Applied Regression Analysis: Linear and General Linear Models

Contents

1 Anorexia treatment. (Adapted from Agresti Ex. 1.24)

For 72 young girls suffering from anorexia, the `Anorexia.dat` file contains their weights before and after an experimental period (Table 1).

Table 1: The first five rows of the anorexia data.

subj	therapy	before	after
1	b	80.5	82.2
2	b	84.9	85.6
3	b	81.5	81.4
4	b	82.6	81.9
5	b	79.9	76.4

The girls were randomly assigned to receive one of three therapies during this period. A control group (c) received the standard therapy, which was compared to family therapy (f) and cognitive behavioral therapy (b). The goal of the study is to compare the effectiveness of the therapies in increasing the girls' weights.

(a) Prepare the data by (1) removing the `subj` variable, (2) re-coding the factor levels of `therapy` as `behavioral`, `family`, and `control`, (3) renaming `before` and `after` to `weight_before` and `weight_after`, respectively, and (4) adding a variable called `weight_gain` defined as the difference of `weight_after` and `weight_before`. Print the resulting tibble.

(b) Explore the data by (1) making box plots of `weight_gain` as a function of `therapy`, (2) making a scatter plot of `weight_gain` against `weight_before`, coloring points based on `therapy` and (3) creating a table displaying, for each `therapy` group, the mean weight gain, maximum weight gain, and fraction of girls who gained weight (i.e. `weight_gain > 0`). Based on these summaries: What therapy appears overall the most successful and why? How effective does the standard therapy appear to be? What is the greatest weight gain observed in this study? Which girls tended to gain most weight (in the absolute sense), based on their weight before therapy? Why might this be the case?

(c) Run a linear regression of `weight_gain` on `therapy` and print the regression summary (print in `R`, without using `kable`). Identify the base category chosen by `R` and discuss the interpretations of the fitted coefficients. Choosing `control` as the base category makes more sense. Recode the factor levels so that `control` is the first (and therefore will be chosen as the base category), rerun the linear regression, and print the summary again.

(d) Directly compute the between-groups, within-groups, and corrected total sums of squares (without appealing to the `aov` function or equivalent) and verify that the first two add up to the third. What is the ratio of the between-groups sum of squares and the corrected total sum of squares? What is the interpretation of this quantity, and what quantity in the regression summaries printed in part (c) is it equivalent to?

1.1 Anorexia treatment Data Analysis

(a) Tibble is printed in the code file.

(b) From Table 2 it can be seen that "family" therapy has highest mean weight gain among other therapies also, the maximum weight gained is highest with this therapy, further the fraction of girls with positive weight gain is highest for this therapy.
From the boxplot in fig 1 it seems that there are quite a few outliers in the behavioral therapy, this is quite useful to know because without this information one might be biased towards selecting behavioral therapy as a successful one.
From the scatter plot in fig 2 it seems that the girls with weight < 80 tend to gain the most weight, but there seems to little correlation between the therapy and weight gain for this argument.

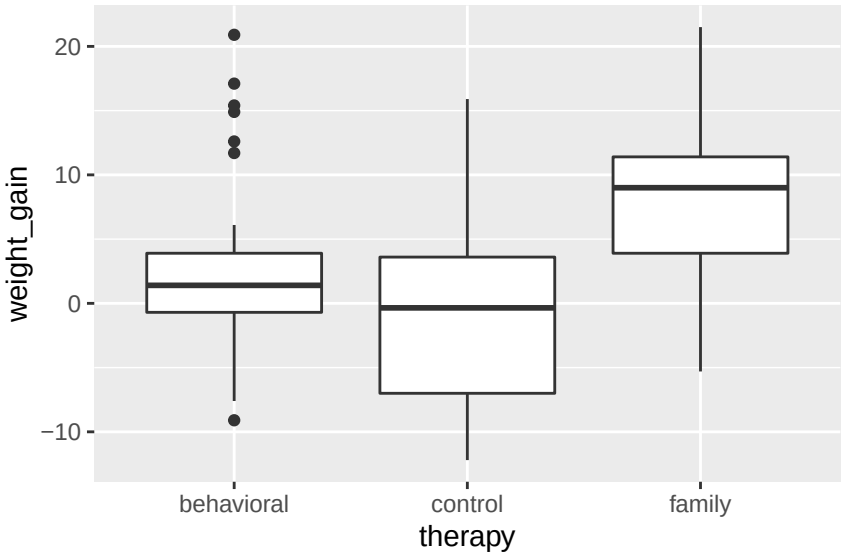

Figure 1: Box plot of weight gain as a function of therapy

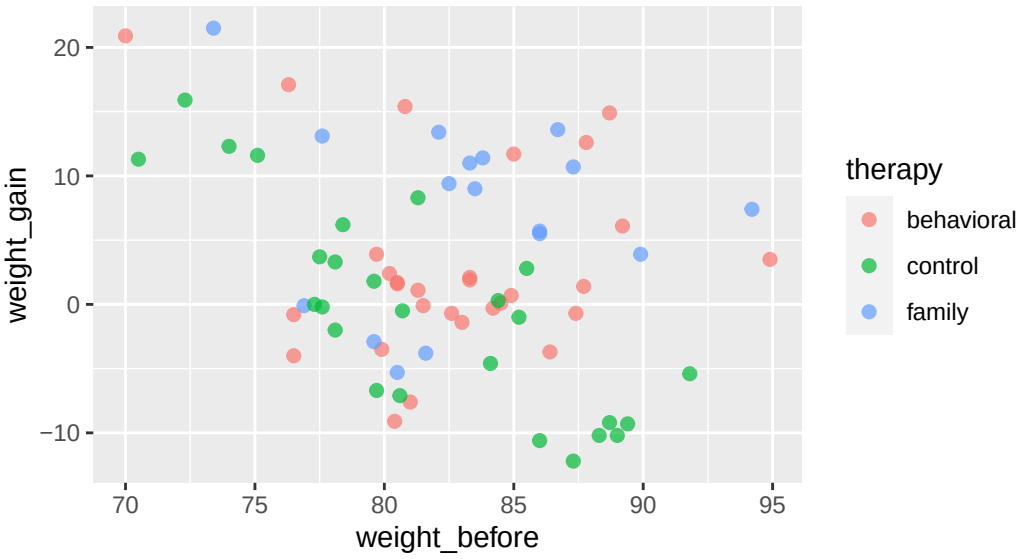

Figure 2: Scatter plot of weight gain vs weight before.

Table 2: Weight gain summary by group

Therapy	Mean Weight Gain	Max Weight Gain	Fraction (Weight Gain > 0)
behavioral	3.01	20.9	0.62
control	-0.45	15.9	0.42
family	7.26	21.5	0.76

(c) From the regression summary, it can be seen that behavioral therapy is chosen as base category, for this case the coefficient of control therapy indicates the difference between the predicted value of weight gain for control therapy and behavioral therapy, similarly the coefficient of family therapy indicates the difference between the predicted value of weight gain for family therapy and behavioral therapy. Expected Values upon change of base category:

$$\widehat{\beta}_0' = \widehat{\beta}_0 + \widehat{\beta}_1$$

$$\widehat{\beta}_1' = -\widehat{\beta}_1$$

$$\widehat{\beta}_2' = -\widehat{\beta}_1 + \widehat{\beta}_2$$

In this problem, we have: ($\widehat{\beta}_0'$, $\widehat{\beta}_1'$, $\widehat{\beta}_2'$) = (-0.450, 3.457, 7.715) ($\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$) = (3.007, -3.457, 4.258), clearly the above relationships hold for these set of parameters.

(d) Calculations for this part are performed using the R script

$$\text{SSR} = \sum_{i=1}^{n} (\widehat{\mu}_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\bar{y}_{c(i)} - \bar{y}\right)^2 = 614.64$$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}_{c(i)}\right)^2 = 3910.74$$

$$\text{SSR} + \text{SSE} = 4525.39$$

$$\text{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 = 4525.39$$

$$\frac{SSR}{SST} = 0.136$$

The above ratio corresponds to Multiple R-squared from linear regression model.

2 Violent crime.

The `Statewide_crime.tsv` file contains information on the number of violent crimes and murders for each U.S. state in a given year, as well as three socioeconomic indicators: percent living in metropolitan areas, high school graduation rate, and poverty rate (Table 3).

Table 3: The first five rows of the crime data.

STATE	Violent	Murder	Metro	HighSchool	Poverty
AK	593	6	65.6	90.2	8.0
AL	430	7	55.4	82.4	13.7
AR	456	6	52.5	79.2	12.1
AZ	513	8	88.2	84.4	11.9
CA	579	7	94.4	81.3	10.5

The goal of this problem is to study the relationship between the three socioeconomic indicators and the per capita violent crime rate.

(a) These data contain the total number of violent crimes per state, but it is more meaningful to model violent crime rate per capita. To this end, go online to find a table of current populations for each state. Augment `crime_data` with a new variable called `Pop` with this population information (see `left_join()` from the `dplyr` package) and create a new variable called `CrimeRate` defined as `CrimeRate = Violent/Pop` (see `mutate()` from the `dplyr` package).

(b) Explore the variation and covariation among the variables `CrimeRate`, `Metro`, `HighSchool`, `Poverty` with the help of visualizations and summary statistics.

(c) Construct linear model based hypothesis tests and confidence intervals associated with the relationship between `CrimeRate` and the three socioeconomic variables, including any relevant tables or plots in your LaTeX report. Discuss the results in technical terms.

(d) Discuss your interpretation of the results from part (c) in language that a policymaker could comprehend, including any caveats or limitations of the analysis. Comment on what other data you might want to gather for a more sophisticated analysis of violent crime.

2.1 Violent Crime Data Analysis

(a) Calculations performed in the R-script, the population data was taken from https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html

(b) Table 4 covariance matrix, from the table it can be seen that there is very insignificant variance in crime-rate among different states, but a significant variance in the percent of people living in the metropolitan areas, whereas there is small variance in high school graduation rate, and poverty rate. From the correlation heatmap (3), it can be seen that, there is no correlation between high school graduation rate and percent of people living in the metropolitan areas as well as crime rate, crime rate seems to be positively correlated with percent of people living in the metropolitan areas and poverty rate, also poverty seems to be negatively correlated with high school graduation rate and percent of people living in the metropolitan areas.

(c) Table 5 shows the regression output. It can be seen that F-statistic is significant and all the t-statistics are also significant based on the p-values. Fig 4 shows the confidence intervals of

Table 4: Covariance matrix

	CrimeRate	Metro	HighSchool	Poverty
CrimeRate	0.000	0.001	0.000	0.000
Metro	0.001	233.353	0.124	-4.888
HighSchool	0.000	0.124	13.076	-7.909
Poverty	0.000	-4.888	-7.909	9.794

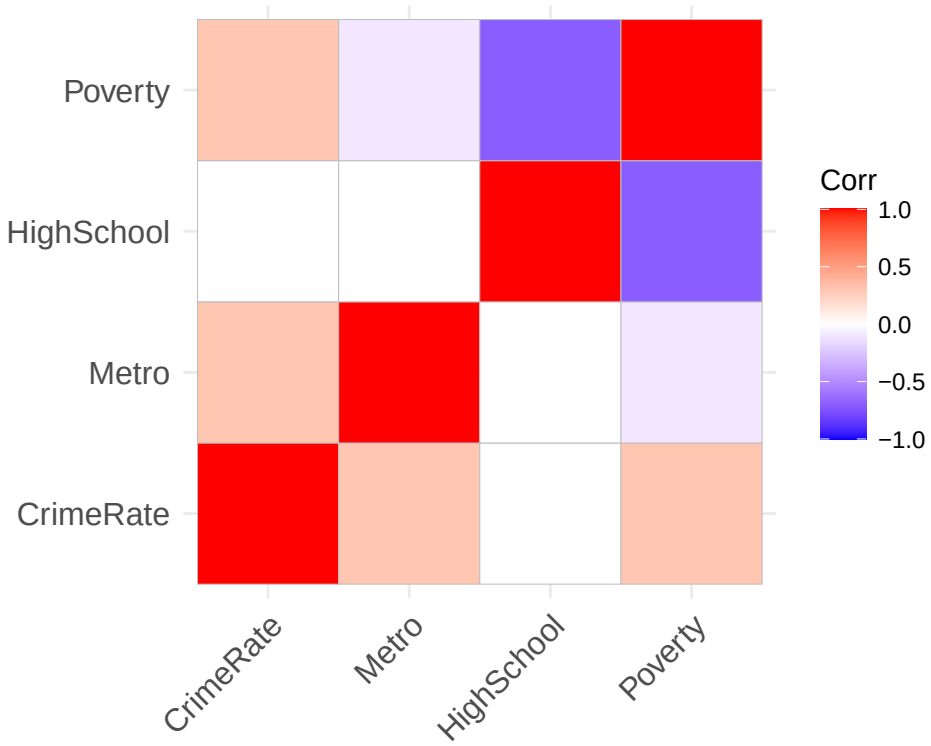

Figure 3: Correlation Plot

the coefficients of linear model. It seems that coefficient for percentage of population living in metropolitan areas has least variance and the other two coefficients have a bigger confidence interval.

Table 5: Results of regressing CrimeRate on Metro, HighSchool and Poverty

	Dependent variable:
	CrimeRate
Metro	0.000008^{***}
	(0.000003)
HighSchool	0.000051^{***}
	(0.000016)
Poverty	0.000083^{***}
	(0.000019)
Constant	-0.005613^{***}
	(0.001558)
Observations	51
R^2	0.348037
Adjusted R^2	0.306422
Residual Std. Error	0.000293 (df = 47)
F Statistic	8.363326^{***} (df = 3; 47)
Note:	*p<0.1; **p<0.05; ***p<0.01

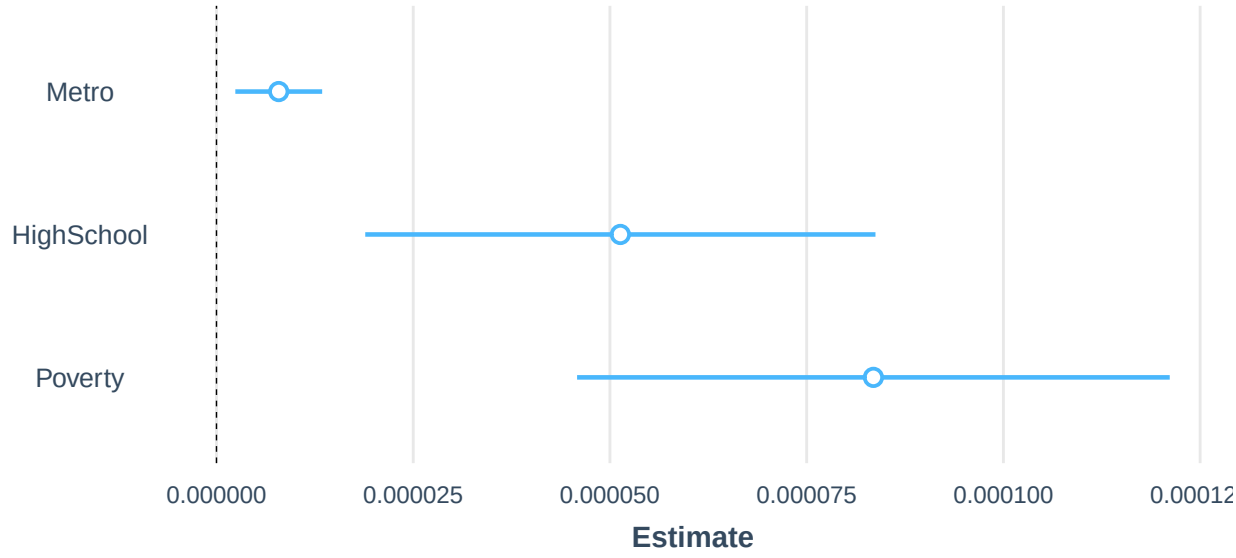

Figure 4: Confidence-Interval Plot for Coefficients

(d) From our analysis of data it seems that all the socioeconomic variables are important in estimation of violent crime in a city,from our model we can conclude that increase in percentage of population living in metropolitan areas, high school graduation rate, and poverty rate causes increase in crime rate per capita. Also we found that our model is better at estimating the crime rate compared average value of crime rate across the states. Because the analysis was done a sample data, we found that our model is less confident in estimating the impact of poverty and high-school graduation rate, for this we suggest city-wise data collection for robust model generation.

3 Does employment differ across economic sectors?

Panel data are datasets where each unit has multiple observations across a period of time. In this problem, we consider a panel dataset of manufacturing firms in the UK (see Table 6) from 1976 to 1984. For each firm, we have the year it was observed (`year`), the sub-sector of the manufacturing industry (`sector`), the logarithm of the number of employees (`emp`), a measure of the average wage in the firm (`wage`), an inflation-adjusted estimate of the company's gross capital stock (`capital`), and an index of value-added output (`output`). The goal of this analysis is to determine whether employment differs across economic sectors when controlling for year, wages, capital, and output.

Table 6: The first five rows of the employment data.

firm	year	sector	emp	wage	capital	output
1	1977	7	5.04	13.15	0.59	95.71
1	1978	7	5.60	12.30	0.63	97.36
1	1979	7	5.01	12.84	0.68	99.61
1	1980	7	4.72	13.80	0.62	100.55
1	1981	7	4.09	14.29	0.51	99.56

(a) How many distinct firms are represented in these data? What is the breakdown of the number of firms by sector? Create a table displaying this information. Additionally, create a plot to visualize the distribution of employment by sector, faceting by year. Comment on any trends you see in this plot.

(b) Use a standard F-test to obtain a p-value for the null hypothesis that mean employment does not vary across sectors, when controlling for year, wages, capital, and output. If this analysis were valid, what would be its conclusion? Why might the analysis not be valid? What are the potential consequences?

(c) We might want to adjust for the fact that each firm is being observed multiple times. Why is it not possible to add firm-level fixed effects to the regression?

(d) At least, we might want to carry out inference robust to error correlations within firms across years. Derive a cluster-robust version of the F-test based on the the ordinary least squares estimate $\widehat{\boldsymbol{\beta}}$, the Liang-Zeger estimate $\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}]$, and the Wald test perspective [Hint: Use the fact that in general, if $\boldsymbol{Z} \sim N(0, \boldsymbol{\Omega})$, then $\boldsymbol{Z}^T \boldsymbol{\Omega}^{-1} \boldsymbol{Z} \sim \chi^2_{\dim(\boldsymbol{Z})}$; there is no need for maximum likelihood theory or Fisher information matrices here.]

(e) Implement the test you proposed in part (d) in an R function called `robust_anova()`. Your function should take arguments `lm_fit`, `lm_fit_partial`, and `cluster`. The first two should be objects outputted by `lm()` on the full and partial models to be compared, and the third should be a formula object specifying the variable(s) to cluster on (the latter is like the `cluster` argument to `vcovCL()`; see the examples at `?vcovCL()`). Your function may call `vcovCL()`; no need to implement Liang-Zeger standard errors from scratch. Apply your function to get a robust analog of the p-value obtained in part (b). Comment on how the conclusion from the robust analysis compares to that of the standard one. [Hint: Use the command `which(!(names(coef(lm_fit)) %in% names(coef(lm_fit_partial))))` to extract the indices of the variables omitted in the partial model, i.e. the variables in S.]

3.1 Solutions

(a) Table 7 summarises number of firms across the sectors, in Fig 5, we can see that the trend of employment across the sectos remains relatively the same from 1976-1982, the drop in employment across sectors can be seen in 1983 and 1984 hinting at a recession.

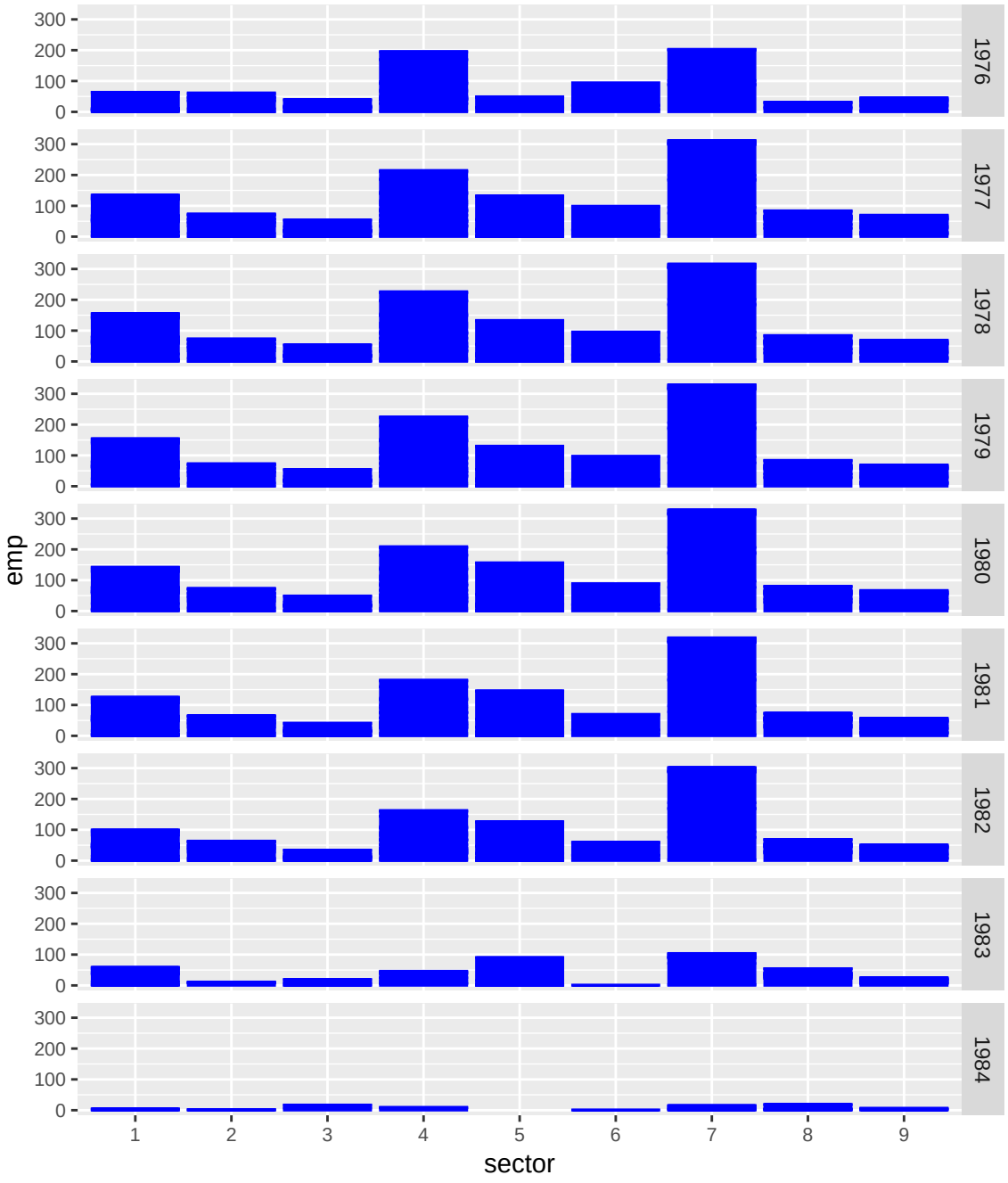

Figure 5: Plot of employment in each sector for every year

(b) From Table 8 (generated with anova in R-script) it can be seen that the p-value of the F test is highly significant, indicating there is dependence of sectors on employment. If there

Table 7: Distinct Firms

Sector	1	2	3	4	5	6	7	8	9
Firms by Sector	17	12	12	29	13	5	16	15	21
Total Firms									140

was no dependence of sector on employment we would see uniform employment across the sectors however this is not the case as seen in Fig 5. This analysis might not be valid because there are different of firms in each sector and employment demand in each firm varies across the years. From Fig 6 it can be seen that there is heteroskedasticity present in the model indicating there is correlation between residuals, to check the dependence of sectors we further check the box plot of residuals against the sectors, from Fig 7, it can be seen that the residuals lie close to zero with very little variance, this might indicate independence of employment on the sector, therefore robust error correlation based test is required. Also from the box plot 7 it seems that sector 3 has few extreme outliers, however this does not dramatically change the residuals within the sector.

Table 8: F-test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1026	79051.96	NA	NA	NA	NA
1018	72087.15	8	6964.81	12.29	0

(c) We might not be able to add firm-level fixed effects to the regression model because of within firm correlations across years. If there was no time component present in the data and if we had aggregate data over years, this could have been possible. Also adding a firm variable will not fix heteroskedasticity.

(d) The least squares estimate is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$$

Consider the Liang-Zeger estimate of the covariance matrix:

$$\widehat{\Sigma} \equiv \text{block} - \text{diag}\left(\widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_C\right), \quad \text{where} \quad \widehat{\Sigma}_c \equiv \widehat{\epsilon}_c\widehat{\epsilon}_c^T$$

Where $\widehat{\epsilon}_c$ is the vector of residuals in cluster c. The estimate of variance of covariance matrix of $\widehat{\beta}$ is given by:

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}] \equiv \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

$$\Rightarrow \widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1})$$

Consider

$$I_S = [e_1, e_2 ... e_s]$$

i.e a matrix of basis vectors for S parameters being tested. Therefore under the null hypothesis we have,

$$\Rightarrow I_S\widehat{\boldsymbol{\beta}} \sim N(0, I_S^T\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]I_S)$$

$$\Rightarrow \widehat{\boldsymbol{\beta}}_S \sim N(0,^T \widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]_{S,S})$$

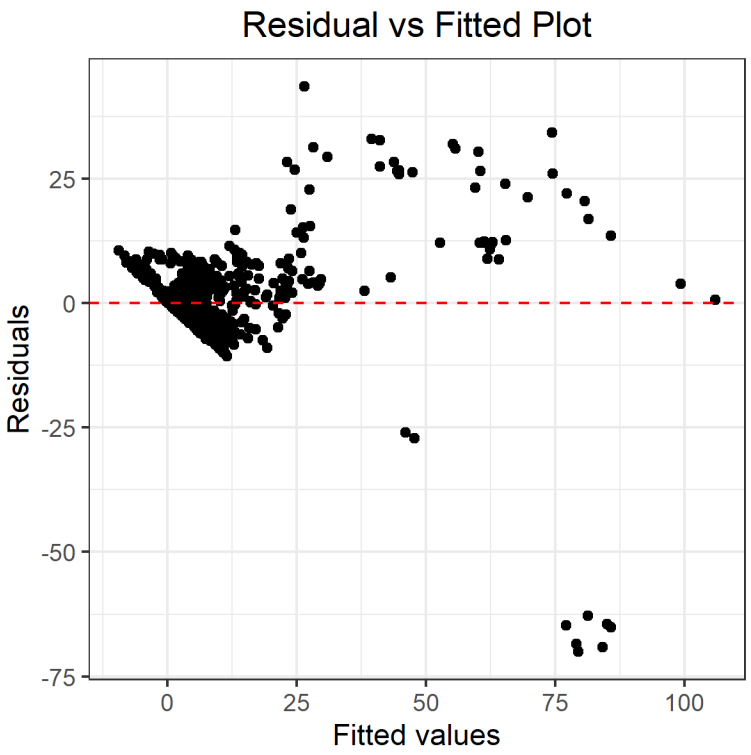

Figure 6: Residuals vs Fitted

Now using the fact that, $\boldsymbol{Z} \sim N(0, \boldsymbol{\Omega})$, then $\boldsymbol{Z}^T \boldsymbol{\Omega}^{-1} \boldsymbol{Z} \sim \chi^2_{\dim(\boldsymbol{Z})}$, we get

$$\widehat{\boldsymbol{\beta}}_S^T \widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}]^{-1}_{S,S} \widehat{\boldsymbol{\beta}}_S \sim \chi^2_{|S|}$$

(e) From Table 9 it can seen that the P-value is not above significance level (0.05), therefore we can say that inference robust to error correlations within firms across the years shows that there is no dependence of sectors on employment. Comparing this to result obtained in part (b), from Fig 8 it can be seen that standard error inflation is modest and only one sector has high inflation, from this we can say that robust error estimation based on clustering indeed detected co-linearity within firm across years. From Table 10, we can see that only three model coefficients are significant based Liang-Zeger covariance method, here we can clearly see the standard error inflation for all significant coefficients compared to estimates from OLS in Table 11.

Table 9: Robust error inference

T Statistic	12.00
P value (Based of χ^2_8)	0.15

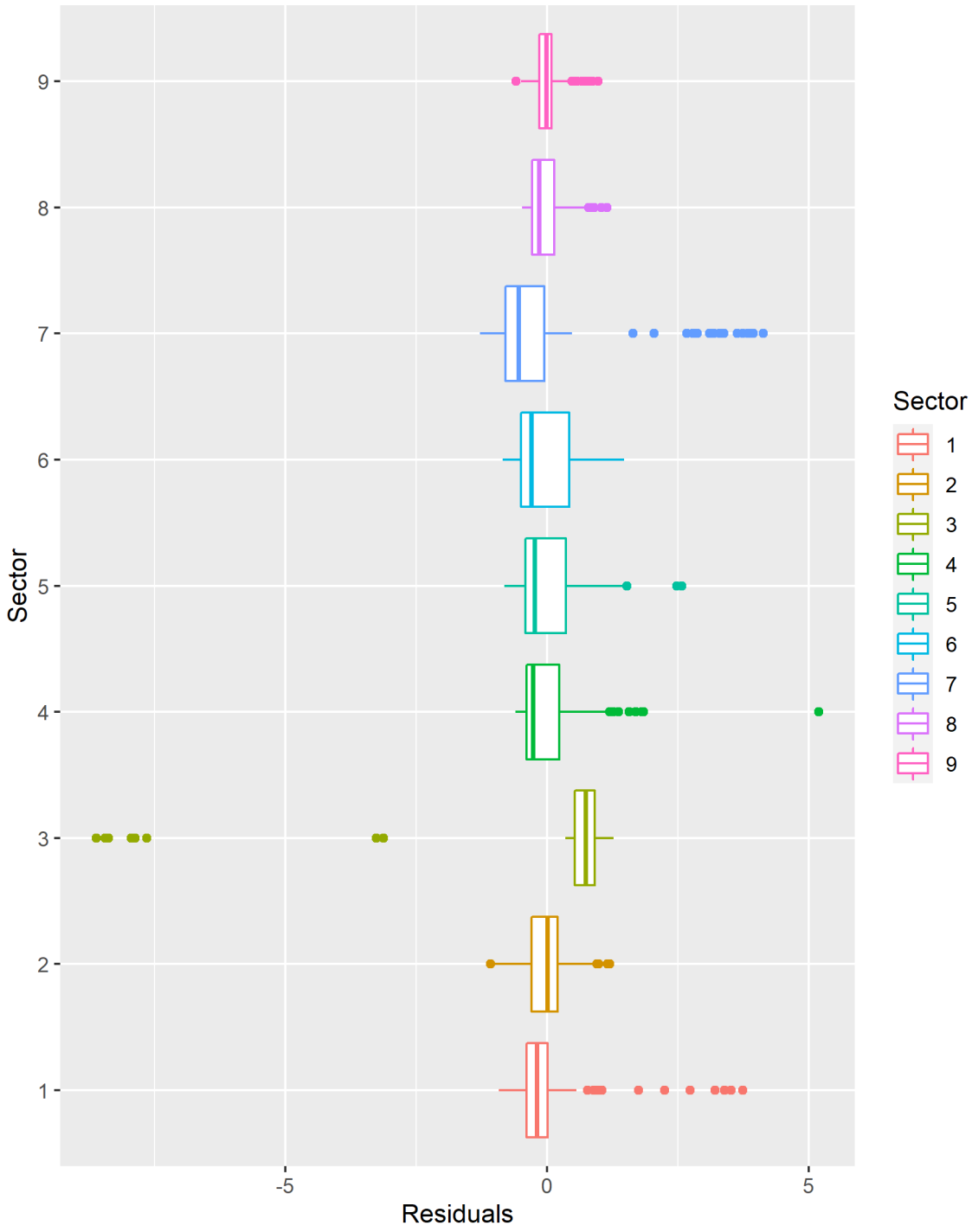

Figure 7: Box-plot of residuals grouped by sector.

Table 10: Summary of predictors in regression analysis using Liang-Zeger covariance method, siginificant at 0.05 level

Predictor	Estimate	Standard Error	p value
(Intercept)	1321.5180704	493.3710349	0.00751336
year	-0.6591101	0.2451087	0.00728261
capital	2.1115754	0.5106140	0.00003836

Table 11: Summary of predictors in regression analysis using OLS variance estimate, siginificant at 0.05 level

Predictor	Estimate	Standard Error	p value
(Intercept)	1321.5180704	329.50773652	0.00006500
year	-0.6591101	0.16488051	0.00006863
wage	-0.2209678	0.06474641	0.00066826
capital	2.1115754	0.04381546	0.00000000
sector3	-9.2228157	1.24375026	0.00000000
sector9	-2.3685525	1.11594926	0.03404042

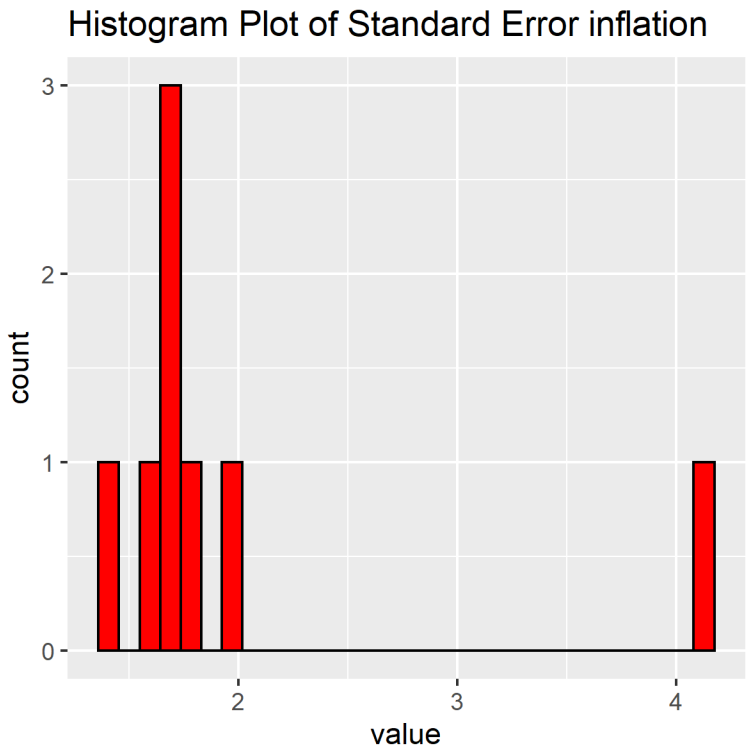

Figure 8: Histogram of Standard Error Inflation

4 Pollution data

In this problem, we will analyze a data set related to pollution (`pollution.tsv`), whose first five rows are shown in Table 12 below. These data contain hourly measurements of nitric oxide (NOx)

Table 12: The first five rows of the pollution data.

date	log_nox	wind
373	4.46	0.86
373	4.15	1.02
373	3.83	1.10
373	4.17	1.35
373	4.32	1.20

concentration in ambient air (in parts per billion) next to a highly frequented motorway. The first column is an integer specifying the date the observation was taken (the hour each observation was taken is not available). The second column is the logarithm of the nitric oxide concentration. The third column is the square root of the windspeed in meters/second. The goal is to learn how NOx concentration depends on wind speed.

(a) Create some plots and/or summary statistics to explore the data. Comment on any trends you observe.

(a) Run a linear regression of `log_nox` on `wind`, and produce a set of relevant diagnostic plots. What model misspecification issue(s) appear to be present in these data?

(b) Address the above misspecification issues using one or more of the strategies discussed in Unit 3. Report a set of statistical estimates, confidence intervals, and test results you think you can trust.

(c) Discuss the findings from part (b) in language that a policymaker could comprehend, including any caveats or limitations of the analysis.

4.1 Pollution Data Analysis

(a) Fig 9 shows the distribution of NOx against wind speed, we can see that there is a negative correlation between NOx and wind speed, Fig 10, highlights this correlation by averaging over days. Fig 11 shows the temporal distribution of NOx against windspeed on different days, clearly the distribution varies everyday, indicating there might be within day correlation. From Fig 12 we can see that there is some heteroskedasticity in the data and from the Fig 13 we can see that no point has high values of residuals and leverage indicating there is no outlier in this data. Also for the linear model fit we only have one predictor i.e windspeed, and from model summary in Tab 11, we can see that this coefficient is significant, therefore we can say there is no model bias.

(b) Fig 14 shows the confidence bands determined by variance estimates obtained with OLS, clustered pairs bootstrap and Liang-Zeger method. It can be seen from the plot that OLS provides the tightest bands among the three and bootstraping and Liang-Zeger provide similar bands. Further Table 13 and 14 summarises the significant coefficients based on the t-statistic with the two variance estimates, also the standard error inflation is not big, meaning the two methods are efficient. Further Table 16 and 17 show the confidence intervals for intercept and

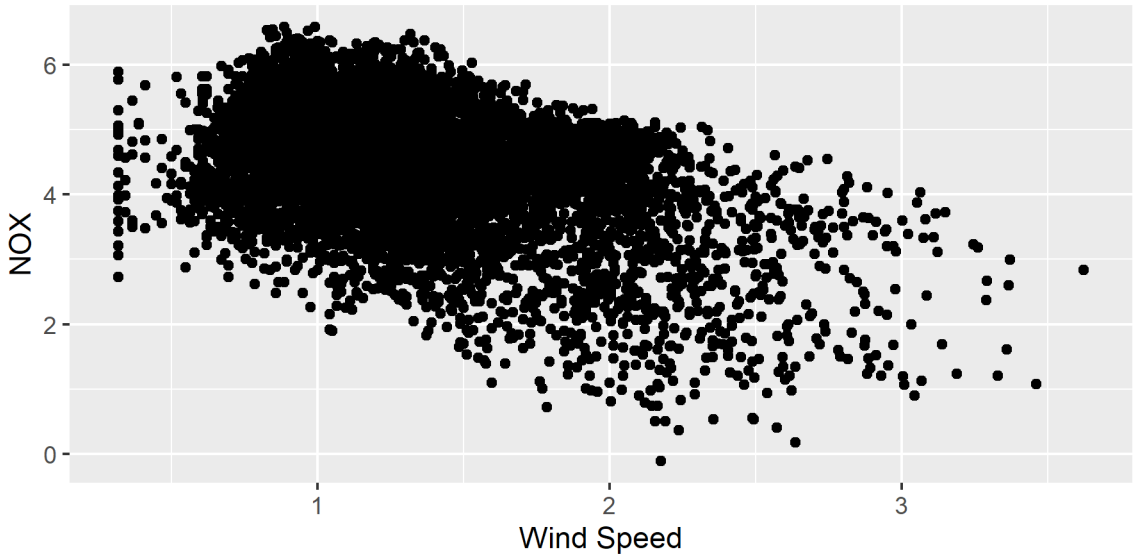

Figure 9: Plot of NOX vs Wind speend

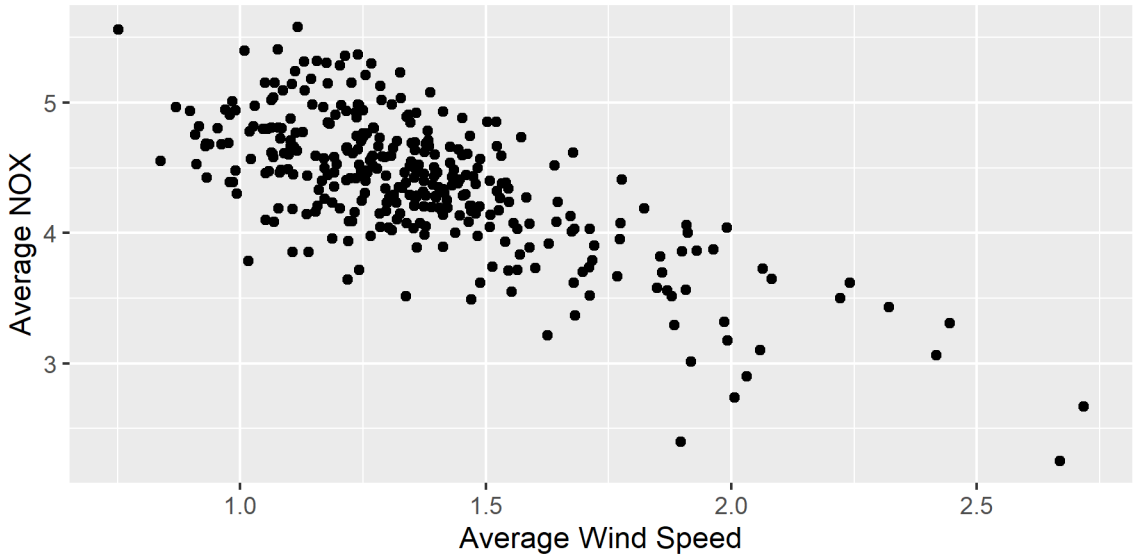

Figure 10: Plot of NOX vs Wind Speed averaged over days

Table 13: Summary of predictors in regression analysis using Liang-Zeger variance estimate

Predictor	Estimate	Standard Error	Statistic
p value			
(Intercept)	5.5588538	0.06475863	85.83958
0			
wind	-0.8644279	0.04775083	-18.10289
0			

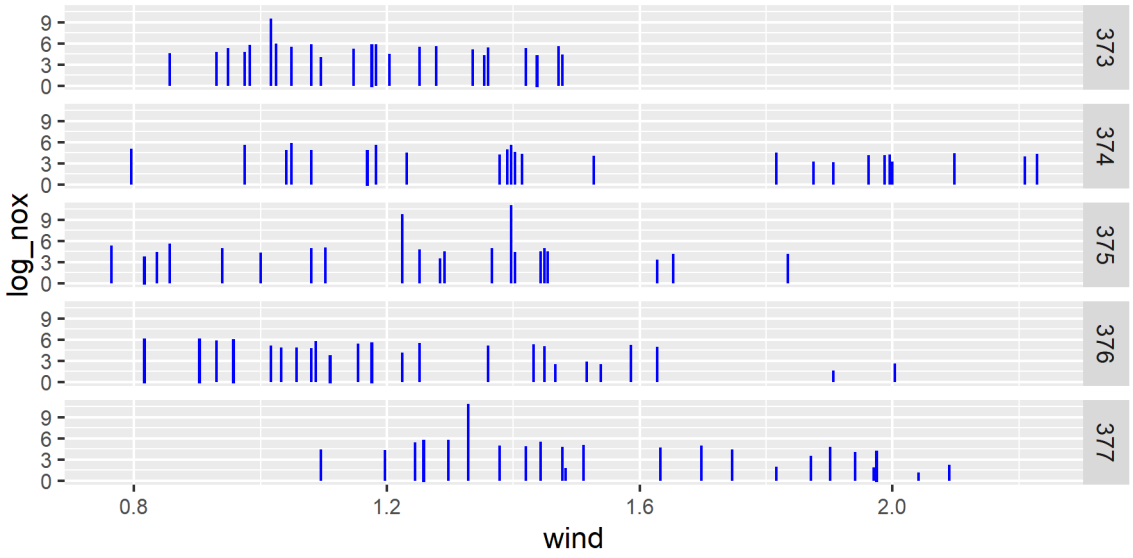

Figure 11: Plot of NOX vs Windspeed on different days

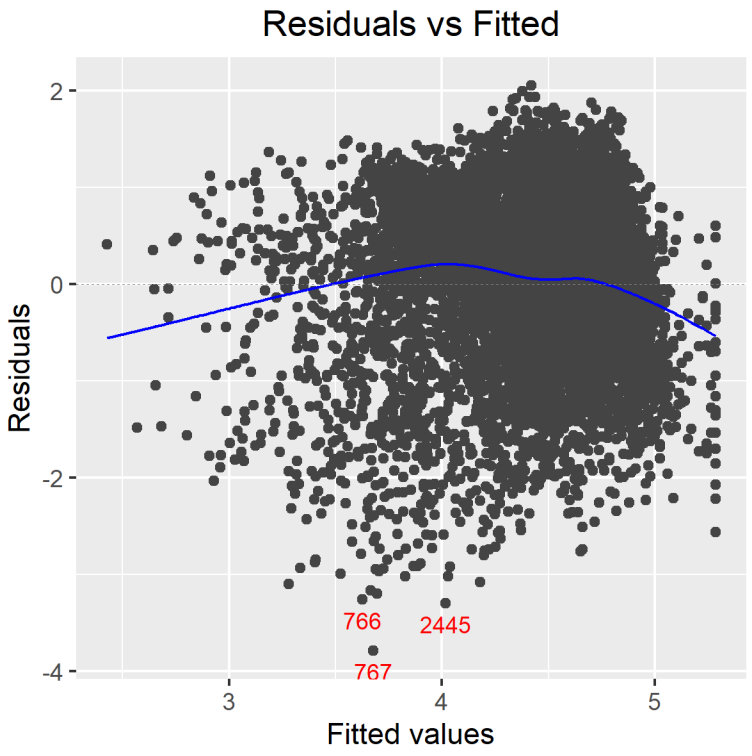

Figure 12: Residuals against fitted values

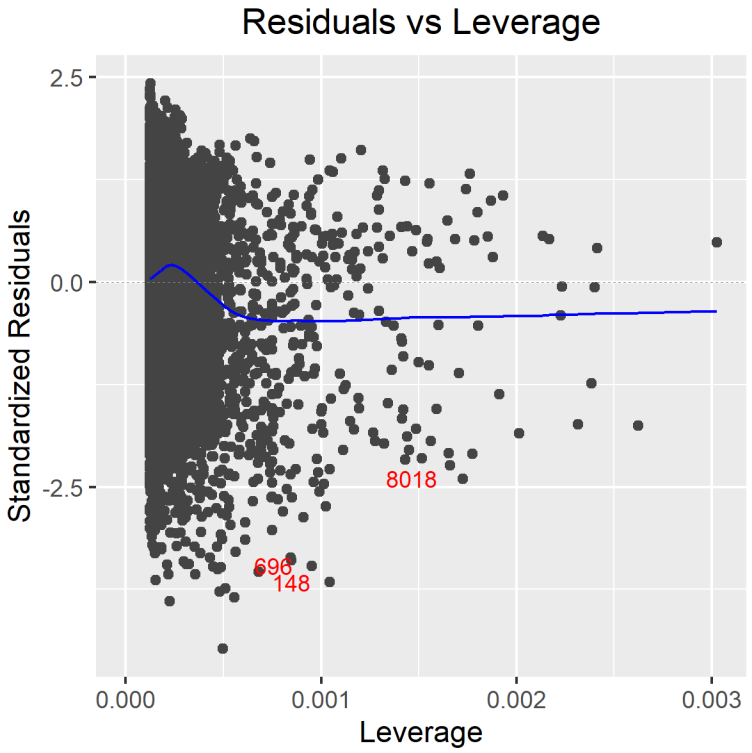

Figure 13: Residuals against Leverage

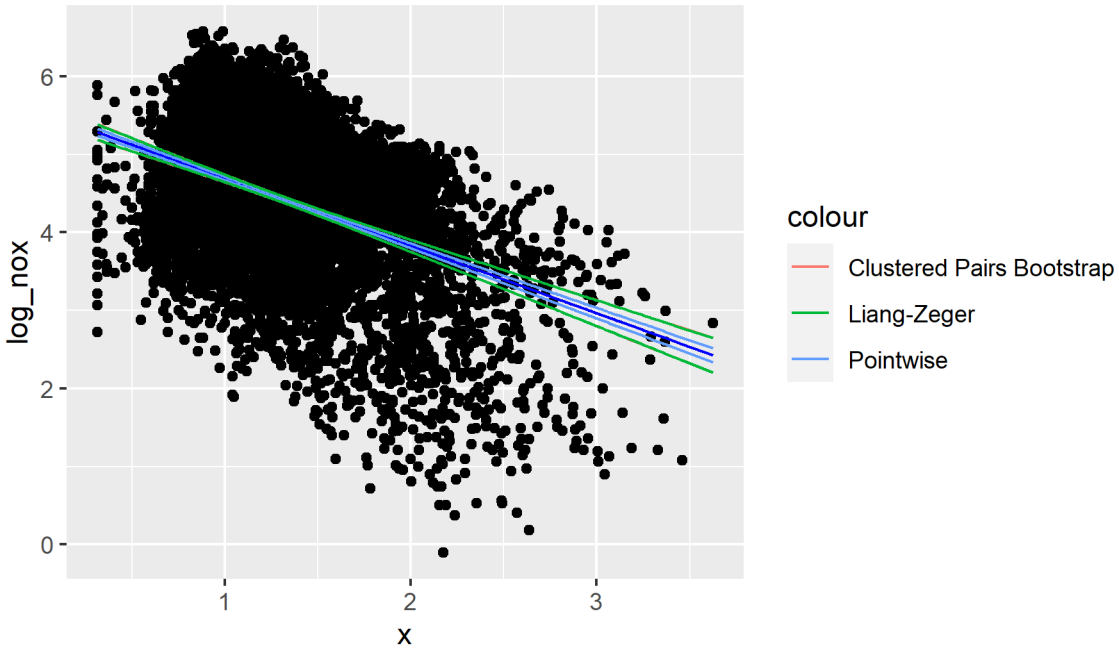

Figure 14: Linear model fit with different variance estimates

the slow respectively, cleary the width is inflated with the two methods but not significantly.

Table 14: Summary of predictors in regression analysis using clustered pairs bootstrap variance estimate

Predictor	Estimate	Standard Error	Statistic
p value			
(Intercept)	5.5588538	0.06662831	83.43082
0			
wind	-0.8644279	0.04858193	-17.79320
0			

Table 15: Summary of predictors in regression analysis using OLS variance estimate

Predictor	Estimate	Standard Error	Statistic
p value			
(Intercept)	5.5588538	0.02911941	190.89856
0			
wind	-0.8644279	0.02018434	-42.82666
0			

Table 16: Confidence interval for β_0

	Bootstrap	Liang Zeger	OLS
lower	5.428245	5.431910	5.501772
upper	5.689462	5.685797	5.615935

Table 17: Confidence interval for β_1

	Bootstrap	Liang Zeger	OLS
lower	-0.9596609	-0.9580318	-0.9039944
upper	-0.7691948	-0.7708240	-0.8248614

(c) From our analysis of dependence of NOx on windspeed we have concluded that, increase in wind speed reduces the NOx amount in the air. Also this decrease is highly significant meaning drop in wind speed will almost certainly cause an increase in NOx concentration. In our analysis we also included the effect of daily pattern, we found that on a given day, the measurments are related with each other and therefore we cannot aggregate the daily data and present it in compressed format.

5 Child development.

Children were asked to build towers as high as they could out of cubical and cylindrical blocks.[1] The number of blocks used and the time taken were recorded (see Table 18 below). In this problem, only consider the number of blocks used and the age of the child.

Table 18: The first five rows of the blocks data.

Child	Number	Time	Trial	Shape	Age
A	11	30.0	1	Cube	4.67
B	9	19.0	1	Cube	5.00
C	8	18.6	1	Cube	4.42
D	9	23.0	1	Cube	4.33
E	10	29.0	1	Cube	4.33

(a) Create a scatter plot of blocks used versus age; since there are exact duplicates of `(Number, Age)` in the data, use `geom_count()` instead of `geom_point()`. Propose a GLM to model the number of blocks used as a function of age.

(b) Fit this GLM using R, and write down the fitted model. Determine the standard error for each regression parameter, and find the 95% Wald confidence intervals for the regression coefficients.

(c) Use Wald, score, and likelihood ratio tests to determine if age seems necessary in the model. Compare the results and comment.

(d) Plot the number of blocks used against age as in part (a), adding the relationship described by the fitted model as well as lines indicating the lower and upper 95% confidence intervals for these fitted values.

Acknowledgment: This problem was drawn from "Generalized Linear Models With Examples in R" (Dunn and Smyth, 2018).

5.1 Child Development Data Analysis

(a) Since the number of blocks used is an integer value, a Poisson regression model for the problem is suitable. Fig 15, shows the number of blocks required as a function of age.

(b) The fitted model reads as:

$$\text{Blocks} \sim \text{Poi}(\mu_i), \quad \log(\mu_i) = 1.34 + 0.14age$$

Table 19 summarises the regression output for the Poisson GLM.

[1] Johnson, B., Courtney, D.M.: Tower building. Child Development 2(2), 161–162 (1931).

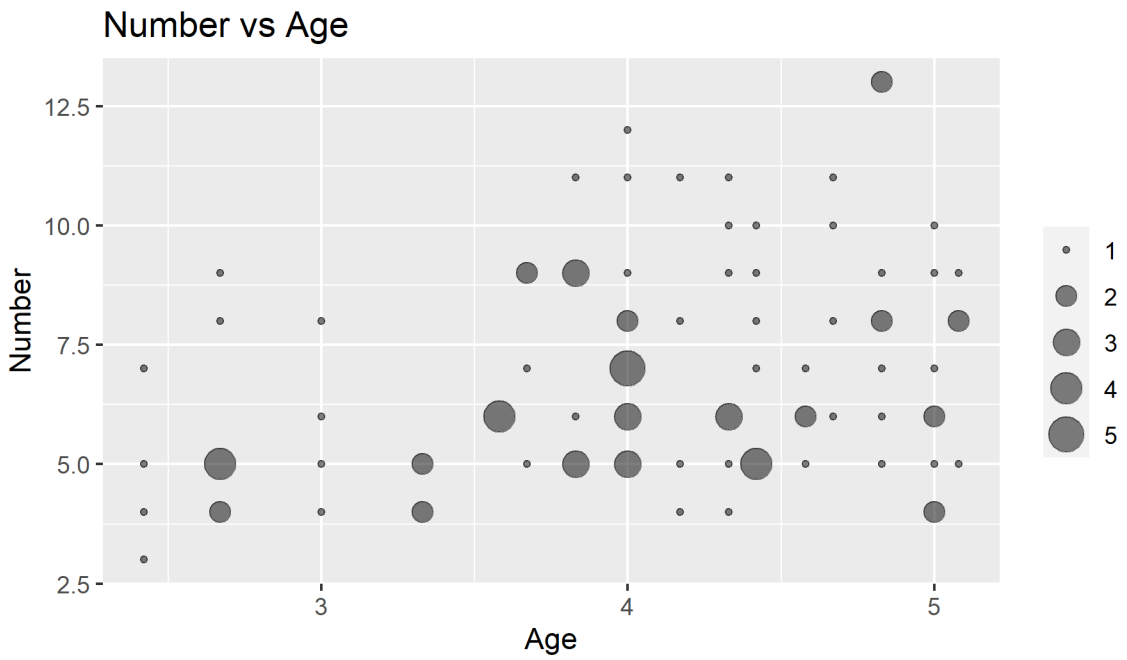

Figure 15: Count plot of number of blocks vs age

Table 19: Poisson Regression Summary

Coefficient	Estimate	Std Err	2.5 %	97.5%
Intercept	1.3447	0.2223	0.91	1.78
Age	0.1415	0.0534	0.037	0.25

(c) From regression summary in Table 19, we have wald type t-statistic=2.650, and the corresponding P-value = 0.00805, for LRT based test we use an intercept-only model and perform ANOVA with fitted model, the corresponding P-value=0.00735, Table 20, summarises these results, we can see that both tests give results on same order of magnitude and we can concluded that age is an important factor because P-values are less than 0.05.

Table 20: P-value Summary

Test	P-value
Wald	0.00805
LRT	0.00735

(d) Fig 16, shows the fitted model and confidence intervals for the model along with the dataset. The mean values (in blue) are obtained by taking the exponential of the fitted values (Calculation shown in the R-script).

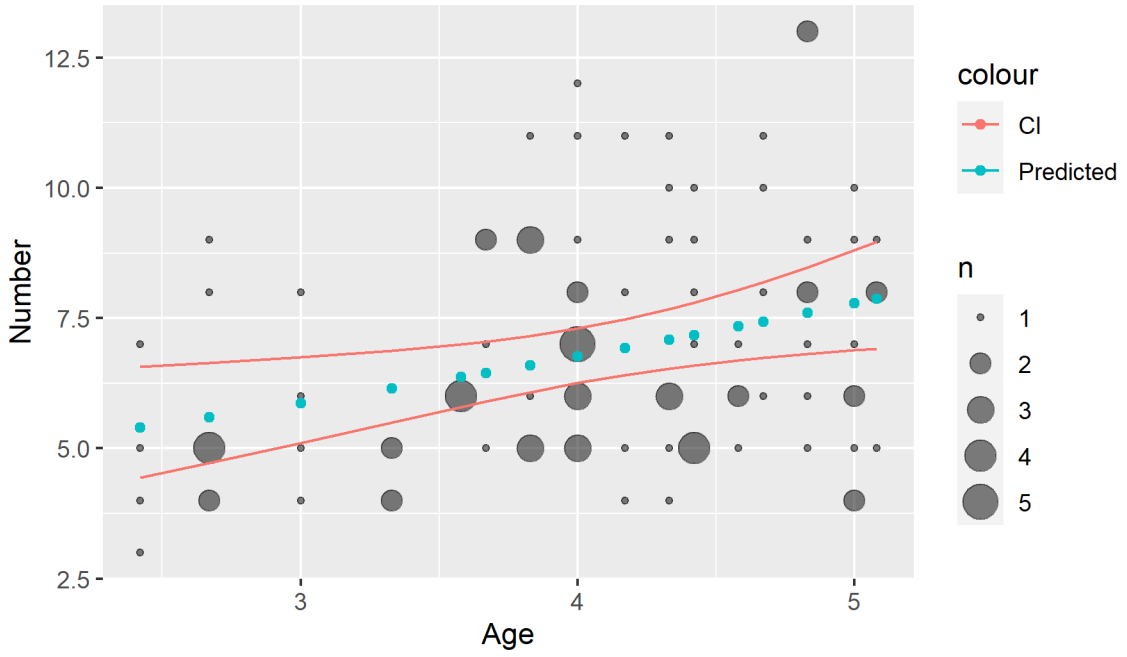

Figure 16: Count plot of number of blocks vs age

6 Testing for association between income and job satisfaction, given gender.

Consider the job satisfaction data (Table 21), which cross-tabulate income, job satisfaction, and gender:

Table 21: The first five rows of the job satisfaction data.

Income	Job.Satisfaction	Gender	Count
<5000	Very Dissatisfied	Female	1
5000-15000	Very Dissatisfied	Female	2
15000-25000	Very Dissatisfied	Female	0
>25000	Very Dissatisfied	Female	0
<5000	A Little Satisfied	Female	3

We'd like to test whether there is a relationship between income and job satisfaction, conditional on gender.

(a) Create a plot to visualize the relationship between income and job satisfaction for males and females, making sure to respect the natural orderings of the income and job satisfaction variables. Comment on the trends you observe in this plot.

(b) Implement the test from Problem 1c on the `job_satisfaction` data. You may use the `glm()` function but not more specialized functions for conditional independence testing. What p-value do you obtain, and what is the corresponding conclusion?

(c) Why may the test implemented in part (b) be underpowered? Propose a test statistic that may be able to more sensitively pick up the relationship between income and job satisfaction.

(d) Note that the conditional permutation test from Problem1e can be applied to calibrate *any* test statistic, not just the Pearson X^2 statistic. Implement the conditional permutation test to calibrate the statistic you proposed in part (c). What is the resulting p-value? What is your conclusion about the relationship between income and job satisfaction, controlling for gender?

(e) Suppose you applied a regular permutation test to search for association between income and job satisfaction, ignoring the gender variable. Would this result in a valid p-value for testing independence between income and job satisfaction, conditional on gender? Why or why not?

6.1 Analysis

(a) From Fig 17 it can be seen that most Females lie in the moderately satisfied category for their jobs across all the income categories, similar is the case for males as seen in Fig 18, both genders also have least people in the very dissatisfied category, we can also see that females on all income levels are nearly uniformly spread in the very satisfied category whereas males with high-income salary tend to be very satisfied. This suggests there is a relationship between the ordering of the two categorical variables and females and males have different job satisfaction level based on their income.

(b) From part 1c, we have:

$$X^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(y_{jkl}-\widehat{\mu}_{jkl})^2}{\widehat{\mu}_{jkl}}; \quad \widehat{\mu}_{jkl} = \frac{y_{j+l}y_{+kl}}{y_{++l}}$$

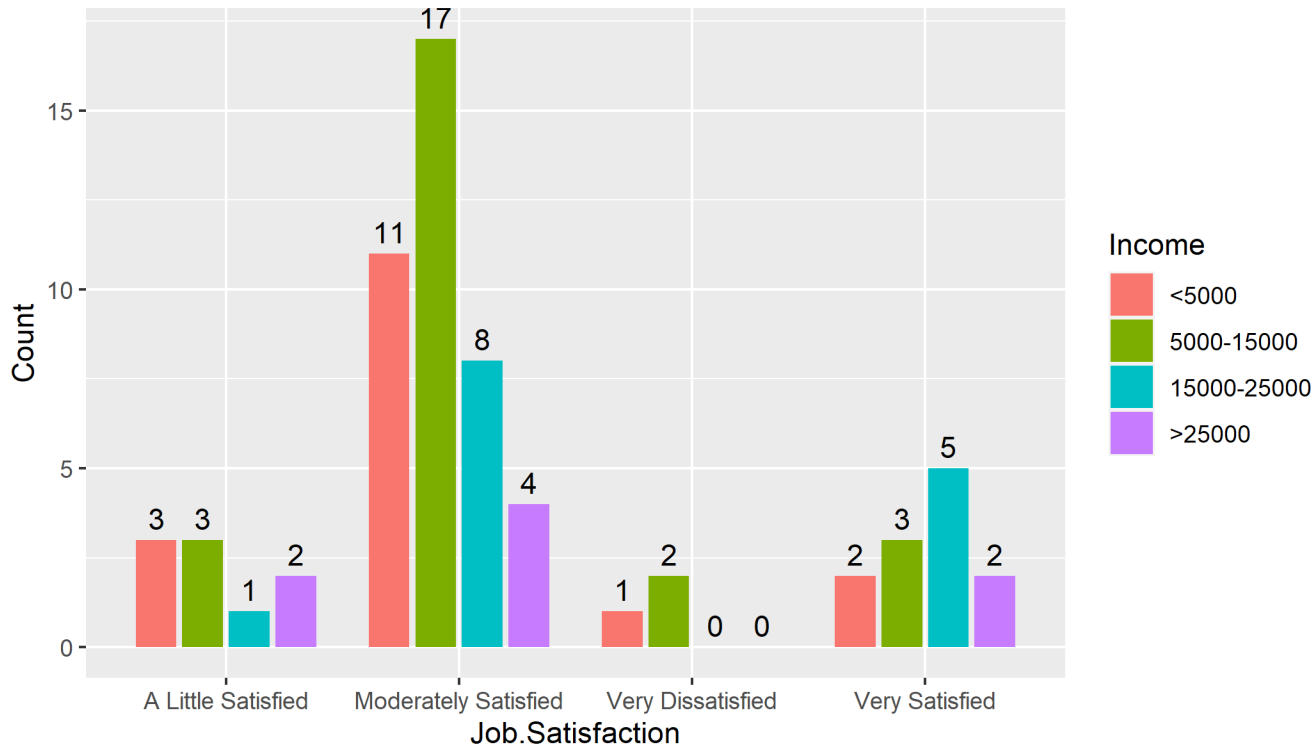

Figure 17: Histogram plot for Females

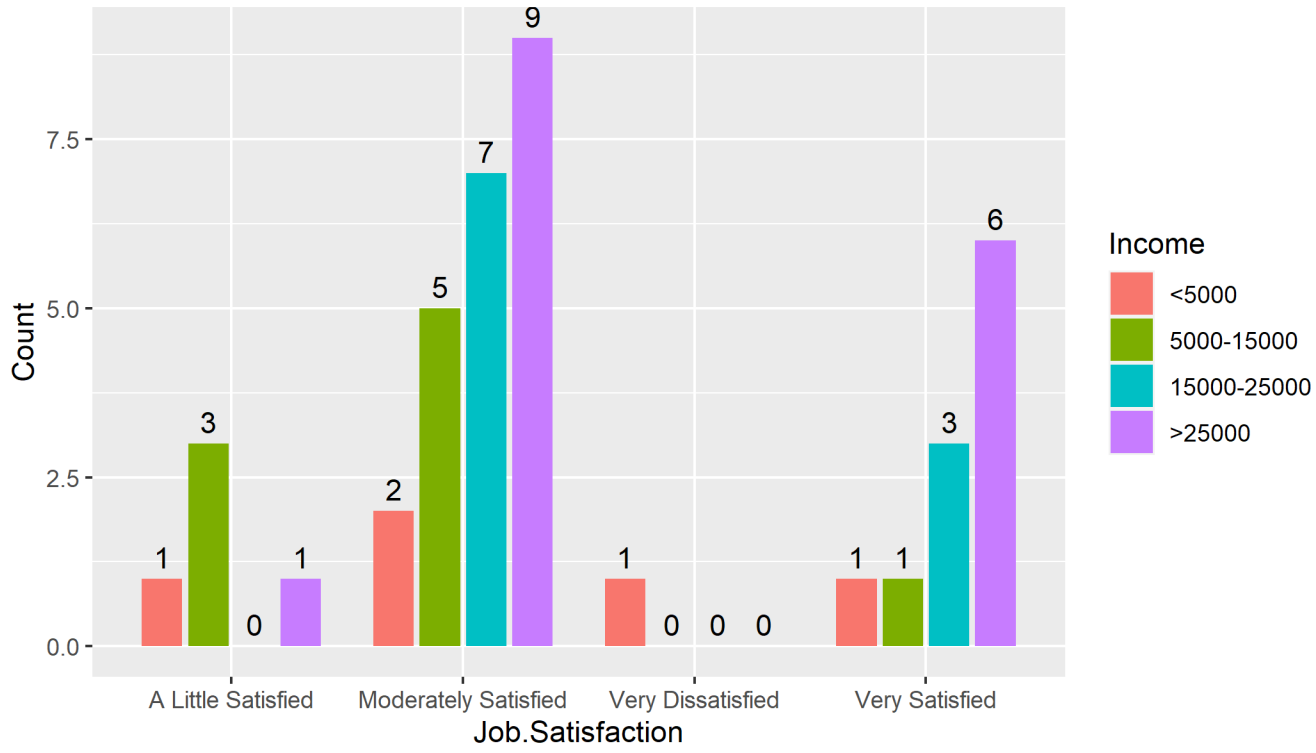

Figure 18: Histogram plot for Males

$$X^2 \sim \chi^2_{n-p}$$
$$n - p = L(J-1)(K-1) = 2 * (4-1) * (4-1) = 18$$

Table 22 summarises the test statistic baesd on above formulation,

Table 22: P-value Summary for score test

Score Test	P-value
21.05	0.277

Since the P-value is very significant, we can conclude that income is independent of job satisfaction conditioned on gender.

(c) The test implemented in part b may be underpowered because the sample size is too small and there is clear ordering in job satisfaction and income, to more accurately pick up the relationship between income and job satisfaction we would want to capture the monotonic relationship between the two categories. To this end, we rank each categorical variable in natural order (1,2,3,4) and use **Kendall tau-b** correlation as the test statistic as recommended in [2]. Which can measure the correlation between ranked variables income and job satisfaction.

(d) To perform a conditional permutation test, we will keep job satisfaction fixed and permute the income of individuals, in each gender category separately, we then combine the two permuted vectors and compute Kendall tau-b correlation between income and job satisfaction. We then define p-value as:

$$\widehat{p}^{\text{perm}} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} 1\left(T\left(\boldsymbol{X}, \boldsymbol{y}_{\tau_b}\right) \geq T(\boldsymbol{X}, \boldsymbol{y})\right)\right)$$

Where B is the total number of permutations. $T(\boldsymbol{X}, \boldsymbol{y})$ is the test statistic for given data and $T\left(\boldsymbol{X}, \boldsymbol{y}_{\tau_b}\right)$ is the test statistic for permuted data. Table 23 shows the calibrated p-value for 5000 permutation, based on this p-value we can reject the null the hypothesis, that job satisfaction and income are independent conditioned on gender.

Table 23: P-value Summary for Kendall tau ranked correlation test test

Test	P-value
Kendall's tau	0.007

(e) The question of independence between job satisfaction and income can be asked for collective population without considering the gender, however this might not yield a correct p-value because we have seen from the data set that males and females rank their satisfaction level differently acrossv income level.

[2] Alan Agresti: An introduction to categorical data analysis

7 Bradley-Terry model for the NBA.

The NBA has 30 teams, and each team plays a total of 82 games during the regular season. If the game between team j and team k takes place in team j's arena, then team j is said to be the "home" team and team k is said to be the "away" team. Suppose that each team $j \in \{1, \dots 30\}$ has an associated parameter β_j representing how good the team is, and β_0 is a parameter representing "home court advantage." Then, a simple model for the outcome of the match between team j and team k is

$$\text{logit}(\mathbb{P}[\text{home team } j \text{ beats away team } k]) = \beta_0 + \beta_j - \beta_k. \tag{1}$$

This model is called the *Bradley-Terry model*. In this problem, we'll be fitting the Bradley-Terry model to data from the NBA 2017-2018 season (Table 24):

Table 24: The first five rows of the NBA data.

game	team	conference	location	result
1	BOS	East	Away	Loss
1	CLE	East	Home	Win
2	HOU	West	Away	Win
2	GS	West	Home	Loss
3	CHA	East	Away	Loss

(a) (Identifiability) This model suffers from an identifiability issue. State the issue and how you could restrict the parameters to resolve this issue. A more subtle identifiability issue would arise if the teams could be split into two groups such that games were played only within groups. Discuss why the latter issue would arise and check whether this issue is a concern in the given data.

(b) (Model fitting) Reformulate the Bradley-Terry model as an ungrouped logistic regression model (i.e. what are the predictors and the response)? Based on this reformulation, transform the data into the format expected by `glm()`, print the resulting tibble (no need for a kable table), and then call `glm()` on this tibble to obtain fitted model parameters.

(c) (Home court advantage) Produce a point estimate and Wald confidence interval for the factor by which the odds of winning increase with home court advantage. Comment on the direction and magnitude of the effect. Produce a scatter plot of home game win percentage versus away game win percentage (this plot should contain 30 points, one per team), and comment on what this plot says about home court advantage.

(d) (Team rankings) Produce a table with 30 rows and four columns: team, win percentage for home games, win percentage for away games, and estimated value of β_j. Order the columns in descending order of $\widehat{\beta}_j$. Comment on the degree of concordance or discordance between the columns in this table. Intuitively, what might be a reason why a team has a relatively higher win percentage but a relatively lower value of $\widehat{\beta}_j$?

(e) (NBA finals prediction) The NBA finals in 2017-2018 were between the Golden State Warriors (GS) and the Cleveland Cavaliers (CLE). The first two games were in Oracle Arena (Oakland, California) while the last two games were in Quicken Loans Arena (Cleveland, Ohio). Golden State won all four of these games and the NBA championship. For each of these four games, what are point estimates and Wald confidence intervals for the probability that Golden State won?

7.1 Analysis

(a) There is an issue of identifiability because we have 31 coefficients and the rank of $C(X)$ matrix is 30, this is because when $X_J = 1$, we have $X_K = -1$, therefore the coefficients cannot be uniquely identified, this can be easily verified by looking at a model structure if we increase/decrease all the coefficients by the same amount, the inference does not change, to mitigate this issue we can set one of the coefficients to zero and estimate other coefficients w.r.t the coefficient which is set equal to zero. We can set any coefficient to zero and the estimation of log odds will not change because we are looking at differences.
If the teams have only played half of the teams in a particular "conference" (east/west), then we cannot identify which team is stronger amongst cross-conference matches, however in this dataset we do have data for east vs west matches and identifiability is not an issue in this problem.

(b) Reported in R-script.

(c) For this problem, we estimate the coefficients w.r.t Phoenix Suns. Table 25 shows the estimate of home court advantage, the positive value suggests the odds of team j winning against team k at the home court are boosted by exp(0.38). Also the confidence interval does not contain zero, meaning there is definite advantage of playing at home court for any team. From Fig 19 we can see that teams tend to win more matches on the home court compared to away games, the line indicates y=x, and most teams have more home wins against away wins, this implies there indeed is a home court advantage.

Table 25: Point estimate and confidence interval for home court advantage

Coefficient	Estimate	Std Err	2.5 %	97.5%
Advantage	0.38	0.0640	0.254	0.505

(d) In the above part we found that teams have higher win percentages at home court than in away games, the added home court advantage gives higher win percentage at home games, this is also true for a weak team, therefore the strength of a team can really be identified by looking at away games win percentage. Table 26 shows top 5 teams according to our model. The extra column of total win percentage out of 82 wins clearly indicates that strength corresponds to total wins and not just home or away wins.

Table 26: Top 5 Teams

Team	Home Wins	Away Wins	Total Wins	Strength (β)
HOU	82.92	75.60	79.26	2.55
TOR	82.92	60.97	71.95	2.10
GS	70.73	70.73	70.73	2.08
BOS	65.85	68.29	67.07	1.86
PHI	73.17	53.65	63.41	1.71

(e) Table 27 summarises the win probabilities for GSW in NBA finals.

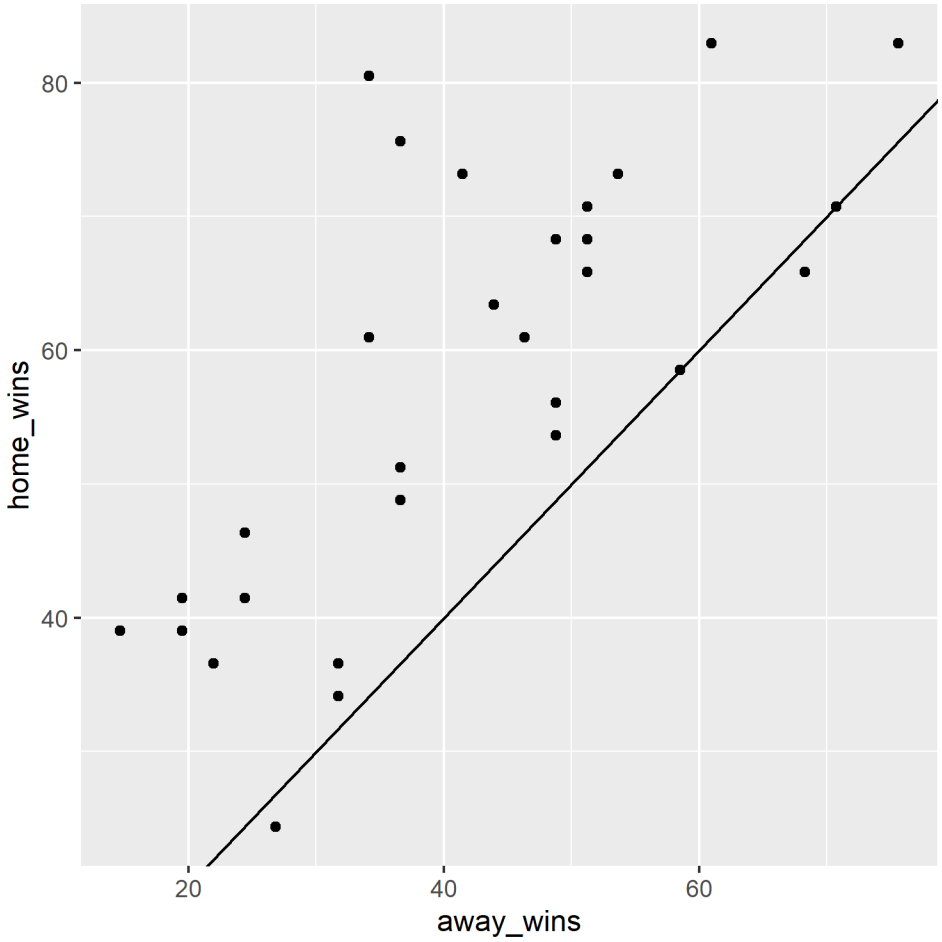

Figure 19: Home wins vs Away Wins

Table 27: Confidence interval for Winning probabilities of GSW

Game	Estimate	2.5 %	97.5 %
1	0.71	0.55	0.83
2	0.71	0.55	0.83
3	0.53	0.69	0.36
4	0.53	0.69	0.36

8 Applying gamma regression

In this problem, you will apply the estimation and inference procedures you derived in Problem **??** and implemented in Problem **??** to data related on the *body mass index* (BMI), defined for a person as their mass (in kilograms) divided by the square of their height (in meters). According to Wikipedia,

> "The BMI is a convenient rule of thumb used to broadly categorize a person as underweight, normal weight, overweight, or obese based on tissue mass (muscle, fat, and bone) and height. Major adult BMI classifications are underweight (under 18.5 kg/m^2), normal weight (18.5 to 24.9), overweight (25 to 29.9), and obese (30 or more)."

The data (`bmi_data.csv`) come from the 2014-16 Eating and Health Module of the American Time Use Survey. They contain information on eating and exercise habits of over 10,000 individuals, as well as their BMI; for more information see the codebook (provided to you as `codebook.pdf`). The goal of the analysis is to identify factors affecting a person's BMI.

(a) (Data cleaning.) This dataset has many missing values, coded as data entries taking on either negative or zero values. Remove all variables from the data for which at least 40% of observations are missing. Then, remove all rows with any missing values. Some of the remaining variables will have only one unique value among the remaining observations; remove these variables as well. Finally, convert all variables except `erbmi`, `ertpreat`, `euexfreq` to factors. [Hint: You may find the `across()` functionality of `dplyr` useful for this problem; see the documentation online.]

(b) (Data exploration.) How many rows and columns are left in the data after the cleaning in part (a)? How many predictors are there? Create a histogram of the response variable `erbmi`, with dashed vertical lines at 18.5, 25, and 30. Using `annotate()`, label the resulting four BMI regions as underweight (UW), normal (NM), overweight (OV), and obese (OB). Based on this plot, comment on the rough fraction of survey respondents falling into the normal BMI range.

(c) (Wald inference.) Using the functions you wrote in Problem **??**, carry out Wald inference for the gamma regression of BMI on the other variables with logarithmic link. Using `kable()`, print the table outputted by `wald_inference()`, with rows ordered by increasing p-value. Why are there more rows in this table than the number of predictors you found in part (b)? [There is no need to clean up the variable names. Do express the p-values in scientific notation; see e.g. here. If you did not complete Problem **??**, you may use the built-in `glm()` function to complete this portion of the problem.]

(d) (Multiplicity adjustment.) Apply the Bonferroni FWER procedure at level $\alpha = 0.05$ to the Wald p-values obtained in part (c). Produce a table like that in part (c), except containing only the Bonferroni-significant predictors. Also, clean up the variable names for readability.

(e) (Interpretation.) Comment on the directions of the effects found in part (d). Do these make sense to you? Why or why not? Interpret the fitted coefficients in terms of the change in BMI associated with a unit increase in the corresponding predictors. Comment on the sizes of the effects. For each Bonferroni-significant predictor, create a plot of how BMI varies based on that predictor, superimposing the three boundary lines as in part (b) for reference. Discuss the trends you find in these plots.

8.1 Application to BMI Data

(a) Implemented in R-script

(b) There are 2591 rows and 15 columns in the dataset after data cleaning. There are 14 predictors and 1 response variable. Based on Fig **??** it can be said that around 30-40 % of the respondents fall into the normal BMI range.

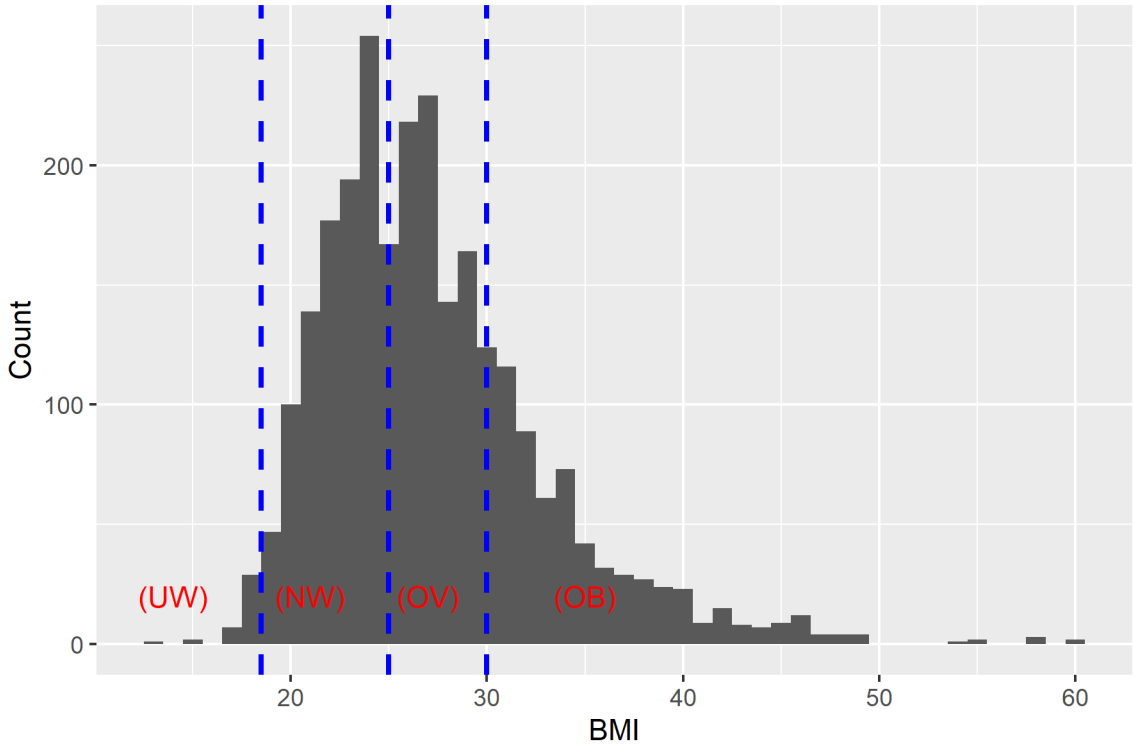

Figure 20: Histogram plot for BMI

(c) Tables 28 summarises the wald test. There are more coefficients in the output because we have categorical data. For every categorical variable, we have c-1 coefficients where c is the number of categories for that variable.

(d) Table 29 shows the significant values based on Bonferroni FWER. The significant variables are: Intercept, Exercise frequency, Household Income Category 2, Beverage Consumption, and category 2. The significance of categorical variables is with respect to their baseline category, the base category for household income is **Income > 185% of the poverty threshold** and for Beverage Consumption is **Yes**. Household Income: Level 2 means **Income < = 185% of the poverty threshold** and Beverage Consumption: Category 2 means **No**.

(e) According to the GLM fit, exercising more frequently would tend to decrease the BMI, also being in a household with Income < = 185% of the poverty threshold the BMI is more likely to be higher than the baseline BMI, people who don't drink Beverages tend to have lower BMI than people who do. All of these effects make sense as exercising and not drinking beverages are known to help maintain a lower weight and being in a household of lower income one is more likely to be eating fast food to save time and work more. Below is the summary of the magnitude and direction of these predictors.

The intercept can be interpreted as the mean response of a person belonging to a household of Income > 185% of the poverty threshold and does not exercise and drinks beverages this

Table 28: Wald Statistics for BMI fit

	Estimates	SE	t_value	p_value
(Intercept)	3.36e+00	4.32e-02	77.74820	0.00e+00
euexfreq	-7.95e-03	1.58e-03	-5.02743	5.31e-07
eeincome12	4.89e-02	9.98e-03	4.89342	1.05e-06
eusoda2	-2.98e-02	8.76e-03	-3.40546	6.71e-04
eufdsit2	6.62e-02	2.21e-02	2.99108	2.81e-03
eustreason3	-2.92e-02	1.31e-02	-2.23303	2.56e-02
eufastfd2	-1.92e-02	8.74e-03	-2.19268	2.84e-02
ertpreat	-1.81e-04	8.77e-05	-2.06131	3.94e-02
eueat2	1.68e-02	8.63e-03	1.94108	5.24e-02
eeincome13	4.33e-02	2.25e-02	1.92466	5.44e-02
eustreason6	-3.71e-02	2.06e-02	-1.80234	7.16e-02
eustores5	3.95e-02	2.82e-02	1.40335	1.61e-01
eustores3	-2.21e-02	2.01e-02	-1.10272	2.70e-01
eugroshp3	1.47e-02	1.48e-02	0.99496	3.20e-01
eustores4	-6.74e-02	6.94e-02	-0.97194	3.31e-01
eustreason5	3.21e-02	3.33e-02	0.96282	3.36e-01
eustreason4	-1.39e-02	1.53e-02	-0.90439	3.66e-01
eustreason2	-8.41e-03	1.04e-02	-0.80815	4.19e-01
eustores2	6.97e-03	1.03e-02	0.67646	4.99e-01
erhhch3	9.55e-03	1.96e-02	0.48656	6.27e-01
eutherm2	-3.97e-03	1.23e-02	-0.32182	7.48e-01
erhhch2	1.03e-02	3.67e-02	0.28080	7.79e-01
eufdsit3	-5.88e-03	4.57e-02	-0.12867	8.98e-01
euprpmel3	-1.78e-03	1.58e-02	-0.11216	9.11e-01
eumilk2	-3.09e-03	3.46e-02	-0.08936	9.29e-01

Table 29: Wald Statistics based on Bonferroni-significant predictor

	Estimates	SE	t_value	p_value
Intercept	3.36e+00	4.32e-02	77.74820	0.00e+00
Exercise Frequency	-7.95e-03	1.58e-03	-5.02743	5.31e-07
Household Income: Level 2	4.89e-02	9.98e-03	4.89342	1.05e-06
Beverage Consumption: Category 2	-2.98e-02	8.76e-03	-3.40546	6.71e-04

corresponds to a value of exp(3.36) = 28.78, which belongs to the overweight category.
The mean response of a person belonging to a household of Income > 185% of the poverty threshold and drinks beverages but also exercises can be written as:

$$\text{Group 1: } \mu = exp(3.36 - 0.00795 * freq)$$

If a person in this category exercises at least once then compared to the baseline category the person's BMI would decrease by exp(-0.000795) = 0.9.

The mean response of a person belonging to a household of Income <= 185% of the poverty threshold and drinks beverages but also exercises can be written as:

$$\text{Group 2: } \mu = exp(3.40 - 0.00795 * freq)$$

If a person belongs to a household of Income <= 185% of the poverty threshold, then the BMI would increase by exp(0.0489) = 1.05.
The mean response of a person belonging to a household of Income > 185% of the poverty threshold and does not drink beverages and also exercise can be written as:

$$\text{Group 3: } \mu = exp(3.33 - 0.00795 * freq)$$

If a person from baseline category stops drinking beverages then the BMI would drop by exp(-0.0298) = 0.97

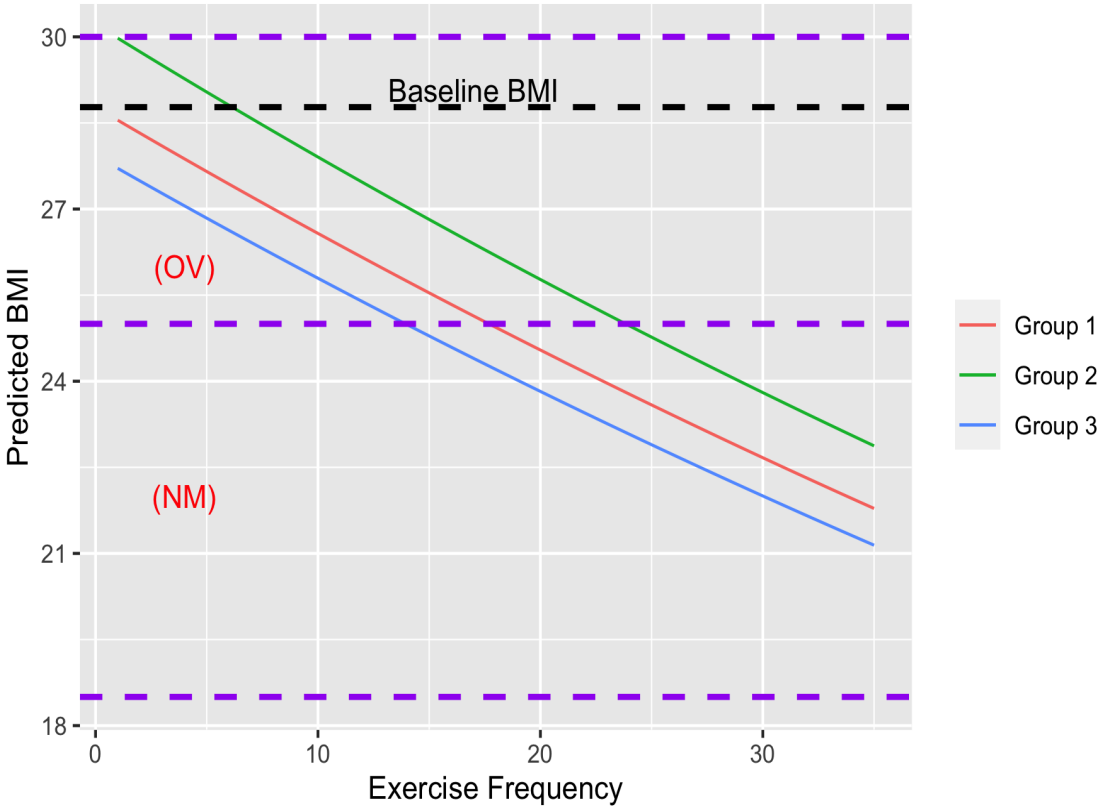

Figure 21: Plot of BMI vs Exercise Frequency

Since exercise frequency is the only continuous and significant variable, the plot has been made as a function of exercise frequency for better visualization. From Fig **??** we can see that there is a clear difference in BMI among different groups defined in the above discussion, notably people belonging to these groups need to exercise more frequently to stay in the normal BMI range according to the model prediction.

9 Regression analysis of factors associated with human population across countries

Introduction

Population plays a pivotal role in shaping the socio-economic landscape of nations, influencing resource utilization, economic development, and overall well-being. Understanding the factors contributing to variations in human population size is crucial for informed policy-making and sustainable planning. In this project, we conduct a regression analysis to explore the relationships between the human population of diverse countries and a set of key variables. Our chosen variables encompass critical aspects such as health, employment, and environment. Each variable represents a feature of a country's demographic, economic, and environmental profile. Through regression, we aim to understand the correlation between these factors and the size of human populations.

Methodology

Data Acquisition and Preprocessing

Data for more than 150 countries was acquired from the World Bank. The data consists of 4 independent variables and the total population as the dependent variable. The data consists of the average value of the indicators given in Table 30 from 1960-2020. The data from the World Bank is available in categories. The categories of interest for this project are Health, Education, and Environmental Factors. We chose to perform regression analysis with these variables because these variables are closely tied to population growth and had the least missing values of different countries out of more than 500 indicators.

Table 30: Description of Independent Variables

Independent Variables
Arable land (hectares per person)
Unemployment, total (% of total labor force) (national estimate)
Fertility rate, total (births per woman)
Fossil fuel energy consumption (% of total)
Population, total
Country Code

The rational behind choosing each variable is as follows:

- **Arable Land (hectares per person)**: The availability of arable land per person can significantly impact a country's ability to sustain its population. Countries with limited arable land may face food production challenges and need to rely on imports.
- **Unemployment, total (% of total labor force) (national estimate):** Unemployment is a crucial economic indicator affecting living standards. High unemployment rates may lead to social and economic issues, influencing population dynamics.
- **Fertility Rate, Total (births per woman):** Fertility rate is a key demographic indicator. It provides insights into a country's population growth and can impact its population size.
- **Fossil Fuel Energy Consumption (% of total):** Fossil fuel consumption measures a country's energy usage and can indicate its industrialization and development. It may also have environmental implications, influencing the overall quality of life.

Data Visualization

The histogram depicting the populations of over 150 countries in Fig 22 reveals an interesting pattern characterized by a heavy left-tailed distribution. Many countries exhibit smaller population sizes in this distribution, forming a cluster towards the lower end of the spectrum. This skewness suggests that most nations in our sample have relatively modest populations. The heavy left tail indicates that there are fewer countries with exceedingly large populations, contributing to the overall asymmetry of the distribution. Such a distributional pattern prompts us to dig deeper into the factors influencing population size, hinting at the prevalence of countries with limited demographic scales. Further analysis and exploration of the variables influencing population dynamics in this context can provide valuable insights into the disparities in global population distribution and inform strategies for addressing the unique challenges faced by countries with smaller populations.

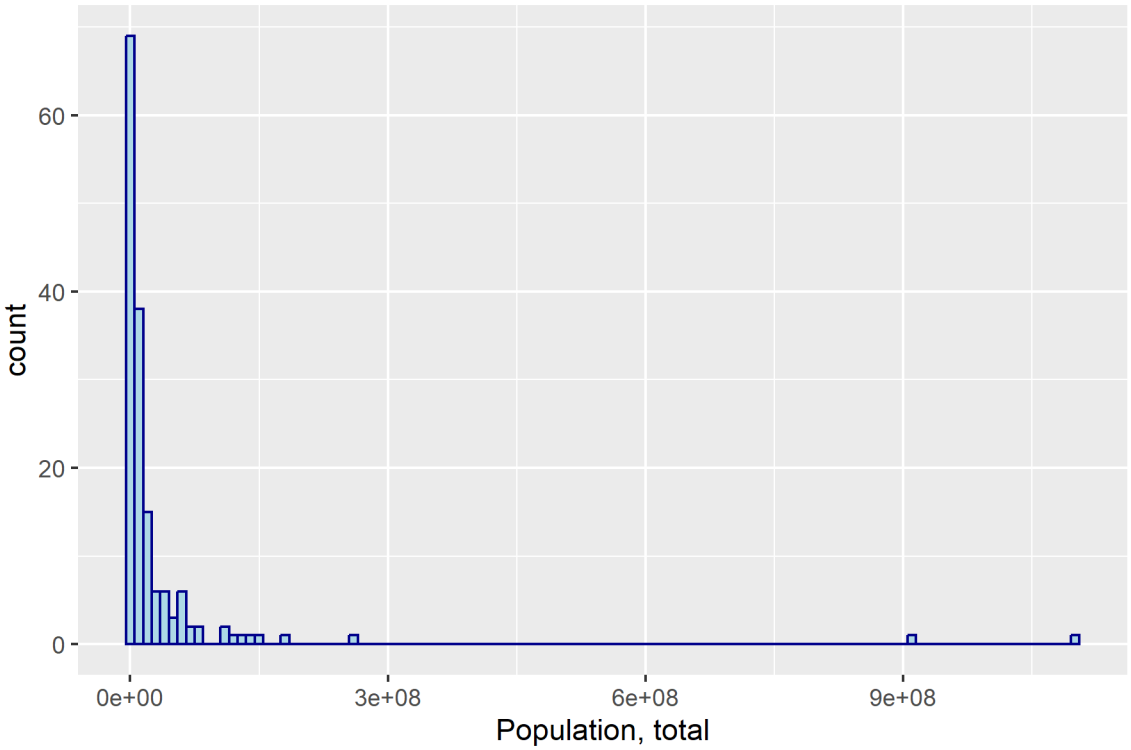

Figure 22: Population Histogram

Linear Regression Analysis

Before applying OLS to this data, we will state the required assumptions for the model:

- **Consistency (OLS1):**In the context of your regression model, consistency implies that the errors (u_i) in predicting population sizes are not related to the chosen independent variables (e.g., arable land, unemployment, fertility rate, etc.). In other words, the model assumes that the influence of unobserved factors on the population is not biased or correlated with the observed factors.
- **Full Rank (OLS2):** The full rank assumption is crucial for estimating the coefficients accurately. In your model, this assumption ensures that the matrix of independent variables

(X_i) has full rank, meaning that the chosen indicators collectively provide sufficient information to estimate the population sizes. In other words, there should not be perfect linear relationships among the chosen indicators.

- **Homoskedasticity (OLS3):** Homoskedasticity is essential in the context of your model to ensure that the variability in population sizes is consistent across different levels of the independent variables. For example, if the variance of errors is not constant, it could indicate that the model is better at predicting population sizes for certain ranges of independent variables and less accurate for others.

We will use the following regression formula for all the models in this study:

$$\begin{aligned}\log(\text{'Population, total'}) \sim\ & \text{'Arable land (hectares per person)'}+\\ & \text{'Unemployment, total (\% of total labor force)'}+\\ & \text{'Fertility rate, total (births per woman)'}+\\ & \text{'Fossil fuel energy consumption (\% of total)'}\end{aligned}$$

Table 32 contains the summary statistics of the regression model used with the logarithm of the population. All the independent variables have significant p-values (see Table 34).
Below we will comment the coefficient estimates based on percentage change in populations based on unit increase in independent varialbes and relate the model with real-world examples:

- **Arable land (hectares per person):** Percentage Change is given by, $100\times(\exp(1.394)-1)\approx 303.38\%$ For a one-unit increase in arable land per person, while controlling for other variables, the model predicts an approximately 303.38 % increase in the logged population. This suggests a substantial positive impact of increased arable land on population size. For example, countries with significant increases in arable land, such as India and China, have enormous populations.
- **Unemployment (% of total labor force):** Percentage Change is given by $100\times(\exp(-0.089381)-1)\approx -8.64\%$ For a one-unit increase in the percentage of unemployment while controlling for other variables, the model predicts an approximately 8.64% decrease in the logged population. This implies that higher unemployment rates are associated with decreased population size. Nordic countries, known for low unemployment rates, have seen relatively smaller population growth, reflecting the negative impact of higher unemployment on population size.
- **Fertility rate (births per woman):** Percentage Change is given by $100\times(\exp(0.183)-1)\approx 20.10\%$ For a one-unit increase in fertility rate while controlling for other variables, the model predicts an approximately 20.10% increase in the logged population. This suggests that higher fertility rates are associated with a significant increase in population size. Niger has one of the highest fertility rates globally. High fertility rates in some African countries can be attributed to factors such as limited access to family planning and cultural preferences for larger families.
- **Fossil fuel energy consumption (% of total):**Percentage Change is given by $100\times(exp(0.026)-1)\approx 2.65\%$. Arab nations rely heavily on fossil fuel energy, particularly in the form of oil, and this has fueled economic growth and migration to these countries leading to population growth.

Generalized Linear Model

The justification for using a Generalized Linear Model (GLM) with a quasi-family and a variance structure of $\text{Var}=\mu^3$ is closely tied to the observed histogram of the population distribution in Fig 22.

- **Addressing Skewed Population Distribution:** The heavy left-tailed distribution observed in the population histogram suggests that the data is not symmetrically distributed. A quasi-family allows flexibility in modeling the distribution of the response variable, and specifying a variance structure of Var $= \mu^3$ is particularly suitable for addressing the observed positive skewness. This choice acknowledges and accommodates the non-normal distribution of population sizes.
- **Improved Model Fit:** By choosing a quasi-family with a variance structure that reflects the observed features of the data, the GLM is better equipped to capture the inherent complexities of the population distribution. This can result in a more accurate representation of the relationship between the independent variables and the logged population sizes.

We define $\{(y_i, \boldsymbol{x}_{i*})\}_{i=1}^n$ as following a generalized linear model based on the exponential dispersion model (EDM), monotonic and differentiable link function g, and observation weights w_i if

$$y_i \overset{\text{ind}}{\sim} \text{EDM}\left(\mu_i, \phi_0/w_i\right), \quad \eta_i \equiv g\left(\mu_i\right) = o_i + x_{i*}^T\beta.$$

The offset terms o_i and observation weights w_i are both known in advance. The free parameters in a GLM are the coefficients β and, possibly, the parameter ϕ_0 controlling the dispersion.
The linear regression model is a special case of a GLM, with $\phi_0 = \sigma^2$ (unknown), $w_i = 1, o_i = 0$, and identity (canonical) link function:

$$y_i \overset{\text{ind}}{\sim} N\left(\mu_i, \sigma^2\right); \quad \eta_i = \mu_i = x_{i*}^T\beta.$$

For this dataset we will be using the quasi family of the generalized linear model, of the following property:

$$E[y_i \mid x_i] = \mu_i; \quad Var[y_i \mid x_i] = \mu_i^3;$$

Table 32 contains the summary statistics of the regression model used with the logarithm of the population. Based on these results we conclude the following (see Table 33 for more details):

- **Arable land (hectares per person):** The coefficient increased in the glm, suggesting a higher positive impact of Arable land on population size when using the quasi family.
- **Fertility rate (births per woman):** The coefficient increased in the glm, suggesting a higher positive impact of fertility rate on population size when using the quasi family.

Instrument Variable Estimation

In this section, we will argue for endogeneity in the model

- **Arable land** may be influenced by population density and economic factors. As population grows, there might be increased pressure on arable land for residential and industrial use. Economic development might impact land use, with shifts from agriculture to urbanization. Factors such as population density and agricultural land are directly tied to arable land per person, therefore we believe the estimate of arable land is biased and underestimated.
- **Fertility rate, total (births per woman):** Fertility rates can be influenced by social, and cultural factors. Factors such as mortality rate and contraceptive prevalence have been known to have direct impact in fertility rate and therefore we believe OLS estimate is biased.

Table 31 contains a list of instrument variables. Below argue that each of these is a valid exogenous variable:

- **Agricultural land (%of land area):** This variable is introduced as it captures the proportion of land area used for agriculture. In contexts where arable land is endogenous, agricultural land as a percentage of total land area provides an additional exogenous measure that may influence population size independently.
- **Population density (people per sq. km of land area):** Population density is introduced to consider the concentration of people in a given land area. It provides an exogenous measure that can influence population sizes independently of arable land and fertility rate.
- **Contraceptive prevalence, any method (%of married women ages 15-49):** This variable is included to represent the prevalence of contraceptive use among married women. It introduces a demographic control that can affect fertility rates, thus serving as an exogenous factor in explaining population sizes.
- **Mortality Rate, Infant (per 1,000 live births):**Infant mortality rate is a critical health indicator. It reflects the healthcare system's effectiveness and the overall well-being of a population. This variable can directly impact the fertility rate and is included as exogeneous variable.

Table 31: Description of instrument Variables

IV Varialbes
Agricultural land (% of land area)
Population density (people per sq. km of land area)
Mortality rate, infant (per 1,000 live births)
Contraceptive prevalence, any method (% of married women ages 15-49)

Table 32 contains the summary statistics of the 2-stage regression model used with the logarithm of the population. Below are comments on use of IV variables (see Table 35 for more details):

- **Arable land (hectares per person):** The coefficient estimate for arable land in the IV estimation (ivreg) is substantially higher compared to both the linear and generalized linear models. This suggests that the instrumental variable approach, which addresses endogeneity issues, results in a larger estimated impact of arable land on population size.
- **Fertility rate (births per woman):** The coefficient has increased by including mortality rate and contraceptive prevalence as IVs, indicating a larger impact on population size compared to the lm and glm. This means that LM and GLM estimates were biased downwards also we can see that the p-value has decreased making the fertility rate a more pronounced factor in population dynamics.

Model Inference

Fig 23 and 19 contain diagnostic plots for the linear model and generalized linear model with quasi-family, please refer to the figure to see the summary. Table 33, 34 and 35 contain model coefficients and heteroskedastic robust standard errors. Based on the confidence intervals we can see that LM and GLM yield similar results, but GLM has wider intervals. IV estimation introduces wider intervals, highlighting potential bias reduction but also increased uncertainty.

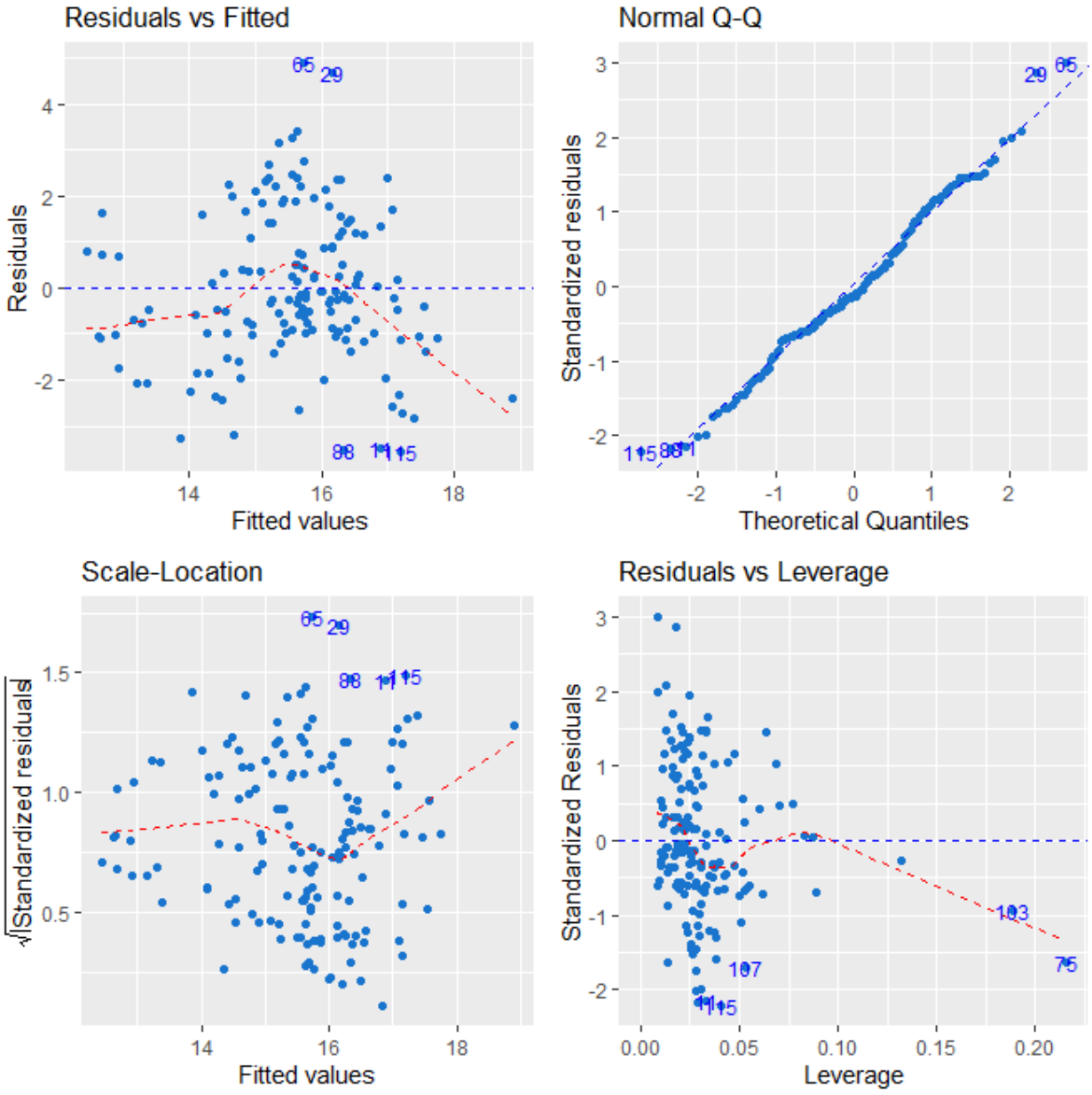

Figure 23: **Diagnostic plots for linear model**: The residual vs fitted plot suggests there is some heteroskedasticity in the data, the Normal Q-Q plot suggests that the model has done reasonably well to establish the correlation between various variables, The standardized residual vs fitted plot confirms there is some heteroskedasticity in the data and the residual vs leverage plot suggests that countries like Kazakhstan, China, India, and Namibia influence the estimation of parameters.

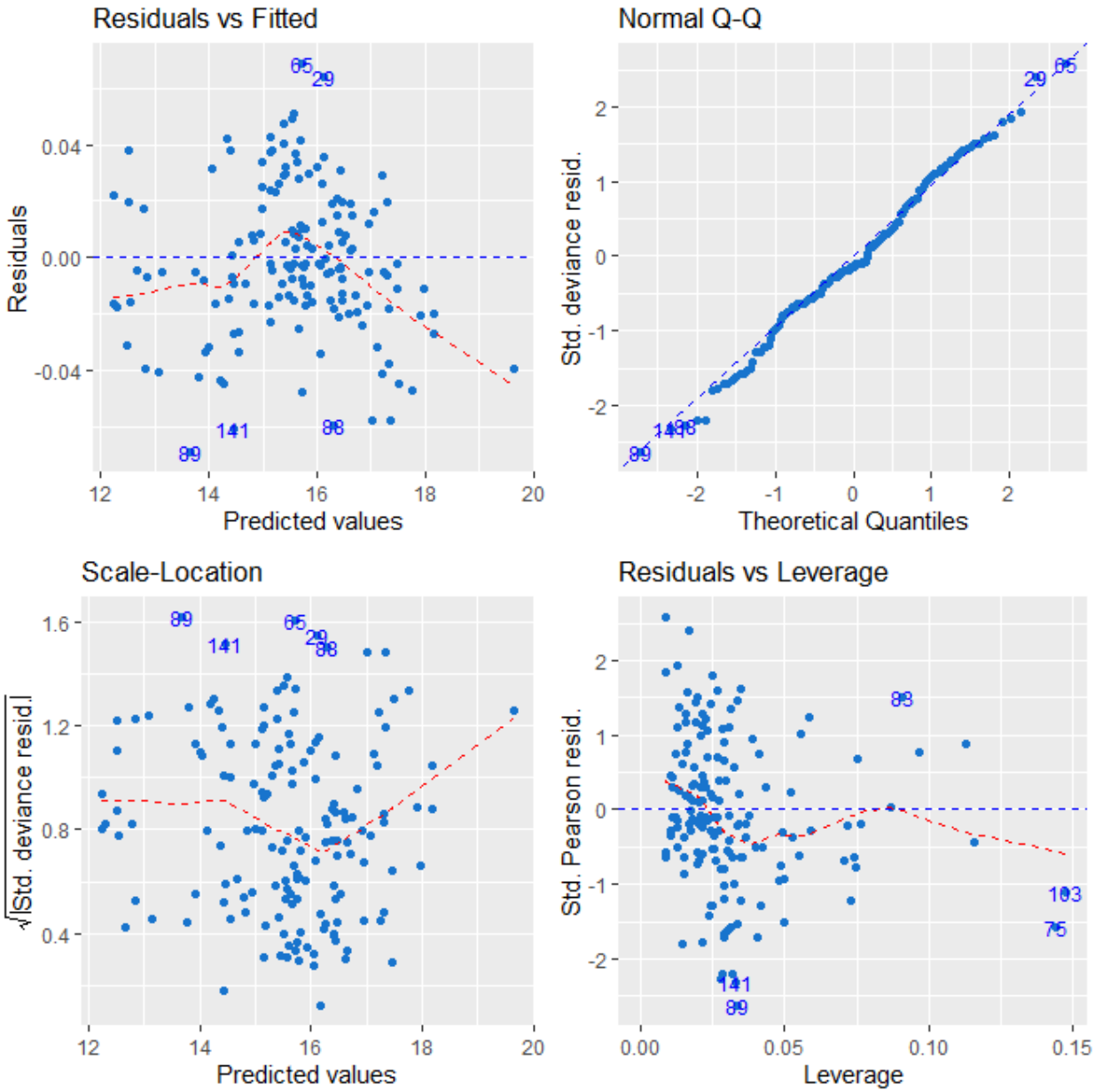

Figure 24: **Diagnostic plots for quasi GLM**, The residual vs fitted plot suggests there is some heteroskedasticity in the data also the quasi glm has reduced residual magnitudes compared to the linear model, the Normal Q-Q plot suggests that the model has done reasonably well to establish the correlation between various variables, The standardized residual vs fitted plot confirms there is some heteroskedasticity in the data and the residual vs leverage plot suggests that countries like Kazakhstan, Saudi, Niger influence the estimation of parameters.

Table 32: Summary Statistics of coefficient estimates using LM, GLM and IV Estimation

	LM	GLM	IV
(Intercept)	13.792***	13.088***	11.795***
	(0.611)	(0.595)	(0.944)
‘Arable land (hectares per person)‘	1.394**	1.770**	6.134**
	(0.474)	(0.550)	(2.084)
‘Unemployment, total (% of total labor force)‘	-0.089***	-0.087***	-0.080**
	(0.023)	(0.020)	(0.029)
‘Fertility rate, total (births per woman)‘	0.183•	0.266**	0.362*
	(0.098)	(0.098)	(0.150)
‘Fossil fuel energy consumption (% of total)‘	0.026***	0.031***	0.025***
	(0.004)	(0.004)	(0.006)

Table 33: Heteroskedastic robust standard errors and confidence intervals for GLM

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	13.088	0.516	25.339	0.000	12.075	14.100
‘Arable land‘	1.770	0.539	3.285	0.001	0.714	2.826
‘Unemployment, total‘	-0.087	0.020	-4.354	0.000	-0.126	-0.048
‘Fertility rate, total‘	0.266	0.082	3.237	0.001	0.105	0.427
‘Fossil fuel‘	0.031	0.004	8.274	0.000	0.024	0.039

Table 34: Heteroskedastic robust standard errors and confidence intervals for linear model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	13.792	0.532	25.932	0.000	12.741	14.843
‘Arable land‘	1.394	0.502	2.779	0.006	0.403	2.385
‘Unemployment, total‘	-0.089	0.020	-4.395	0.000	-0.130	-0.049
‘Fertility rate‘	0.183	0.082	2.242	0.026	0.022	0.345
‘Fossil fuel‘	0.026	0.004	6.466	0.000	0.018	0.034

Table 35: Heteroskedastic robust standard errors and confidence intervals for IV estimation

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	11.795	0.785	15.026	0.000	10.244	13.346
‘Arable land‘	6.134	1.810	3.388	0.001	2.557	9.711
‘Unemployment, total‘	-0.080	0.024	-3.401	0.001	-0.126	-0.034
‘Fertility rate, total‘	0.362	0.143	2.529	0.012	0.079	0.645
‘Fossil fuel‘	0.025	0.004	5.547	0.000	0.016	0.034

Limitations

In the dataset, variables like arable land and fertility rate might be endogenous, as they can be influenced by unobserved factors such as cultural practices, economic policies, or technological

changes. Omitting these factors might lead to biased OLS estimates (we tried to eliminate this bias using IV estimation). The dataset may lack certain important variables that influence both the dependent variable (population) and the independent variables. For example, factors like government policies, and education levels, potentially give biased coefficients. The accuracy of measurements for variables such as arable land, fertility rate, and others is critical. If there are measurement errors in the dataset, OLS estimates may not precisely reflect the true relationships, impacting the validity of the results. The dataset contains variables that exhibit temporal autocorrelation, such as population dynamics where one time period may influence subsequent periods and OLS assumptions are violated, but since we deal with averaged data, we are completely ignoring this aspect. The dataset may contain outliers or influential observations that disproportionately affect OLS estimates. The assumption of a linear relationship between independent and dependent variables may not hold in all cases. For instance, the impact of arable land on population size may not be strictly linear, and a non-linear functional form might be more appropriate.

Conclusion

In this study, we undertook a comprehensive analysis of factors influencing the population of various countries, employing a regression framework with a focus on variables such as arable land, unemployment, fertility rate, and fossil fuel energy consumption. The Ordinary Least Squares (OLS) regression served as the primary modeling technique. The inclusion of diagnostic plots and alternative model specifications, such as a Generalized Linear Model (GLM) and instrumental variable (IV) estimation, allowed for an in-depth exploration of the relationships. While OLS provided valuable insights, limitations such as potential endogeneity, omitted variable bias, and the assumption of linearity were recognized. Diagnostic plots were used in assessing the validity of model assumptions and identifying influential observations.