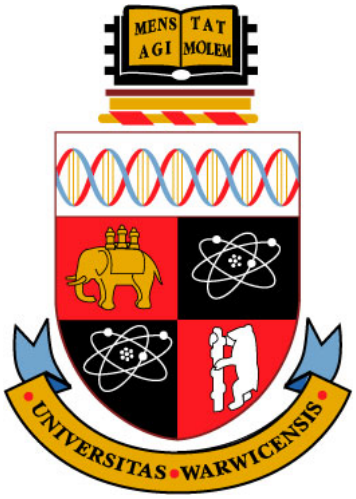

Effects of the COVID-19 Pandemic on Learner Drivers

ST421 Data Science Project

David Ferreira

Supervisor: Prof. Jane Hutton

Departments of Statistics and Computer Science

University of Warwick

Year of Study: 2021/2022

Abstract

The COVID-19 pandemic was an unprecedented phenomenon that affected a large number of aspects of our daily lives. One of these factors is learning to drive. The lockdown measures in the United Kingdom made it more difficult for young adults to complete their driving lessons. More specifically, the proportion of hours spent learning to drive with instructors relative to learning to drive with mentors or parents decreased by 64% during the lockdowns and 15% after. This means that the quality of lessons being received by young adults has lowered. Additionally, we found that after lockdowns ended the number of hours required to pass the practical exam increased by 12%. These findings indicate that the COVID-19 lockdowns have negatively impacted the learning to drive process. These figures were found using Statistical Modelling techniques on a dataset of 15,748 learner drivers. We utilised Generalised Statistical Models, specifically, a Negative Binomial model for the total number of hours and a Gamma Model for the proportion of hours.

Keywords

Data Analysis
Transport
Driving
Statistical Modelling
Generalised Linear Models

Acknowledgements

I would like to thank my supervisor, Professor Jane Hutton, for spending time every week offering guidance and assistance in completing this project. Additionally, the two third years, Vanessa Rodriguez and Dodzia Daraz, who undertook similar projects, and helped in the data cleaning and exploration stages of the project.

In addition, Sritika Chowdhury, Jill Weekley and Shaun Helman, who work with TRL, worked together with us and provided assistance by comparing findings and results to help us work in the right direction. Finally, they, as well as Prof. Hutton, listened to and provided feedback for our presentations.

Finally, I would like to thank my friends and family for supporting me throughout the year.

Disclaimer

This project was completed in collaboration with TRL, thus the results from this project are from the provisional Driver2020 data and should not be circulated. The results should also not be taken as official results from the study.

Contents

Chapter 1

Introduction

This project aims to use data analysis to evaluate how learner drivers' experiences were affected by the COVID-19 lockdowns. We will look at how the data was acquired, processed and analysed. This project was done with help from TRL, or the Transport Research Laboratory. They handled data collection entirely and did their own analysis in parallel with us.

Driver2020 is the name of the project headed by TRL. The question their project initially sought to answer was, how effective certain interventions were in the safety record of novice drivers in their first year of driving. To answer this question a prospective, randomised controlled [1], longitudinal study [2] took place that invited learner and novice drivers to participate in a year long trial, in which they were randomly assigned these interventions and were monitored periodically through questionnaires that were sent out.

The data collection has since concluded and the data acquired through these surveys was made accessible for use in this project, with some participant data censored for privacy, in order to answer questions similar to the one posed by TRL.

The study was proposed before the COVID-19 pandemic had begun, however, when it began, so did the large-scale lockdowns that took place in the UK. Hypothesising that these measures would have an impact on learner drivers, TRL added some additional questions into their surveys to record the effects of lockdowns on participants. These unprecedented events meant that there were more questions that could be looked into with the data that was available, thus the question this project aims to answer was created.

1.1 Background

The Driver2020 project builds off of another study, "A review of interventions which seek to increase the safety of young and novice drivers" (Pressley et al, 2017) [3]. What is meant by interventions is futher elaborated in Section 1.1.1.

1.1.1 Interventions

The Driver2020 project, as well as the preceding study [3] mention several interventions that had the objective of increasing safety for novice drivers. The first 4 were recommended for evaluation by the study, while the fifth was added for this project as it is similar to the intervention the government uses across Great Britain currently.

1. Mentor Agreement: a set of materials designed to assist parents and other mentors in setting limitations on post-test driving for newly qualified drivers, to manage access to high risk driving situations such as driving at night, driving in bad weather, and carrying similar-age passengers.
2. Logbook: a set of materials designed to assist learner drivers in achieving greater amounts of on-road practice before passing their practical driving test.
3. Telematics: an app-based intervention to provide feedback to newly-qualified drivers on their driving style, mimicking as closely as possible telematics-based insurance products.
4. Hazard Perception Training: an e-learning course designed to increase hazard perception skill (delivered pre-test and post-test).
5. Education: a classroom session with follow-up support designed to target a range of attitudes and behaviours known to be associated with post-test risk.

1.1.2 COVID-19 Lockdowns

The UK lockdowns enacted many restrictions on citizens. These included travel restrictions, self-isolation, mask mandates and so on [4]. In addition to this, there were multiple lockdowns, during which driving lessons and tests were suspended. Table 1.1 lists the start and end dates of when these were suspended. Note, these are not necessarily the start and end dates of the lockdowns themselves.

Restriction	Start Date	End Date
England		
First	2020-03-19	2020-07-22
Second	2020-11-05	2020-12-02
Third	2021-01-05	2021-04-22
Scotland		
First	2020-03-20	2020-09-14
Second	2020-12-26	2021-05-06
Wales		
First	2020-03-20	2020-08-17
Second	2020-10-24	2020-11-09
Third	2021-12-20	2021-04-22

Table 1.1: Dates for restrictions on learning to drive

The dates are quite different for each country, however, due to the lack of participants from Scotland and Wales, we will focus more on England when referring to lockdowns [5].

1.1.3 Learning to Drive

There is a specific process for learning to drive in the UK.

1. In most cases you need to be 17. However, this study includes 425 participants who are 16, as there are some special cases.
2. You must apply for a provisional license, which costs around £40.
3. The government helps you find driving instructors who are approved by the Driver and Vehicle Standards Agency. There is no minimum number of lessons you must take and you are allowed to practise with family or friends who qualify.
4. You must then take the driving theory test, which consists of multiple choice questions about the highway code, traffic signs and driving practises; and a hazards perception test. Booking this test costs £23.
5. You then take your practical test, which costs around £70. The test takes around 40 minutes and involves testing what you've learnt in your lessons.

Knowing this process helps us better understand the problem we are tackling. For example, knowing that there is no minimum number of lessons, we can say that the number of hours a participant is essentially a measure of how quickly they are able to learn to drive.

1.2 Surveys

Participants in the study were sent several surveys to fill out at different points in time. The study followed the participants for up to 12 months after registration. Since participation was not mandatory some participants did not respond, as the only incentive to answering them was a £5 voucher (per survey). This resulted in a large drop off in collected data, which increased as time passed. There were also follow-up calls for those who did not respond, but the only information collected was collision-related if they answered.

This included one when they registered for the study, containing questions related to personal details and characteristics. Then one when they passed their driving test, with questions related to their learning process and their driving style. Finally, in intervals of 3, 6 and 12 months participants were asked to fill identical surveys that were similar to the test pass survey but included more details, such as miles driven and details on any accidents.

1.2.1 Registration

The registration survey was conducted when participants applied to partake in the study. However, due to the nature of the study participants were recruited at different stages in their driving timelines. Some interventions required participants that had yet to take their driving test and some which had already taken it. Incidentally, it would be possible to recruit everyone in the former category, but this would require waiting until they had passed their test to measure post-test interventions, which would take 6-12 months. In that time road safety measures could have developed, which would introduce problems when comparing the two groups. This resulted in two participant groups, which are labelled in the data as novices and learners.

Novices were recruited after their driving test took place, while learners were recruited as they were beginning to learn to drive. They were still given the same survey, however. Most of this survey is irrelevant to our project, as it contains participants' personal details, which are unable to be disclosed. The only 2 factors we have from this are age and gender, both of which are used quite often in our analysis.

1.2.2 Test Pass

The test pass survey was conducted when participants passed their driving exam. For novices this was immediately after registration, while for learners it was whenever they passed. Since we are looking at effects on learner drivers, this survey is the most important

and the majority of data analysis will involve the results acquired from it.

1.2.2.1 Time-Variable Questions

Firstly, this survey, as well as the post-test surveys, contain questions related to factors that can change over time. This includes education, employment, access to a vehicle and insurance. Below is the structure of the first two, while the remaining two are simple 'yes or no' options.

1. Are you in full-time education?

   If yes: Projected qualification

   If no: Highest past qualification

   - University Higher Degree (MSc; PhD) or Chartered status
   - First degree level qualification (BA; BSc; PGCE)
   - Higher education diploma (HNC, HND, Nursing or Teaching qualification)
   - A Level; AS Level; NVQ Level 3; GNVQ Advanced or equivalent
   - GCSE; CSE, NVQ Levels 1&2; GNVQ Foundation & Intermediate or equivalent
   - None of the above

2. Which of the following best describes your employment status?

   - Full-time
   - Part-time
   - Self-employed
   - Unemployed

1.2.2.2 Driving Style Questions

Secondly, this survey also contains questions related to participants' subjective evaluations of their own driving styles. This takes the form of 12 questions, grouped into blocks of 3, which TRL has determined to be similar. Each has a value scale ranging from 1 to 7. The following are the aspects of participants' driving styles that are evaluated:

Attentiveness	Experience	Patience	Selfishness
Carefulness	Irritability	Responsibility	Speed
Decisiveness	Nervousness	Safety	Tolerance

Thirdly, this survey contains questions related to the learning to drive process.

1. Number of hours spent learning to drive with and without a driving instructor
   Scale from 0 to 150 hours for each
2. Proportion of learning mileage driven in different locations
   - Residential areas
   - Busy towns or cities
   - Country roads/lanes
   - Dual carriageways
   - Motorways
3. Proportion of learning mileage driven with abnormal circumstances
   - With additional passengers
   - In the dark
   - On wet roads

1.2.2.3 Lockdown Impact Question

Finally, the survey contains a question that was added to the survey after the COVID-19 lockdowns began to take effect. It asked participants whether they had been impacted in certain ways by the pandemic. The question had the 10 following options:

1. No impact

2/3. Practical/Theory test was postponed or cancelled

4/5. Unable to book practical/theory test when ready

6/7. Unable to get intended driving experience with/without driving instructor

8/9. Able to get more than intended driving experience with/without driving instructor

10. Able to get intended driving experience but it took longer

1.2.3 Post-test

There were 3 post test surveys that participants could fill out at 3, 6 and 12 months after their test pass date. The primary aim of these were to record any collisions and details on how they happened. They also recorded other factors that were present in the test pass survey. They also record factors that may have affected the number of collisions, including, miles driven, propensity to speed or drive while tired, and more. There are also questions that relate to the interventions that were randomly assigned to participants.

Chapter 2

Methodology

We want to understand the data that we have acquired, so the first step is to select appropriate outcome and response variables, while ignoring those that are not useful. Once this is done we visualise the data to see how the variables change with time and interact with each other.

2.1 Data Cleaning

The data we receive as an output from the surveys are not ready to be analysed in R. We need to reformat it before doing anything else. Then, we check for inconsistencies in the data, usually caused by human error by participants filling out the survey.

2.1.1 Formatting

We were sent the outputs of the surveys, which were saved as '.xlsx' files, also known as 'Microsoft Excel Open XML Spreadsheet'. While this is not as preferable as '.csv' files, R has tools that allow us to read these files into tables that we can then manipulate. However, there were several formatting issues that needed to be fixed before any analysis is done.

	Q23. Appr	Q24. Thinking about all the driving you have done ir				Q25. Think
		Q24.1. In r	Q24.2. In b	Q24.3. On	Q24.4. On	Q24.5. On motorways	
6424	10	5	30	30	25	17
3338	40	10	25	20	5	33
2115	20	10	30	20	20	70
6471	15	5	30	25	25	5
10000	20	20	20	30	10	30

Figure 2.1: Example of two rows of columns in the spreadsheet

The first issue is that several questions in the survey had sub-questions that resulted in two rows being used in the spreadsheet for the same question, as seen in Figure 2.1. While this format is easy to interpret by eye, R is unable to pull columns from more than one row when converting the spreadsheet into a table. Figure 2.2 shows an example of how R reads data with 2 rows of column headings. This is a problem, because R requires every entry in a column to be numeric, for the column to be labelled as numeric. Without this, we are unable to perform numeric operations, such as summations or average calculations, on these columns in the dataset. So, because the second row of questions is treated as a row of data, this needs to be fixed before we move onto analysis.

Q24.1 <fctr>	Q24.2 <fctr>	Q24.3 <fctr>
Q24.1. In residential areas	Q24.2. In busy towns or cities	Q24.3. On country roads/lanes
10	5	30

Figure 2.2: Example of two rows of columns in R

We fix this issue by writing an algorithm to merge the two rows into one. This algorithm favours sub-questions, so Q24 is replaced by Q24.1 and so on. This results in a single row which maps exactly to the original two, so R can now create a table with the correct columns.

The second issue is that each column name contains the full question from the survey. While this may be useful for checking which question a column refers to - it means that referencing columns in the code would be difficult. Below is an example of a line of code in R that references question 30 from the survey.

```
select("Q30. When driving in the last 6 months, how often have each of the
following things happened to you?")
```

As can be seen this is too long to reference. This can be fixed by renaming the columns more appropriately. It is possible to do this manually, however, we have elected to use an algorithm to cut everything apart from the question number from the column name. This way it is short and the question can be referenced quite easily. A key will be included so that anyone not familiar with the questions is able to interpret the code. Below is an example of a line of code referencing a column name that has been shortened.

```
select(Q30)
```

The data is now successfully converted to a table, however, the data types of each column are not correct. This is one of the issues of using '.xlsx' files, every column is set to 'character' by default. So we change any containing to 'numeric' type, those containing choices become 'factor' and finally those with words are left as 'character'. The column type is important when creating graphs, calculating averages and modelling.

2.1.2 Consistency

One question in the survey resulted in outputs that aren't consistent and are therefore not easily interpreted when being analysed.

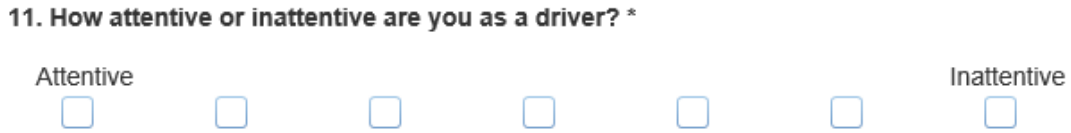

Figure 2.3: Example of scale in driving style questions

These are the driving style questions, as seen in Figure 2.3, which ask the user to rate aspects of their style on a scale from 1 to 7. In the survey the extremes of the scale are labelled with what that value implies rather than the value itself, otherwise the scale would be unclear. This has resulted in the output of this question labelling 1s and 7s with their respective character label, meaning we cannot do any numeric analysis without changing it. Since each question has its own labels and there is no pattern between them, we are required to change these values manually.

Another issue with these questions is that when they are displayed in a graph or analysis, without manual labels it is unclear what a 1 or a 7 refers to. Since every question seems to have a positive extreme and a negative extreme of each trait, we have chosen to denote a 1 as the positive and vice versa. This required reversing the values of traits where 1 was the negative extreme. The calculation was simple: $8 - value$.

Q11. How	Q12. How	Q13. How
Q11.1.	Q12.1.	Q13.1.
Attentive	Careful	2
Attentive	2	Decisive

(a) Example of character labels

	Q11.1	Q12.1	Q13.1	
)	1	1	2	
)	1	2	1	

(b) Example of fixed labels

2.2 Exploratory Data Analysis

Once the survey data is cleaned and prepared, we need to select and then understand the important variables and how they interact with each other. We can understand them by visualised how they are distributed through graphs and tables.

2.2.1 Variable Selection

Before we explore our variables, we have around 50 to choose from, so we need to narrow down which ones are useful. We also need to decide on dependent and independent variables for the modelling process.

2.2.1.1 Dependent Variable

We want to find the effect of lockdowns on learner drivers. There are many ways to measure an effect, however, the questions on the test-pass survey were created to answer TRL's question related to the number of collisions post-test, so the information we have on the learning-to-drive process is very limited.

There are only 2 good candidates for our dependent variable, questions 4 and 5. They ask about the hours spent learning to drive with and without a driving instructor, respectively.

The only other candidate that was considered is question 23, confidence in driving ability. However, there is no objective way to measure confidence, whereas hours spent learning is a time measure. Since it is subjective, two participants could have the same level of confidence, but answer differently, meaning the measure is not accurate.

The optimal dependent variable would be the proportion of candidates who failed their driving test compared pre and post-lockdowns. However, we only have access to the candidates who succeed.

We can only compare changes in number of hours and we need to ascertain whether an increase is positive or negative. Intuitively, an increase is positive as it allows the candidate more time to study and thus increases their likelihood to pass. However, since we are only looking at those who succeeded, a decrease in number of hours is the positive change. This is because, we do not have access to the number of failures before passing, so we can only infer that someone with less hours was able to learn in a shorter time than others. This is not a definitive measure, unlike the one mentioned above, a participant with a given number of hours could have potentially passed their test with less, there are also factors of luck involved, however, it is the best measure available to us.

To summarise, our dependent variables are the number of hours spent learning to drive with and without an instructor and we have determined that a decrease in this value is positive change.

Limitations The largest issue with our dependent variables is that they are a measure of time, however, we only have an end date. The lack of start date means that we

are unable to produce a time interval to check whether a participant had their learning interrupted by a lockdown. Luckily, we have Question 10, described in Section 1.2.2.3, however, this only measures a self-described impact. Ideally, we would want a more objective measure, to verify the subjective one. For an applied explanation of this example, refer to Section 2.2.4.1.

We could estimate the start date with some degree of accuracy, by using the date that pre-test participants were recruited for study. As mentioned in Section 1.2.1, pre-test participants were recruited when they were beginning to learn to drive. This is an estimate, because that date is not guaranteed to be the date when they began to learn. Additionally, the number of pre-test participants is only around 30% of the total cohort, meaning this estimate is likely to not be useful when looking at overall trends.

2.2.1.2 Independent Variables

We want to determine whether lockdowns had an effect on learning hours, so we need to measure the difference in people affected by lockdowns and those that were not and verify that this difference is significant when compared to differences in other factors.

Firstly, we need to determine these other factors. The simple ones to include are differences in demographics, however, we only have access to some of these, age, gender, employment, education and location. Apart from these, we have factors from the survey which include, driving style, proportion of road types and proportion of additional factors. Some of these factors are likely insignificant in affecting learning hours.

Secondly, there are several ways to measure the effect of lockdowns. We could compare participants that started to learn to drive before and after lockdowns. This is a simple and broad estimate, so it does not take into account that there are some people who were unaffected by lockdown procedures. A better measure is one of the questions that was added to the survey by TRL. This is question 10, which asks candidates whether the COVID-19 pandemic had any impacts on their learning process. We will be using this question a lot, as it is the best variable we have for lockdown effects and it is further discussed in Section 1.2.2.3.

2.2.2 Variable Creation

While we have many variables to choose from, some factors can be modified to create new variables that allow us to access more information.

2.2.2.1 Pre-covid and Post-covid

We have access to the dates on which participants passed their driving theory tests and we know the dates on which lockdowns began in England [4]. We can also acquire the dates for Wales and Scotland, however, due to the low number of participants from those two places, for simplicity we will focus mainly on England moving forwards. Additionally, later on we find that a participant's country has little impact on their learning hours, so this decision is acceptable.

We do not know when participants began to learn to drive, however, we do know when they finished. Therefore, we can create an indicator that tells us whether participants learned before or after lockdowns began. By checking whether participants' test pass date is before the first lockdown date, we create an indicator variable called 'pre-covid' which we can use later on.

Additionally, we can do the inverse and note when participants pass their test after lockdowns. This measure is not as good as pre-covid, as we don't have a start date for learning to drive. Thus this variable is left censored.

2.2.3 Summary Statistics

First, we look at a summary of our dependant variables.

Learning Hours	**Total**	**Mean**	**Variance**	**Median**
With Instructor	615,068	42.44	953.56	35
Without Instructor	435,950	30.08	1448.87	16

Table 2.1: Summary statistics of learning hours

As we can see from Table 2.1, hours with an instructor tends to be higher on average, both in mean and median, with a larger total. We can see that the participants spent more time learning to drive with their instructors, rather than with other mentors. We want to figure out whether the pandemic has had any impact on these two numbers, whether it increased their disparity, or simply caused an increase or decrease independent of each other.

We can hypothesize that the pandemic would have caused an overall increase in both variables and also a shift in disparity. Specifically, that the increase in disturbances from lockdowns would have caused an increase in stress and time between lessons, that lead to an increase in learning hours. By shift in disparity, we mean that the proportion of hours without an instructor increased, due to a lack of availability of instructors, for example.

We will see in Section 3 whether these hypothesises are correct.

Then, we look at the distribution of participants for most factors in our data. In Table A.1 we can see the number of participants that are part of certain groups. We have omitted any participants with missing data, reducing the total observations from 15,748 to 14,436. This table is located in Appendix A due to its large size.

We will discuss how each factor is distributed,

- **Gender**: There is a majority of participants who are female. Ideally the proportion would be equal, but that is not the case here. It is not enough to make one of the groups insignificant, however. We can see that the female group tends to have higher values for number of hours. This means we want to include this in our model, to remove the confounding aspect.

- **Age**: As can be seen there is a large majority of participants in the 16-18 group. Additionally, most of these are found in ages 17 and 18, while 16 has the lowest proportion of any age. Looking at the mean hours we can see that 19-21 and 22-24 are very have very similar means, as opposed to 16-18 which is quite different. We could rearrange these groups to be more suitable.

- **Country**: A large majority of participants are from England. Due to the fact that lockdowns are different for each country, for visualisations and discussions involving time-frames and lockdowns, we will focus on England, due its majority. Additionally, it seems that the mean hours for every country are very similar. This means that this variable will likely be insignificant for the most part.

- **Time-frame**: This variable measures the time-frame of when participants finished their practical test. Note that, in the modelling, this variable is split into 2. We will explore these means further in the section with Table 2.2.

- **Lockdown Impacts**: These are the most important variables, as they measure the impacts of the lockdown measures as described by participants. We can see that some of these have little effect, while, for example, impact 2 increases the mean overall, as does impact 6. These impacts will be covered better in later sections, but currently it is difficult to ascertain anything due to the confounders.

- **Full-time Education**: We can see that this variable seems to have a large impact on both hour count categories. The proportion is also not skewed enough to make the results insignificant, therefore, it is likely to be useful in our model.

2.2.4 Visualisation

2.2.4.1 Mean learning hours over time

We now want to visualise how our dependent variable changes over time. Figure 2.5 shows how the average total number of hours spent learning changes over time in England. We have chosen to just look at England, due to the large proportion of participants from there. Each point on the graph is the average for a given date. Additionally, we can see the lockdown dates start and end dates displayed. The purple line shows the average number of hours for before the lockdown.

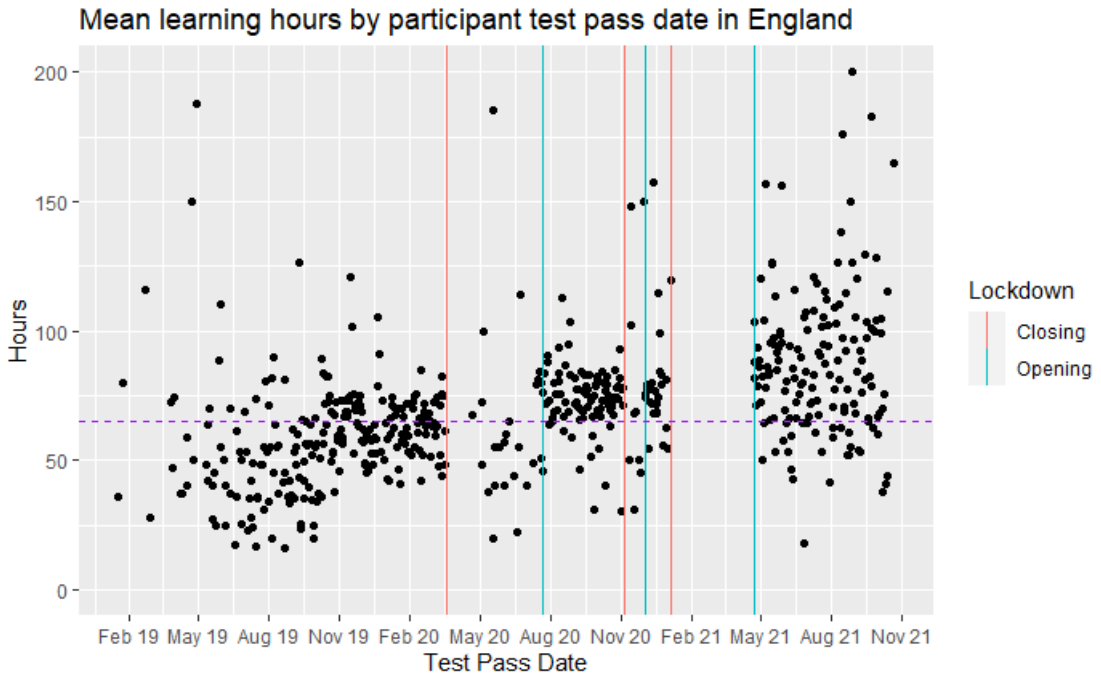

Figure 2.5: Mean learning hours by test pass date in England

We can see that there are clusters of points between and outside lockdowns. This is because most learner drivers were unable to take their test during lockdowns. This is discussed more in Section 1.1.2. It appears that, the average number of hours increases as time passes. We know this because the majority of the points for each cluster after lockdowns begin, is above the purple line. This supports our earlier hypothesis that the total number of learning has increased due to lockdowns. However, this is not enough to finalise the claim, since there is a possibility of confounding variables. This graph is also useful for observing the variance of our dependent variables. It seems that the time frame with the highest variance is after lockdown 3. The graphs for Scotland and Wales have similar trends, however the variance for them is much higher, as well as the number of observations being much less than England's, so any information taken from them is unreliable, so they are not included.

Additionally, we should acknowledge a limitation of our dependent variable, which is that we do not have a time interval for a participant's learning process. We have the number of hours and an end date, but no start date. This issue can be explained by example; if we have two participants, one with test-pass date in November 2020 and one in August 2020. Both of these times are between lockdowns 1 and 2, however, one is near the end and the other at the beginning. We do not know their start dates, and the number of hours is not enough to predict this. Say, they both have 75 total hours spent learning, we cannot make any conclusions about their start date, since lessons are often cancelled, delayed or skipped, randomly. Thus, it is entirely possible for their start date to be the same, meaning they were equally affected by lockdowns. This limitation is further discussed in Section 2.2.1.1.

We can see the general outline of the mean number of hours changes over time, but we are unable to separate the number of hours by with, and without instructor on the same graph. We could include multiple graphs, but a table allows us to record the exact averages for comparison. Table 2.2 shows the average hours spent learning per time frame, separated by with instructor, without and total number. These time frames do not include the times within lockdowns, as those are special cases.

	Hours			**Proportion**
Time Frame	**With Inst.**	**Without Inst.**	**Total**	**(With/Without)**
Before Lockdowns	43.26	21.62	64.88	2.0
Between Lockdown 1 and 2	41.57	34.43	76.0	1.21
Between Lockdown 2 and 3	38.60	37.81	76.41	1.02
After Lockdowns 3	48.79	36.61	85.40	1.33

Table 2.2: Mean number of learning hours per time frame in England

We can see that the average total number of hours correspond to what we observed in Figure 2.5. But now we can also see the values separated by instructor and their proportions. The general trend appears to be that 'with instructor' decreased during lockdowns, but then after it increased to higher than its original value. 'Without instructor', however, seems to have increased by around 60% when lockdowns began and then stayed around that value, even after lockdowns ended. Finally, we can see that the ratio of with to without instructor was 2:1, but that has decreased over lockdowns to around 1.33:1. These changes support the previous hypothesis that, learning hours will increase overall, however, the proportion of hours with an instructor will decrease, but again, we can not say for certain.

2.2.4.2 Age Distribution

We have categorised age into 3 groups, 16-18, 19-21, 22-24. This is because these ages best represent the schooling separations. Typically, the first is high-school age, the second is university age and the last is post-university age. However, the ages are distributed equally among these groups, therefore, we use Figure 2.6 to observe the distribution.

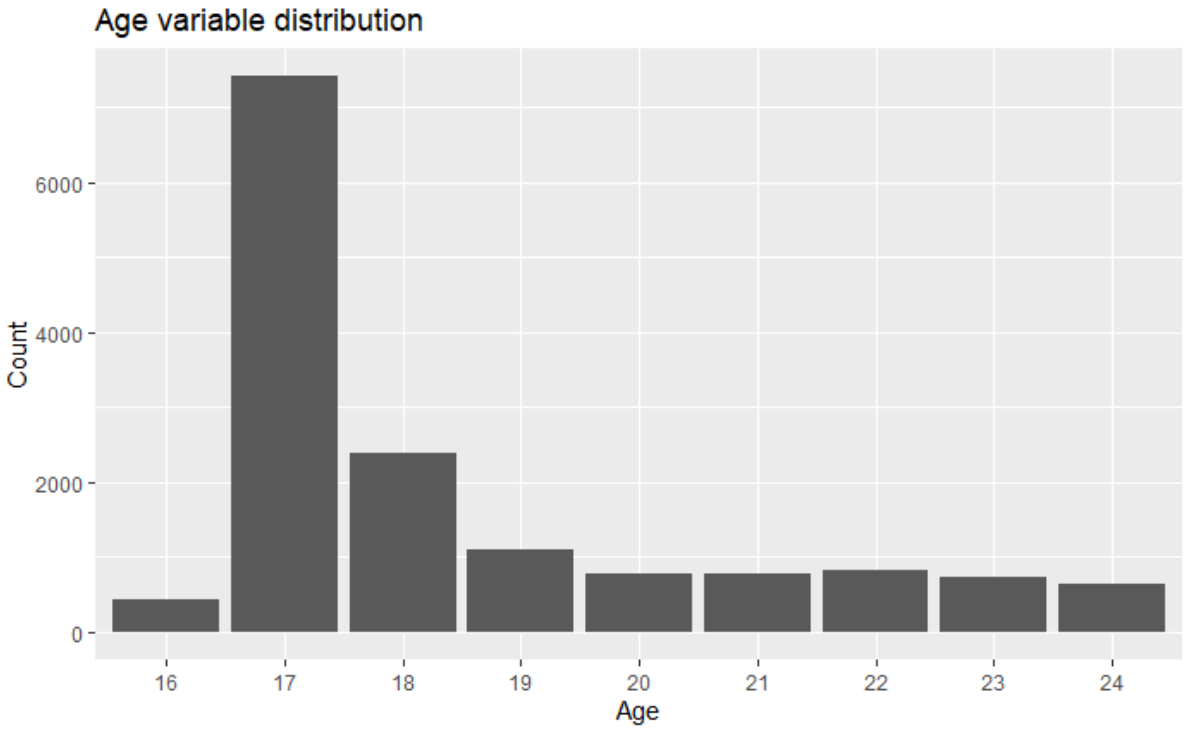

Figure 2.6: Age distribution

As can be seen, the majority of ages lie in ages 17 and 18. The other ages are almost equally distributed with 16 slightly below and 19 slightly above. We can infer that the lack of 16 year old participants is due to the fact that in the majority of cases, people are required to be 17 to learn to drive [5].

2.3 Modelling

The program used in this project for modelling is R, which utilises the Iteratively Reweighted Least Squares, or IRLS, algorithm to estimate the regression parameters. The following section explains the mathematics behind how they are estimated, as well as the structure of the models we will be using. This section utilises mathematical derivation explained in several statistics textbooks [6][7][8].

2.3.1 Linear Regression

A linear model assumes a response y_i, $i = 1, ..., n$ (for n observations), that is modelled by a linear function of p explanatory variables x_j, $j = 1, ..., p$, plus an error term ϵ_i.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i$$

For linear regression we make 2 important assumptions, the response variable is continuous and normally distributed.

2.3.2 Generalised Linear Models

A generalised linear model, henceforth referred to as GLM, has the same structure as a linear model, however, it does not have the restriction of a continuous and normally distributed response. Instead, the outcome can be categorical or count data, for example. The distribution could also be skewed, but that is not as relevant to this project. To ensure these are true, every GLM has the following three components.

1. **Random Component**: There is only one random component of the model, which is the response variable. In GLMs it follows an exponential family distribution and its density can be written in the following form:

   $$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \theta_i)\right],$$

   where ϕ and θ_i are called the dispersion and canonical parameters respectively. It can also be shown that,

   $$\mathrm{E}[Y_i] = b'(\theta_i) = \mu_i$$

   $$\mathrm{Var}[Y_i] = \phi b''(\theta_i) = \phi V(\mu_i)$$

   where $V(\mu_i)$ is called the variance function.

2. **Systematic Component**: This component specifies that the explanatory variables, $\mathbf{x}_j$, can be expressed in terms of the linear predictor, η_i. It is a linear combination of $\mathbf{x}_j$ and the unknown regression co-efficients β_j, $j = 1, ..., p$. Additionally, linear refers to the linearity of the co-efficients.

   $$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

3. **Link Function**: This function, $g()$ specifies the link between the random and

systematic components. It describes how the mean, μ_i, depends on the linear predictor η_i.

$$g(\mu_i) = \eta_i$$

2.3.2.1 Parameter Estimation

Now that we have our model, we want to estimate the regression parameters, $\boldsymbol{\beta}$. To do this we calculate the Maximum Likelihood Estimates of each parameter. This is done by calculating the log-likelihood of the response variables,

$$L = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)$$

then the derivative with respect to the regression parameters, also called score functions, and equating it to 0. However, $\frac{\partial L}{\partial \beta_j}$ cannot easily be calculated, so we use the chain rule to instead find,

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

From before, we know that $\mu_i = b'(\theta_i)$ and $\text{Var}[Y_i] = \phi b''(\theta_i)$, so can find the first derivative, noting that $L = \sum_i L_i$,

$$\frac{\partial L_i}{\partial \theta_i} = \frac{(y_i - b'(\theta_i))}{\phi} = \frac{(y_i - \mu_i)}{\phi},$$

and the (reciprocal of the) second,

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\text{Var}[Y_i]}{\phi},$$

We know that $\eta_i = \sum_j \beta_j x_{ij}$, so the fourth derivative is,

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Finally, since $\eta_i = g(\mu_i)$, the third derivative depends on the link function. We will denote it as $g'(\mu_i)^{-1}$ So, now we have that,

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}[Y_i]g'(\mu_i)} = 0, \text{ for } j = 1, ..., p$$

We can rewrite this in matrix form as,

$$\mathbf{X}^T\mathbf{W}\mathbf{D}^{-1}(y - \mu) = 0,$$

where $\mathbf{W}_{n\times n}$ is a diagonal matrix, such that $w_{ii} = (g'(\mu_i)^2 \text{Var}[Y_i])^{-1}$
and $\mathbf{D}_{n\times n}$ is a diagonal matrix, such that $d_{ii} = g'(\mu_i)^{-1}$

Now, we would utilise the IRLS algorithm to calculate estimates for the regression parameters, however, since we are only using models with a canonical link, this algorithm is equivalent to the Newton-Raphson Method, which is simpler.

The Newton-Raphson method uses iteration to maximise an approximation of the log-likelihood to calculate $\hat{\boldsymbol{\beta}}$. First, define $\mathbf{u}_{p\times 1} = \frac{\partial L}{\partial \beta}$, then we have the Hessian matrix, $\mathbf{H}_{p\times p}$, where $h_{ab} = \frac{\partial^2 L}{\partial a \partial b}$. Let $\mathbf{u}^{(t)}$ and $\mathbf{H}^{(t)}$ be the respective matrices evaluated at $\boldsymbol{\beta}^{(t)}$, the t^{th} guess for $\hat{\boldsymbol{\beta}}$. The process uses the first two terms of the Taylor Expansion of $L(\boldsymbol{\beta})$ near $\boldsymbol{\beta}^{(t)}$,

$$L(\boldsymbol{\beta}) \approx L(\boldsymbol{\beta}^{(t)}) + \mathbf{u}^{(t)T}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \mathbf{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}).$$

To get a formula for the next guess we solve,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = 0,$$

by rearranging for $\boldsymbol{\beta}$ to get,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbf{H}^{(t)})^{-1}\mathbf{u}^{(t)}.$$

However, we want the Fisher Scoring Method, which is similar to the above, except we use $\mathbf{I}^{(t)}$ instead of $\mathbf{H}^{(t)}$, with elements $-\text{E}[\frac{\partial^2 L}{\partial a \partial b}]$ evaluated at $\boldsymbol{\beta}^{(t)}$. So, the formula is,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\mathbf{I}^{(t)})^{-1}\mathbf{u}^{(t)}.$$

Multiply both sides by $\mathbf{I}$ to get,

$$\mathbf{I}^{(t)}\boldsymbol{\beta}^{(t+1)} = \mathbf{I}^{(t)}\boldsymbol{\beta}^{(t)} + \mathbf{u}^{(t)},$$

then replace $\mathbf{I}$ and $\mathbf{u}$ with their decomposed forms to get,

$$\begin{aligned}(\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})\boldsymbol{\beta}^{(t+1)} &= (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})\boldsymbol{\beta}^{(t)} + \mathbf{X}^T\mathbf{W}^{(t)}(\mathbf{D}^{(t)})^{-1}(y - \mu^{(t)}) \\ &= \mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)},\end{aligned}$$

where $\mathbf{z}^{(t)} = \mathbf{X}\boldsymbol{\beta}^{(t)} + (\mathbf{D}^{(t)})^{-1}(y - \mu^{(t)})$.

The $\mathbf{z}^{(t)}$ vector is also called the *working response* and $\mathbf{W}^{(t)}$ is also called the *working weights matrix*. With this, we are able to implement the IRLS algorithm in 3 steps:

1. Choose a reasonable starting value, $\mu^{(0)}$
2. Calculate $\mathbf{z}^{(t)}$ and $\mathbf{W}^{(t)}$ to get $\boldsymbol{\beta}^{(t+1)}$
3. Repeat step 2 until $\boldsymbol{\beta}^{(t)}$ converges to $\hat{\boldsymbol{\beta}}$

2.3.3 Regression Models for Count Data

The data that we are modelling are the learning hours of drivers. These are 'count' variables, meaning the values these variables can take are non-negative integers. Additionally, the numbers do not represent rankings or categories, they represent the amount or *count* of something.

2.3.3.1 Poisson Regression Model

The Poisson Regression model assumes that Y_i has a Poisson distribution. This is useful when dealing with count data. The probability mass function is,

$$f(y_i; \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!},$$

which can be written in exponential family form,

$$f(y_i; \mu_i) = \exp\left(y_i \ln\left(\mu_i\right) - \mu_i - \ln\Gamma(y_i + 1)\right),$$

from this we can derive the mean and variance,

$$\begin{aligned} \mathrm{E}[Y_i] &= \mu_i, \\ \mathrm{Var}[Y_i] &= \mu_i. \end{aligned}$$

We then need to choose a link function to use, which ensures that the transformed mean is positive, since the mean of a Poisson distribution is positive. The most commonly used is the log-link, which is what we have chosen to use. This ensures that the mean is positive, since the predictions are exponentiated,

$$\log\left(\mu(x)\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p,$$

$$\mu(x) = e^{\beta_0}(e^{\beta_1})^{x_1}...(e^{\beta_p})^{x_p}.$$

The interpretation of the co-efficients of this model are different than linear regression models, we have that an increase of 1 in any x_j, results in the intercept being multiplied

by e^{β_j}. For categorical variables, an increase of 1 is the equivalent of that factor present.

2.3.3.2 Overdispersion

Overdispersion is when the variance of a statistical model is higher than expected. This can lead to inaccurate predictions. For count data specifically, it is highly likely for observed count data to have high variance. Poisson distributions only have one free parameter, the mean, variance is assumed to be equal to it, therefore, it is very likely for predictions from a Poisson model to be overdispersed. In linear regression, this is not the case, since Normal distributions have a free parameter for variance.

2.3.3.3 Negative Binomial Regression Model

There are two methods to derive a Negative Binomial Model[**Hilbe**]. One where overdispersion is constant and one where it is variable, that are called NB-I and NB-II, respectively. R utilises NB-II, so we will focus on that version. In order to derive the distribution, we look at a Poisson model, but the mean parameter is distributed according to a Gamma distribution. Additionally, since the overdispersion is variable, the mean parameter is given by $\lambda_i u_i$ instead of just λ_i. We then derive the distribution,

$$f(y_i;\mu_i,\alpha) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i}.$$

This can then be rearranged into exponential family form, which will be omitted for simplicity, to give a mean and variance of,

$$\begin{aligned} \mathrm{E}[Y_i] &= \mu_i, \\ \mathrm{Var}[Y_i] &= \mu_i + \alpha\mu_i^2. \end{aligned}$$

As can be seen, the variance is now proportional to the quadratic of the mean and is dependent on the scale parameter of the Gamma distribution. This serves as a dispersion parameter of sorts that allows the variance to be different than the mean.

The link function and co-efficient interpretation of this model will be the same as the Poisson model.

2.3.4 Regression Models for Positive, Skewed Data

The Gamma Regression model assumes that Y_i has a Gamma distribution. This is useful when dealing with real, positive, right-skewed data. The probability mass function is,

$$f(y_i; \mu_i, \phi) = \frac{1}{y_i \Gamma(1/\phi)} \left(\frac{y_i}{\mu_i \phi} \right)^{1/\phi} \exp\left(-\frac{y_i}{\mu_i \phi} \right),$$

which can be re-arranged into exponential family form,

$$f(y_i; \mu_i, \phi) = \exp\left\{ \frac{y_i/\mu_i - (-\ln \mu_i)}{-\phi} + \frac{1-\phi}{\phi} \ln y_i - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi} \right) \right\}.$$

This gives a mean and variance of,

$$\begin{aligned} \mathrm{E}[Y_i] &= \mu_i, \\ \mathrm{Var}[Y_i] &= \mu + \phi \mu_i^2. \end{aligned}$$

Chapter 3

Results

3.1 Modelling

We want to create models to verify our 2 hypotheses:

1. Number of overall hours increased due to pandemic
2. Proportion of hours without driving instructor increased due to pandemic

To do this we need 2 separate models, one for the total and one for the proportion.

3.1.1 Model Choice

The dependant variables in our question are the number of learning hours. These variables are count variables, meaning they measure the amount of something and are non-negative integers. This means that for the total model we should use a Poisson Model. However, recalling Section 2.3.3.2, one of the assumptions that is made is that the mean is equal to the variance. In our dataset we know that the means of our dependant variables, as seen in Table 2.1, are 42.4 and 30.1, while the variances are 953.6 and 1448.9 for with and without instructors, respectively. As can be seen, the disparity between these parameters is very large, which causes overdispersion.

Our solution to this issue is to use a Negative Binomial Model. This model is similar to the Poisson, but does not have the same restriction on mean being equal to variance and is described in Section 2.3.3.3. We will include both models to compare results.

Additionally, as can be seen in Figure 3.1, the learning hours variables are skewed right. This means that the data is not normally distributed, instead the mean is lower than

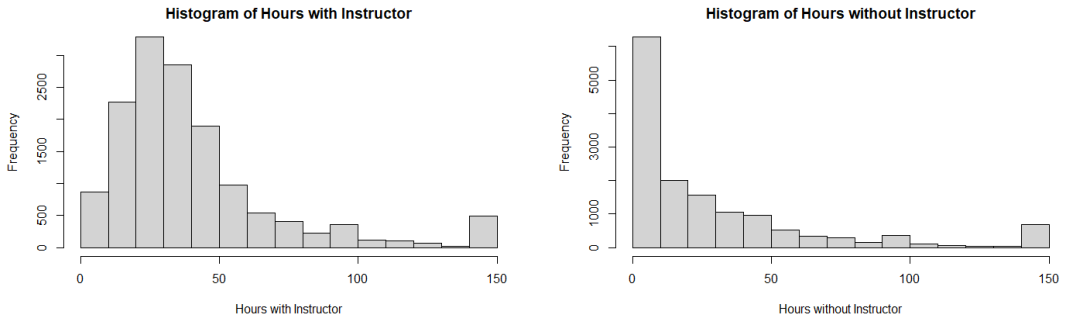

Figure 3.1: Histograms for Dependent Variables

expected. For data like this we can also use an Inverse-Gaussian distribution with a log-link, or a Gamma distribution with an inverse-link. We can compare the AIC that each of these models gives and choose one with a lower value.

Model	**Total Model AIC**
Poisson	295364
Negative Binomial	130561
Gamma	128526
Inverse Gaussian	130690

Table 3.1: Model AIC comparison

From Table 3.1, we can see that the Poisson model has a very high AIC, due to the overdispersion, so we can disregard that model. The 3 other models have very similar AIC, so we choose to use the negative binomial. The reason for this is that it was designed for count data, which is what our variables are, as well as only working for positive data, meaning values of 0 are not observed.

For our proportional model, we can't use Poisson or Negative Binomial, since the variable is not an integer, however, the data is still right skewed. Therefore, we will use a Gamma model with a log-link for this variable.

3.1.2 Model Results

We have fitted two models, one for total hours and one for proportion between with and without instructor. The interpretation of the co-efficients are as mentioned in Section 2.3.3.1. For our model, every independent variable is a factor variable, meaning it can take 1 of the given options. The model outputs are shown in Figures A.1 and A.2. They are included in Appendix A, due to their large size. They already have the estimates exponentiated for simplicity. So, in order to calculate the mean number of hours or proportion, we simply take the intercept estimate and multiply it by any factors that

are present. We can also evaluate individual factors by the value of their estimate, remembering that a decrease in hours is a positive impact, while a change in proportion is neutral.

1. **Gender**: This appears to be one of the higher factors that affects learning hours, as can be seen being male decreases learning hours by 21% and increases the proportion of time without an instructor by 19%. We can infer that females have a higher proportion of hours with a driving instructor and overall more hours required to pass the test. The gender factor had no interactions with any other variables, meaning its an independent factor. This effect is likely due to societal factors, but by including it we have removed its influence and can better ascertain whether lockdowns had any effect.

2. **Age and Education**: We hypothesize that older participants would be busier and thus have a higher numbers of hours. One reason for this would be education, so we have introduced an interaction with age and education status, which is significant, giving weight to our hypothesis. The effects of age on total number of hours is minimal, but only for those not in full-time education. Ages 19-21 have a 4% increase, while 22-24 have no effect. However, for those in education the increase in hours are 7% and 15% for the same two ages and the decrease for the baseline, which is ages 16-18 is 8%. So we see that education has a large impact when looked at in conjunction with age. We can conclude from this that those in education are busier than those that are not, and as you get older the level of business increases, which maps on to reality, since university levels of education require more time put in than high-school. By busier, what we mean is that participants have less time for lessons per a given time span and must therefore make that up with additional lessons, thus increasing their overall hours. As for the proportion of hours, we find that ages 19-21 and 22-24, who are not in full-time education, have an increased proportion for hours with an instructor of 36% and 44% respectively. Additionally, those not in education have a decreased proportion in hours with instructor for ages 16-18 of 27%, while ages 19-21 and 22-24 have increases of 76% and 31%. From this we can infer that 16-18 year-olds have a lot more hours with mentors/parents than an instructor, even more so if they are not in education. This could be because younger participants have to rely on others' monetary contributions to learn to drive with an instructor, while learning from others is free. This applies more so for those not in education, as that could be correlated to lack of household income. When looking at ages 19-21, they are more likely to get taught by an instructor, but even more so when they are in education. This could be because students tend to move away from home to go to university and thus have no mentors/parents to

learn from. Finally it seems that education does not affect those of ages 22-24 as much. The interpretation for this finding is difficult to ascertain, as is the last one.

3. **Country**: It appears that the country a participant is from does not have much effect on the number of hours. Specifically, Wales has almost no effect at all, while Scotland has an increase of 4%, which is very small. This could be due to a variety of societal factors, but it is useful to isolate that effect from the lockdown effect. However, for the proportion it appears that both Scotland and Wales have an increased proportion of learning without an instructor. It could be that instructors are more common in England, the exact reason is difficult to ascertain.

4. **Lockdown Impacts**: We have 9 lockdown impacts from Question 10 in the survey. Option 1 was removed as it stated 'No Impact', which is not useful, apart from verifying that it did indeed have no impact on the number of hours.

   (a) Practical cancelled: We can see that this is the largest negative impact. The effect is an 11% increase in the number of hours. This is what we would expect, since a participant would likely need more lessons in between the cancelled test and the actual one. This may not occur if the actual test is booked very soon after but, due to some lockdowns lasting several weeks, this was usually not the case. Additionally, it has had the effect of increasing the proportion of time learned from an instructor. The implications of this are difficult to ascertain.

   (b) Unable to book practical: This option is very similar to the previous, except less severe, which is reflected in its value only being a 5% increase in hours. A cancellation usually means the next available date is further along in time and it occurs without notice, so no planning is involved. Conversely, being unable to book may not delay the test by as long and it allows the participant to plan around this setback.

   (c) Theory cancelled/unable to book: These options are about the theory exam, which is a separate test from the practical. We would expect to see no effect on hours from these 2 options, which is the case. However, the proportion of time spent learning from a parent or mentor has increased by 19% and 30% for the respective options. The reason for this is unclear.

   (d) Less/more driving experience than intended: These 4 options seem quite self-evident, however, we see that some results are unexpected. For less driving experience, there is a 3% and 6% increase and for with and without instructor respectively. One possibility is that the participant did not get less absolute driving experience, but instead less driving experience per week for a longer time span. This would result in a larger number of hours, but the participant may feel like they got less experience, however, this interpretation is covered

almost exactly by the final option, 'Intended driving experience (but longer). Another possibility is that this option implies the absolute number of hours, but the participant is simply incorrect when describing their experience. They may have felt like they got less experience than intended due to other factors. The effect of this option is difficult to infer, as the statement in the question has multiple interpretations. For more experience, we would expect to see an increase, however, instead we have a 6% decrease and a 30% increase for with and without instructor respectively. Again this may be related to perceived experience being different than number of hours. With an instructor who is trained for this job, it may be that time spent with them feels more experiential than time spent with a parent or mentor. This could mean that less hours are required from an instructor to feel the same amount of experience gained, hence the decrease in number of hours. Finally, for the proportion we have even more unexpected results. We find that those who answered that they had less driving experience with an instructor had an increased or unchanging proportion of time spent with an instructor (this conclusion can be made from the Confidence Intervals found in the table). This has no reasonable explanation, it likely due to error of judgement by the participant. Again, those that reported less driving experience with a parent or mentor were recorded to have had a higher proportion of time spent with a parent or mentor. Again this may be an error. However, for more driving experience, we have that the reported change is the same as the recorded estimate. We see large changes of 61% and 69% increases for their respective category. This is an expected result.

(e) Intended driving experience (but longer): The effect of this option is a 14% increase in the proportion of hours with an instructor and overall hours not changing significantly. Since participants self described having their intended driving experience, we can take this to mean that no change in the overall hours was expected. However, the reason it took longer may be because the lack of availability of instructors due to lockdown limitations, as we see an increase in their proportion.

5. **Pre-covid and post-covid**: This factor removes the effects of pre and post-covid dates on the estimates. And we can see the overall changes with confounding factors removed. From this we can see that the number of overall hours during lockdowns did not change from before. However, after they increased by 12%. We could see this as the effects of the lockdowns being delayed until after they had ended. As for proportion, we can see that before the lockdowns, the proportion was 57% higher for instructors than during. And after it increased by 33%. These two factors line

up mostly with the hypothesis stated previously. The proportion of parent/mentor lessons went up a lot during lockdowns and was higher after lockdowns than when they first began. Additionally, the overall hours spent learning did not increase during the pandemic, but did so after they had ended.

3.1.3 Conclusion

We have successfully removed the effects of any confounding variables, including age, gender, country, and so on. Thus, we are able to address our hypothesis made in Section 2.2.3. Firstly, the overall number of learning hours did NOT increase during the pandemic, but did so afterwards. This could be attributed to effects of the pandemic only presenting themselves after some time. Additionally, we found that the proportion of hours spent learning with an instructor decreased and thus the time spent with parents/mentors increased during the pandemic. This effect also lasted until after the pandemic, although not as much of a difference. This all supports the idea that the pandemic made it more difficult to learn with an instructor, as many of them stopped working during lockdowns and as such finding one was more difficult. This led to more people learning from their parents or mentors with the same number of hours as they did before. Unfortunately, we do not have the number of failed driving tests, or else we could compare the number of failures before the lockdowns with during. If the number increased we could infer that learning from instructors helps more than with untrained people, but if the number didn't decrease, then we could say that instructors are not as important as they should be.

With the information that the learning process decreased in quality after the pandemic, where quality means that the proportion of hours taught by an instructor decreased, we can use the data we have to answer more questions. For example, does this lowering in quality impact the number of accidents in the first year of driving? If it does not, then what does? We should be able to utilise the conclusions from this study to answer more prevalent questions.

Chapter 4

Project Management

This project was headed by Professor Jane Hutton and was conducted in conjunction with the Transport Research Laboratory.

4.1 Teamwork

TRL had proposed several questions to be looked into. This meant that there were other students in the department who worked with the same dataset to complete their own projects. Two third-year students, Vanessa and Dodzia worked with me, in order to cut down on the time spent cleaning the data. Eventually, our paths diverged, as their questions dealt with aspects mine did not. Having multiple people tackle the same problem is useful, as we tended to use different methods. If we got the same results then we knew they were more reliable than if we were working alone. In addition to this we could also steer each other in the right direction when there was a problem that was new to one of us. Working alone can often result in tunnel vision, where it becomes difficult to view your work from another perspective, while Prof. Hutton's feedback was extremely useful, conversing with people who are working in tandem with you can help alleviate this lack of perspective.

4.2 Schedule

In order to complete work on time, I created a rough outline of what should have been completed by certain times. This meant having Data Cleaning and Exploration completed by the end of the first term, leaving the entirety of Term 2 to refine the modelling process. Unfortunately, it did not exactly go according to the plan and the objectives from the

first term overran into the second. However, working during the holidays meant that this was not an issue.

4.2.1 Meetings

Meetings were held almost every week from when the project began. The other two students working on similar projects also attended these meetings, meaning we could discuss our ideas to streamline our process in cleaning the data and working towards answering our questions. These meetings involved discussion on what we had done for the preceding week and planning on what to do for the proceeding week. This helped orchestrate a structure to the project, as without these weekly meetings, sticking to the schedule would have been a lot more complicated.

Occasionally, we also had meetings with members of TRL, which were liaised by Prof. Hutton, to discuss our progress. Since they were working towards similar goals and had a lot more experience, their input was extremely helpful. Additionally, they provided feedback on our presentations before we gave them.

Appendix A

Additional Plots

	Total Hours		
Predictors	*Estimate*	*CI*	*p*
(Intercept)	72.95	70.01 – 76.03	**<0.001**
Male	0.79	0.77 – 0.80	**<0.001**
Aged 19-21	1.04	1.00 – 1.09	**0.042**
Aged 22-24	1.00	0.97 – 1.05	0.813
Scotland	1.04	1.01 – 1.08	**0.024**
Wales	1.01	0.97 – 1.06	0.552
Practical Cancelled	1.11	1.08 – 1.14	**<0.001**
Theory Cancelled	1.03	0.99 – 1.06	0.183
Unable to book practical	1.05	1.03 – 1.08	**<0.001**
Unable to book theory	0.97	0.93 – 1.01	0.113
Less Driving Experience with Instructor	1.03	1.00 – 1.06	**0.033**
Less Driving Experience with Parent/Mentor	1.06	1.01 – 1.10	**0.010**
More Driving Experience with Instructor	0.94	0.90 – 0.98	**0.002**
More Driving Experience with Parent/Mentor	1.30	1.27 – 1.34	**<0.001**
Intended Driving Experience (but longer)	0.98	0.95 – 1.01	0.155
Full-time Education	0.93	0.90 – 0.96	**<0.001**
Pre-covid	0.99	0.96 – 1.03	0.680
Post-covid	1.12	1.08 – 1.16	**<0.001**
19-21:Full-time Education	1.07	1.01 – 1.13	**0.018**
22-24:Full-time Education	1.15	1.08 – 1.24	**<0.001**
Observations	14436		
R^2 Nagelkerke	0.156		

Figure A.1: Negative Binomial Model Estimates for Total Learning Hours

	Proportion of Hours		
Predictors	*Estimate*	*CI*	*p*
(Intercept)	3.78	3.23 – 4.43	**<0.001**
Male	0.81	0.76 – 0.88	**<0.001**
Aged 19-21	1.36	1.15 – 1.62	**<0.001**
Aged 22-24	1.44	1.23 – 1.69	**<0.001**
Scotland	0.78	0.69 – 0.89	**<0.001**
Wales	0.79	0.68 – 0.94	**0.006**
Practical Cancelled	1.33	1.21 – 1.47	**<0.001**
Theory Cancelled	0.81	0.71 – 0.93	**0.002**
Unable to book practical	1.11	1.00 – 1.22	**0.040**
Unable to book theory	0.70	0.60 – 0.81	**<0.001**
Less Driving Experience with Instructor	1.10	1.00 – 1.23	0.061
Less Driving Experience with Parent/Mentor	0.75	0.65 – 0.87	**<0.001**
More Driving Experience with Instructor	1.61	1.36 – 1.91	**<0.001**
More Driving Experience with Parent/Mentor	0.31	0.28 – 0.34	**<0.001**
Intended Driving Experience (but longer)	1.14	1.03 – 1.27	**0.014**
Full-time Education	0.73	0.65 – 0.82	**<0.001**
Pre-covid	1.57	1.39 – 1.78	**<0.001**
Post-covid	1.33	1.16 – 1.53	**<0.001**
19-21:Full-time Education	1.76	1.41 – 2.20	**<0.001**
22-24:Full-time Education	1.31	0.99 – 1.76	0.064
Observations	11064		
R^2 Nagelkerke	0.362		

Figure A.2: Gamma Model Estimates for Proportion of Hours with/without Driving Instructor

Factor	N(%)	***Hours*** With Instructor	Without Instructor
Total	14,436	42.44	30.04
Gender			
Male	5,444 (37.7%)	33.63	27.41
Female	8,992 (62.3%)	47.79	31.69
Age			
16-18	9,799 (67.9%)	36.99	33.77
19-21	2,540 (17.6%)	54.59	22.15
22-24	2,097 (14.5%)	53.21	22.13
Country			
England	12,641 (87.6%)	42.55	29.51
Scotland	1,133 (7.8%)	42.19	34.29
Wales	662 (4.6%)	40.77	32.75
Time-frame			
Pre-covid	6,238 (43.2%)	43.07	21.77
Post-covid	1,122 (7.8%)	48.72	39.04
During Covid	7,076 (49.0%)	40.89	35.90
Lockdown Impacts (1.2.2.3)			
Q10.1	936 (6.5%)	37.0	35.59
Q10.2	4,246 (29.4%)	47.98	34.71
Q10.3	1,180 (8.2%)	33.88	45.09
Q10.4	4,315 (29.9%)	41.65	38.49
Q10.5	1,017 (7.0%)	31.81	44.06
Q10.6	3,116 (21.6%)	44.83	37.87
Q10.7	899 (6.2%)	44.97	36.10
Q10.8	798 (5.5%)	42.24	29.94
Q10.9	2,109 (14.6%)	43.45	25.44
Q10.10	2,509 (17.4%)	43.30	32.98
In full-time Education			
Yes	9,849 (68.2%)	38.75	32.31
No	4,587 (31.8%)	50.36	25.15

Table A.1: Summary statistics of factors

Bibliography

[1] J M Kendall. "Designing a research project: randomised controlled trials and their principles". In: *Emergency Medicine Journal* 20.2 (2003), pp. 164–168. ISSN: 1472-0205. DOI: 10.1136/emj.20.2.164. eprint: https://emj.bmj.com/content/20/2/164.full.pdf. URL: https://emj.bmj.com/content/20/2/164.

[2] Lauren Thomas. *What is a longitudinal study?* Oct. 2021. URL: https://www.scribbr.com/methodology/longitudinal-study/.

[3] A. Pressley. *A review of interventions which seek to increase the safety of young and novice drivers*. Department for Transport, 2016.

[4] Alex Nice Jess Sargeant. *Coronavirus lockdown rules in each part of the UK*. URL: https://www.instituteforgovernment.org.uk/explainers/coronavirus-lockdown-rules-four-nations-uk.

[5] Driving and Transport. *Learn to drive a car: step by step*. URL: https://www.gov.uk/learn-to-drive-a-car.

[6] Alan Agresti. *Categorical data analysis*. eng. A Wiley-Interscience publication. New York [u.a.]: Wiley, 1990, XV, 558 S.

[7] Joseph M. Hilbe. "Generalized Linear Models." In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Springer, 2011, pp. 591–596. ISBN: 978-3-642-04898-2. URL: http://dblp.uni-trier.de/db/reference/stat/stat2011.html#Hilbe11.

[8] James W. Hardin and Joseph Hilbe. *Generalized Linear Models and Extensions, 3rd Edition*. 3rd. StataCorp LP, 2012. URL: https://EconPapers.repec.org/RePEc:tsj:spbook:glmext.