ROC curve provides a comprehensive tool to assess the trade-off between sensitivity and specificity. It is particularly useful for evaluating the performance of diagnostic tests and predictive models, especially in scenarios where the classes are imbalanced. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

AUC (Area Under the ROC Curve)

The AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$. It provides a single scalar value that summarizes the performance of the classifier across all possible thresholds. The AUC value lies between 0 and 1.

- AUC = 1: This is an ideal case where the classifier is able to perfectly distinguish between all the positive and the negative classes.
- AUC = 0.5: This scenario represents a model with no discrimination capacity, effectively random guessing. An AUC of 0.5 suggests that the model fails to distinguish between the classes.
- AUC $<$ 0.5: This indicates a model with worse-than-random predictions, although this situation is less common and typically suggests some error in model setup or data processing.

Plot

Plot the ROC-curves for all six models in the same plot as figure 8, and present a table with their AUC-values, including 95 % confidence intervals, shown as table 17.

Assessing Model Performance Using AUC

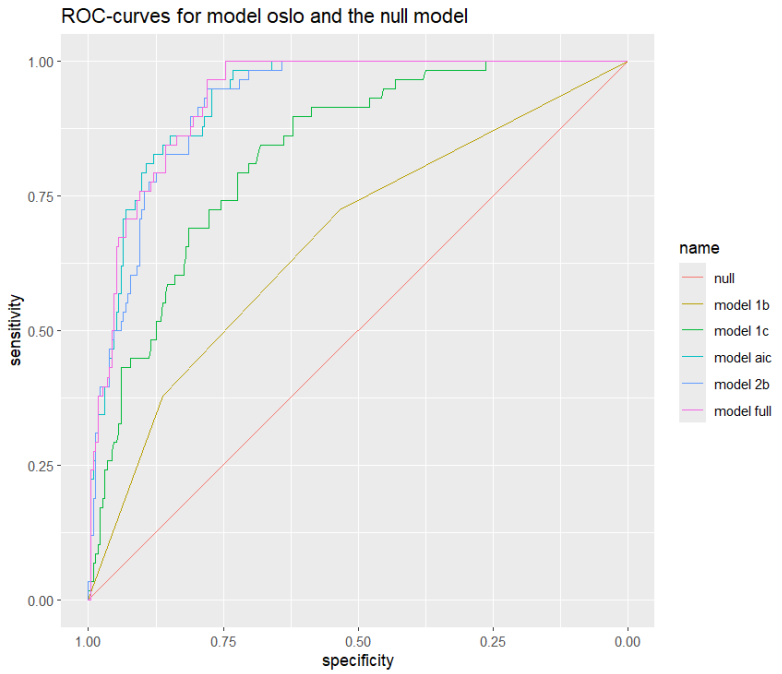

Figure 8: ROC-curves

Table 17: AUC Scores and Confidence Intervals for Models

Model	AUC	Lower CI	Upper CI
null	0.5000	0.5000	0.5000
1b	0.6677	0.5921	0.7432
1c	0.8274	0.7736	0.8812
aic	0.9283	0.8987	0.9578
2b	0.9203	0.8893	0.9514
full	0.9318	0.9037	0.9598

The AUC value serves as a comprehensive measure of the model's ability to correctly classify the positive and negative cases across all possible thresholds, and it does not depend on any specific threshold selection.

- Excellent Performance: AUC values between 0.9 and 1.0 are considered excellent, often indicating that the model has a high measure of separability between the positive and negative classes.
- Good Performance: AUC values between 0.8 and 0.9 suggest good discriminatory ability.
- Fair Performance: AUC values between 0.7 and 0.8 are deemed fair, showing some discriminatory ability but not strong.
- Poor Performance: AUC values between 0.6 and 0.7 are considered poor and often require model re-evaluation or enhancement.
- Fail/Random Guess: AUC values below 0.6 are troubling, indicating little to no discriminative power, often comparable to random guessing.

Upon examining Table 17, we observe that the AUC value of model2b exceeds 0.9, which is close to the AUC values of the aic and full models. This suggests that model2b is a high-performing model. Further analysis of other performance metrics corroborates the conclusion that find in 2(b).

4.3 3(c). Optimal thresholds:

For each of the five models (not the null model), find the optimal threshold for p, where the distance to the ideal model is minimized. Use these new thresholds to calculate new confusion matrices and a new version of Table 16, with the optimal thresholds added, shown as Table 18.

After comparing table 16 and table 18, an interesting observation is that, although the accuracy decreases after adjusting the thresholds, the sensitivity increases, with model2b having the highest sensitivity. This demonstrates that model2b is the best-performing model in terms of sensitivity.

Table 18: Confusion Matrix Comparative of Models after threshold

Model	Accuracy	Acc > NIR	Kappa	McNemar's P-Value	Sensitivity	Specificity
cm_thre_1b	0.5724	1	0.1622	3.032×10^{-16}	0.7241	0.5345
cm_thre_1c	0.7379	0.9958	0.3871	4.913×10^{-09}	0.7931	0.7241
cm_thre_2b	0.8276	0.1346	0.5675	1.672×10^{-07}	0.8966	0.8103
cm_thre_aic	0.8517	0.01432	0.6055	7.341×10^{-05}	0.8621	0.8491
cm_thre_full	0.8552	0.0094649	0.6082	0.0003867	0.8448	0.8578

4.4 3(d). Conclusion

Based on a comprehensive review of all results, we have selected model2b (bic results) as the optimal model for several compelling reasons:

1. **Minimal Parameter Count:** Model_2b has the fewest parameters, which often signifies stronger generalization capabilities. This indicates a more streamlined model that avoids overfitting while maintaining robust predictive performance.
2. **Performance in the Confusion Matrix:** The accuracy rates of models aic, full, and 2b are comparable, each achieving over 80%.
3. **High Sensitivity:** Model_2b surpasses the other models in sensitivity, demonstrating its superior ability to accurately classify positive instances. This trait is particularly valuable when the cost of false negatives is high, thereby making model_2b the most reliable choice among the compared models.

These factors collectively confirm model2b as the best choice among the considered models, striking an optimal balance between complexity, performance, and the ability to accurately identify positive cases. The -estimates and their corresponding 95 % confidence intervals of Model.2(b) are shown in Table 19.

Comments on the variables and the possible reasons:

- **Urban**

  Urban areas with a high percentage of the population residing in city regions typically indicate higher population densities and well-developed public transportation systems. Therefore, there may be a negative correlation between urban residency and cars number.

- **Apartments**

  Generally, a higher number of apartments in an area indicates higher population density and possibly lower income levels. In urban areas with efficient public transportation, residents may rely less on private vehicles. Thus, there might be a negative correlation between the number of apartments and the cars number.

- **Persons per Household (Persperhh)**

  A higher number of persons per household usually means more people sharing vehicles. This suggests that there could be a negative correlation between the number of persons per household and cars number.

- **Income**

  Typically, higher income levels enable more families to afford and actually possess more vehicles. Wealthier individuals and households tend to live in areas where cars are more necessary and feasible. This suggests that there might be a positive correlation between income and cars number.

Table 19: -estimates and their corresponding 95% confidence intervals in Model.2(b)

Variable	Estimate	2.5%	97.5%
(Intercept)	-43.23069	-93.8371	9.2722
Urban	-0.06686	-0.1142	-0.0231
log(Apartments)	-4.04116	-6.0526	-2.1861
Persperhh	-15.62964	-21.5160	-10.5287
log(Income)	16.58025	6.2595	27.2246

5 Team Roles

Yifei Zhang is responsible for problem analysis, code writing, and drafting report part 1 & 3.

Jialu Xu is responsible for problem analysis, code writing and review, and drafting report parts 2.

6 AI Usage Statement

In this project, AI is utilized solely for the following purposes:

1. Translation from Chinese to English.
2. Generating tables.
3. Consultation and resolution of certain code-related issues.