SOST10142 & 20142
Applied Statistics for Social Scientists

Essay Assignment

Student ID: 10915102

Contents

Introduction

In 2019 the Crime Survey for England and Wales found that disabled adults were over 2.5 times more likely to have experienced domestic violence compared to non-disabled adults (Office of National Statistics, 2013). As violent crime declines, those with developmental disabilities remain at disproportionate risk (Petersilia, 2001). Unfortunately, research on this remains somewhat scarce and there are many theories as to what causes people to be victimised (Fisher et al., 2016). Hans von Hentig, a criminal psychologist, who was one of the first to focus on characteristics of the victim rather than the relationship between the victim and perpetrator, identified four categories of people he considered more vulnerable to crime, the mentally disabled, the young, the old, and females (Petersilia, 2001).

This paper aims to study the relationship between disability and crime compared to other factors, using the Life Opportunities Survey (LOS) data from between June 2010 and March 2012, otherwise known as Wave Two (Office of National Statistics, 2013). The LOS was a longitudinal survey conducted by the Office of National Statistics that ran from 2009 to 2014 in three, two-year waves, collecting data on impairment and life limitations and respondent's ability to participate in different areas of life (Office of National Statistics, 2013). For this paper, Wave Two was used as the publicly available data set containing only the first and second waves was more complete than the data set containing the more recent third wave. The LOS is available through the UK Data Service (Office for National Statistics, 2016).

Data description

This data set contains a large amount of data, spanning two "waves". While the entire data set contains 47572 observations over 2051 variables, approximately half of both the observations and variables are for Wave Two data, which is the wave focused on for the purposes of this study. The data set was then narrowed down to the variables of interest, these fell into three categories, limitations, crimes, and other descriptors.

Limitation Variables

All limitation variables recorded the frequency of limits to the amount of kind of activities the respondent could do. These were recorded on a scale of one to five. One being the highest frequency of limitation originally, however, this was recoded to better track overall limitation experienced. There were fourteen limitation variables used as follows:

- W2SEELIM: difficulty seeing (near or farsightedness)
- W2HEARLIM: difficulty hearing
- W2SPKLIM: difficulty communicating with others
- W2MOBLIM: difficulty being mobile
- W2PAINLIM: suffering from constant or spells of pain
- W2CONDLIM: difficulty resulting from chronic health conditions
- W2BRETHLIM: difficulty breathing
- W2LRNLIM: difficulty learning
- W2INTELLIM: intellectual difficulty or developmental delays
- W2BEVLIM: social or behavioural difficulty
- W2MEMLIM: difficulty remembering or periods of confusion
- W2MENLIM: difficulty resulting from a long-term mental health condition

- W2DEXLIM: difficulty grasping, lifting, or holding objects
- W2OTHLIM: any other difficulty resulting from a physical condition, or mental or physical health.

All limitation variables were aggregated into a single variable LIM during data cleaning. LIM is the sum of all limitation values, recoded for 5 to represent constant impact of the limitation on respondent activities.

Crime Variables

All crime variables asked about types of crime personally experienced in the last twelve months, responses were a binary Yes/No. There were seven types of crime covered as follows:

- W2CRIM01: Theft of a motor vehicle or bicycle
- W2CRIM02: Other item stolen from respondent
- W2CRIM03: Someone entering respondent's home without permission
- W2CRIM04: Deliberate damage to respondent home, vehicle, or belongings
- W2CRIM05: Violent force used or threatened against respondent
- W2CRIM06: Any other crime

During data cleaning all crime variables were aggregated into a single variable, CRIMEXP. CRIMEXP is a reflection of the crime a given respondent experienced in the last twelve months. While it tracks types of crime, rather than pure amount, CRIMEXP still provides valuable information as to the severity of crime experienced by a respondent.

Other Descriptors

These variables are other potential explanations for crimes experienced by respondents. There are four other descriptor variables:

- W2SEX: Sex of respondent, coded Male/Female
- W2DVAGE: Age of respondent, derived variable, coded to 80
- W1ETHREP: Respondent ethnicity for original participants
- W2ETHREP: Respondent ethnicity for new entrants in wave two
- W2HHLDDV: Respondent's individual weekly income

W1 and W2 ETHREP were combined into a single ETHREP, as ethnicity was not recorded separately for each wave.

Data Cleaning

To clean the data, all W2 variables listed above were selected from the larger data set and all observations where W2CASE was NA, indicating the observation wasn't wave two, were removed, leaving 30483 observations. Next, non-answers (-9,-8,-7) were converted to NA values. The row sums were then calculated for crime and limitation variables. Finally, binary variables were calculated for age (0-55, 56+), limitations (Y/N), crime experienced (Y/N), and ethnicity (White, Unknown or Non-White. Precise data cleaning steps are described within the appendix.

Data Exploration

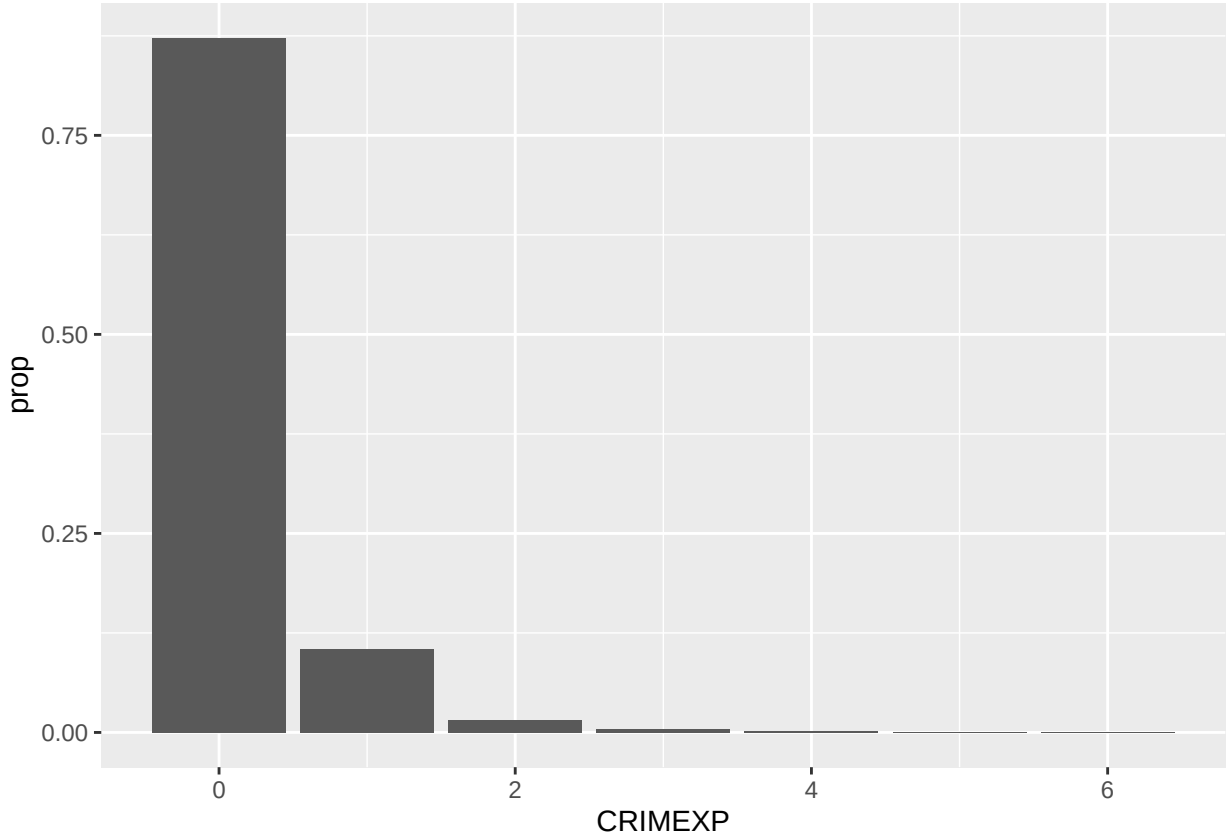
Crime Experienced

The values for types of crime personally experienced in the last twelve months is below. The vast majority of survey respondents (87.3%) experienced zero crime in the last twelve months.

```
table(final_data$CRIMEXP)
```

```
##
##     0     1     2     3     4     5     6
## 16948  2044   306    89    26     5     1
```

```
ggplot(final_data,aes(x=CRIMEXP,y=..prop..,group=1)) +
  geom_bar()
```



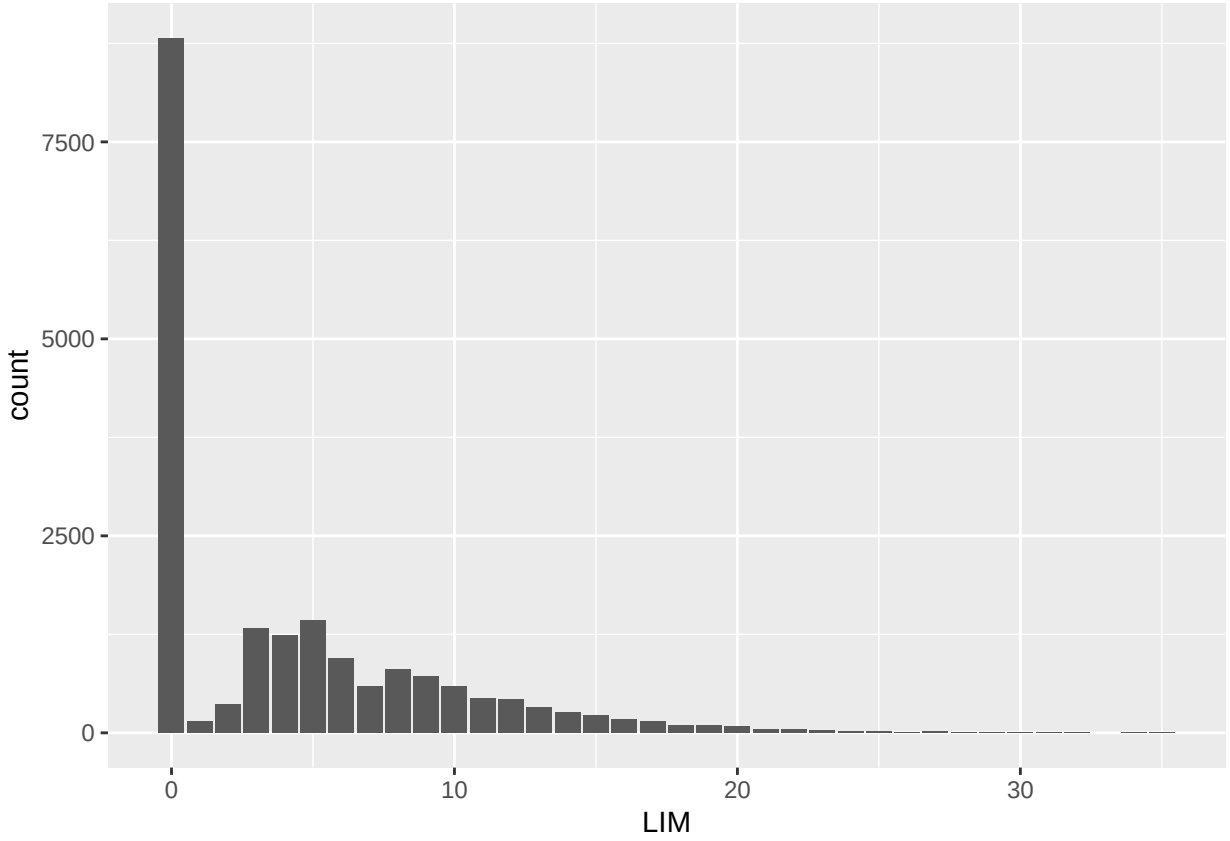
Severity of Limiting Conditions

LIM measures the severity of the limiting conditions, as evident by both the bar chart and summary table, this data is, unsurprisingly, skewed heavily to the right, with 64.3% of respondents at zero.

```
summary(final_data$LIM)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   3.000   4.208   7.000  35.000
```

```
ggplot(final_data,aes(LIM))+
  geom_bar()
```



Sex

The survey respondent's sex is fairly evenly split, with 52.2% female respondents.

```
table(final_data$W2SEX)
```

```
##
##     1     2
##  9144 10275
```

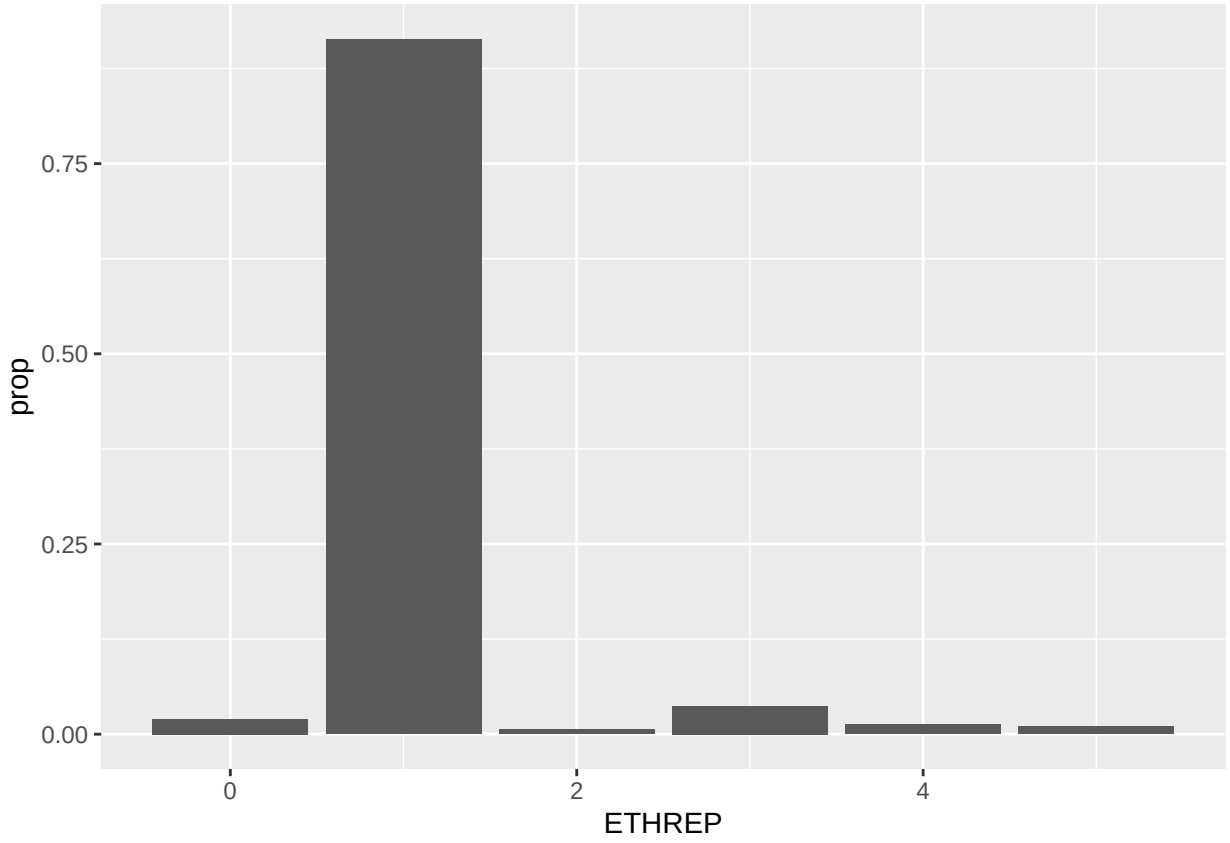
Ethinicity

The large majority (89.9%) of survey respondents are reported as white. The next highest category being Asian/Asian British at 4.7%.

```
table(final_data$ETHREP)
```

```
##
##     0     1     2     3     4     5
##   387 17736   127   716   252   201
```

```
ggplot(final_data,aes(x=ETHREP,y=..prop..,group=1)) +
  geom_bar()
```



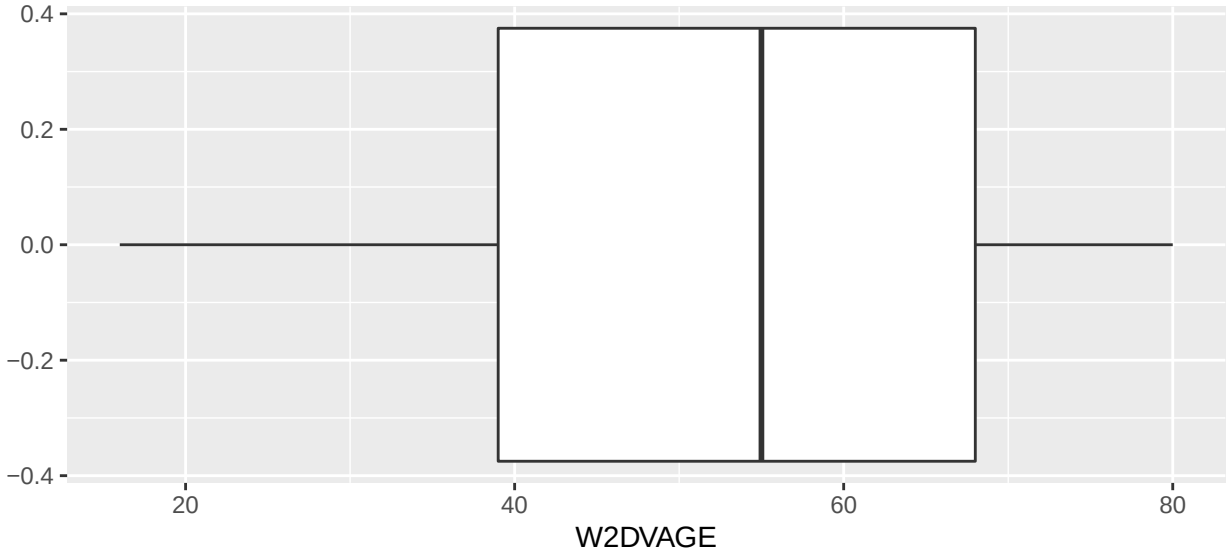
Age

Age appears to skew older, but most respondents are between 25 and 65, a reasonable range The values range between 15 and 80. 80 is capped within the original data, and respondents under 15 were filtered out as they were not asked about crime personally experienced.

```
summary(final_data$W2DVAGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   39.00   55.00   52.63   68.00   80.00
```

```
ggplot(final_data,aes(W2DVAGE))+
  geom_boxplot()
```



Income

Individual weekly income was also intended to be a part of this study, but as it is missing for 96.0% of respondents, it cannot be reliably studied and will not be included going forward.

```
summary(final_data$W2HHLDDV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0   323.0   577.0   737.1   962.0  3355.0   18416
```

Statistical Analysis

For the purpose of the statistical tests below, the following hypotheses and significance level will be used.

H_0: Age, sex, limitations, and ethnicity all have no impact on $\mu_{crime\ experienced}$

H_a: At least one of age, sex, limitations, and ethnicity have some impact on $\mu_{crime\ experienced}$

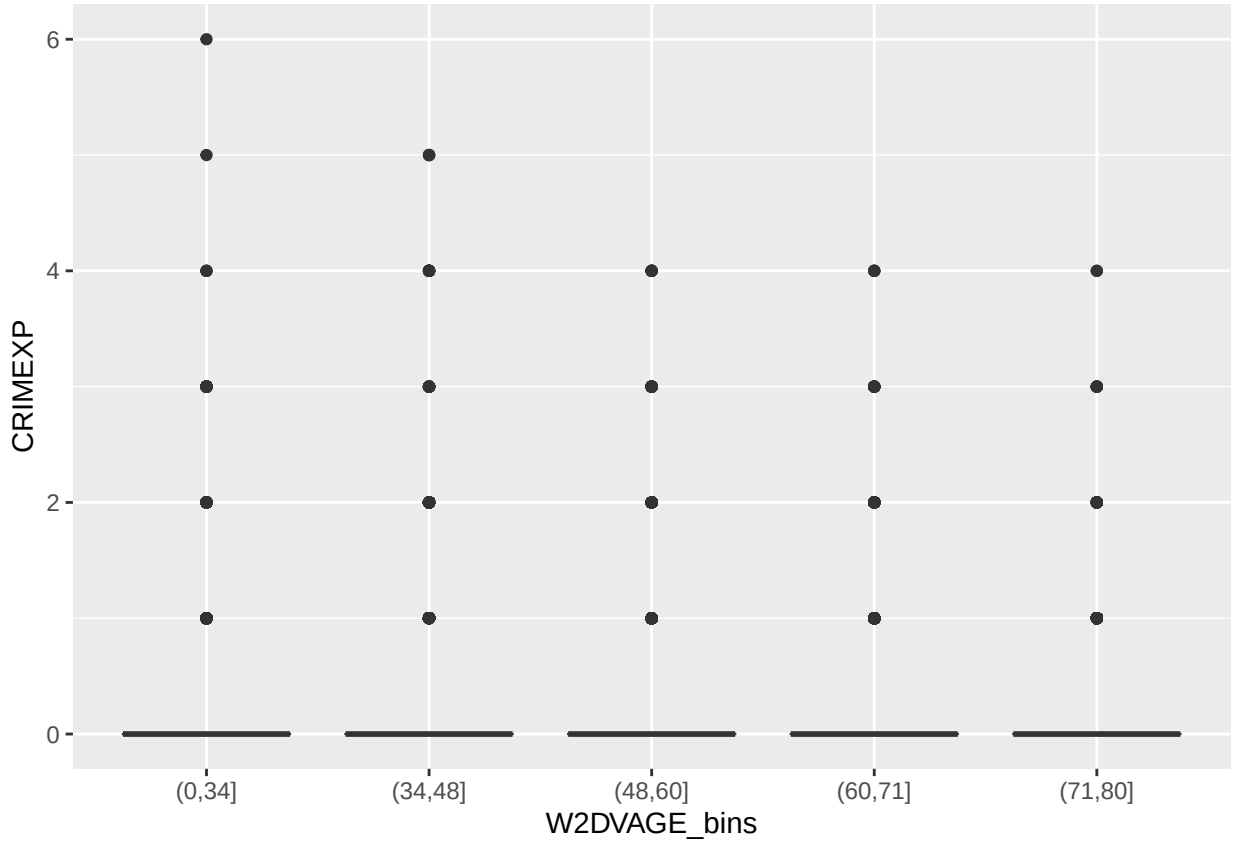
$\alpha = 0.05$

ANOVA and t-test

A t-test analyses the difference of means to determine if the difference is significant or not, it is used for small random samples from a normal population distribution or large random samples (Agresti, 2018). This test is generally robust to violations of the assumption of normality and R can conduct the test without assuming equal variance (Agresti, 2018). As this test only compares two groups, analysis of variability or ANOVA is a way to essentially perform a t-test on three or more categories. ANOVA assumes equal variance within each group tested, so variables without equal variance are tested using a Welch test.

Age

```
ggplot(final_data, aes(x=W2DVAGE_bins,y=CRIMEXP))+
  geom_boxplot()
```



All boxplots presented comparing variables to crime experienced are largely uninformative, due to the small means and large outliers. However, it is interesting to observe the apparent decrease in outliers for older age groups.

```
print(xtable(group_by(final_data, W2DVAGE_bins) %>%
    summarise(mean = mean(CRIMEXP, na.rm = T), sd = sd(CRIMEXP,
        na.rm = T))), comment = FALSE, type = "latex")
```

	W2DVAGE_bins	mean	sd
1	(0,34]	0.19	0.51
2	(34,48]	0.22	0.55
3	(48,60]	0.17	0.47
4	(60,71]	0.13	0.41
5	(71,80]	0.08	0.33

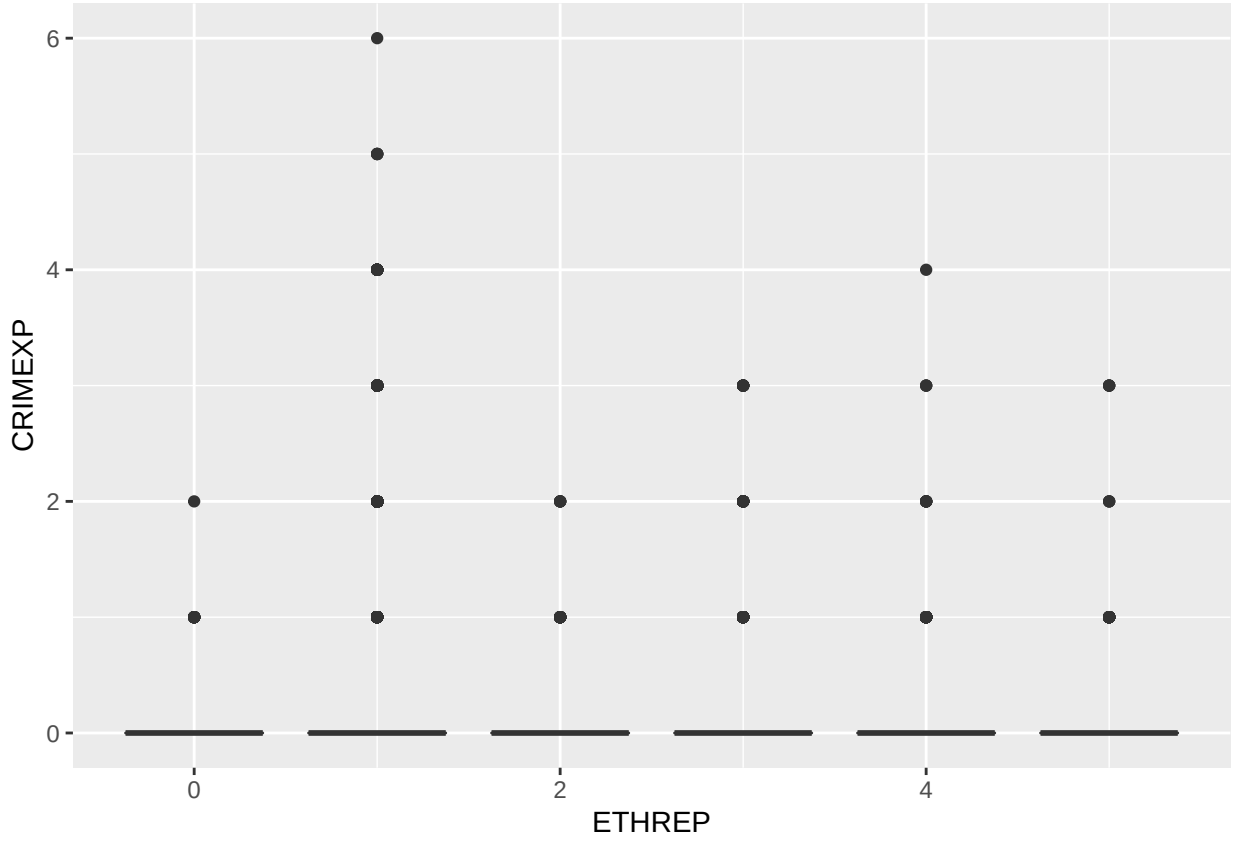
While the mean increases as age does at first, there is a fall in mean crime experienced after respondents between 35 and 48. There is also a decrease in standard deviation as respondents get older, thus a Welch test was used to conduct this test. The test resulted in a p-value of $< 2.2e\text{-}16$.

```
oneway.test(CRIMEXP ~ W2DVAGE_bins, data=final_data, var.equal=FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  CRIMEXP and W2DVAGE_bins
## F = 57.734, num df = 4.0, denom df = 9641.7, p-value < 2.2e-16
```

Ethnicity

```
ggplot(final_data, aes(x=ETHREP,y=CRIMEXP,group=ETHREP))+
  geom_boxplot()
```



Overall, outliers present seem related to sample size of the ethnic group, with the exception of Asian or Asian British (4) respondents. Despite having a very similar sample size to Chinese or Other (5) respondents, Asian or Asian British respondents exhibit higher outliers.

```
print(xtable(group_by(final_data, ETHREP) %>%
    summarise(mean = mean(CRIMEXP, na.rm = T), sd = sd(CRIMEXP,
        na.rm = T))), comment = FALSE, type = "latex")
```

	ETHREP	mean	sd
1	0.00	0.04	0.22
2	1.00	0.15	0.46
3	2.00	0.24	0.51
4	3.00	0.24	0.57
5	4.00	0.27	0.61
6	5.00	0.22	0.57

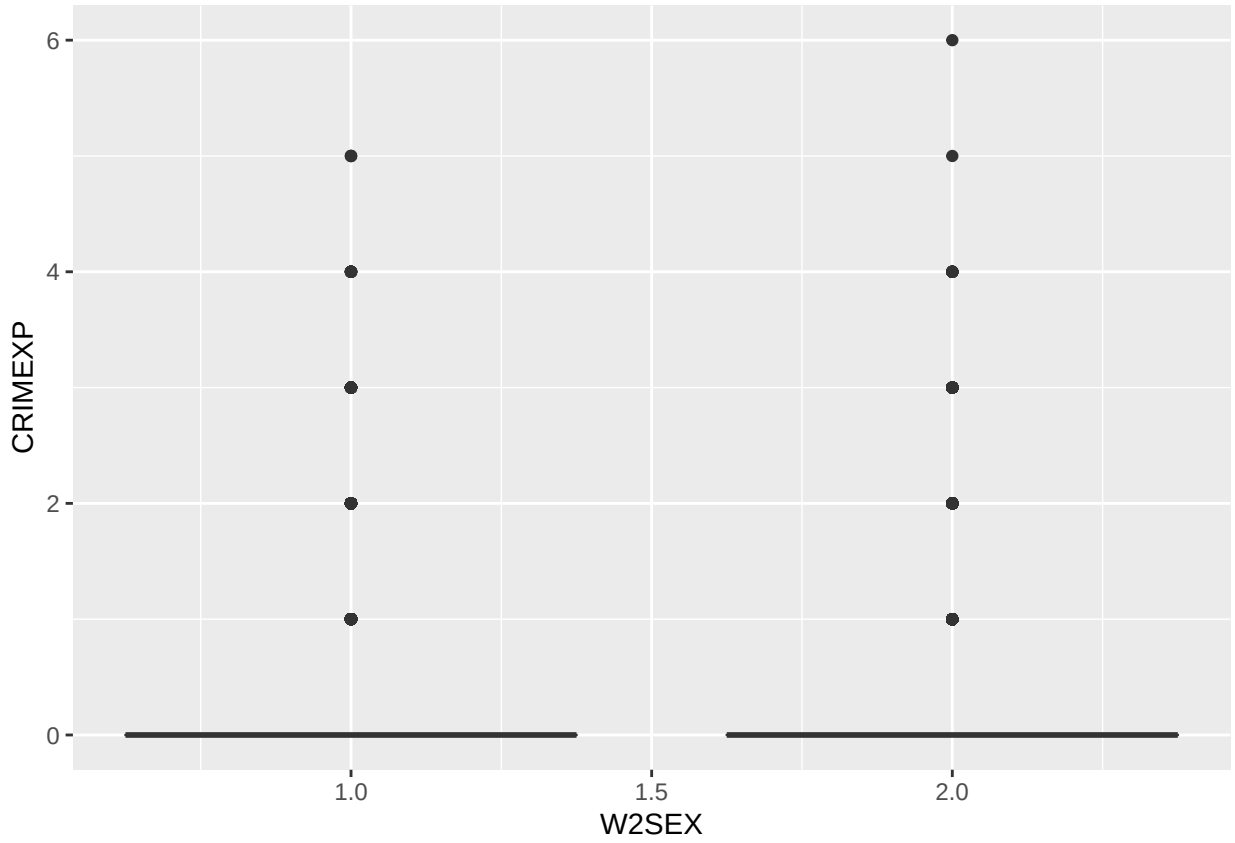
With noticeable lower group means for unknown and White respondents, all other ethnic groups appear to have similar means. Due to the variation in standard deviations, a Welch test was conducted resulting in a p-value less than 2.2e-16.

```
oneway.test(CRIMEXP ~ ETHREP, data=final_data, var.equal=FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  CRIMEXP and ETHREP
## F = 24.973, num df = 5.00, denom df = 576.66, p-value < 2.2e-16
```

Sex

```
ggplot(final_data, aes(x=W2SEX,y=CRIMEXP,group=W2SEX))+
  geom_boxplot()
```



Both groups appear generally equal, with one higher outlier for female respondents.

```
print(xtable(group_by(final_data, W2SEX) %>%
    summarise(mean = mean(CRIMEXP, na.rm = T), sd = sd(CRIMEXP,
        na.rm = T))), comment = FALSE, type = "latex")
```

	W2SEX	mean	sd
1	1.00	0.18	0.49
2	2.00	0.14	0.43

The means for male (1) and female (2) respondents are very similar, it is surprising that females experienced less crime than males. As W2SEX only has two possible values, a t-test was conducted, which resulted in a p-value of 1.171e-08.

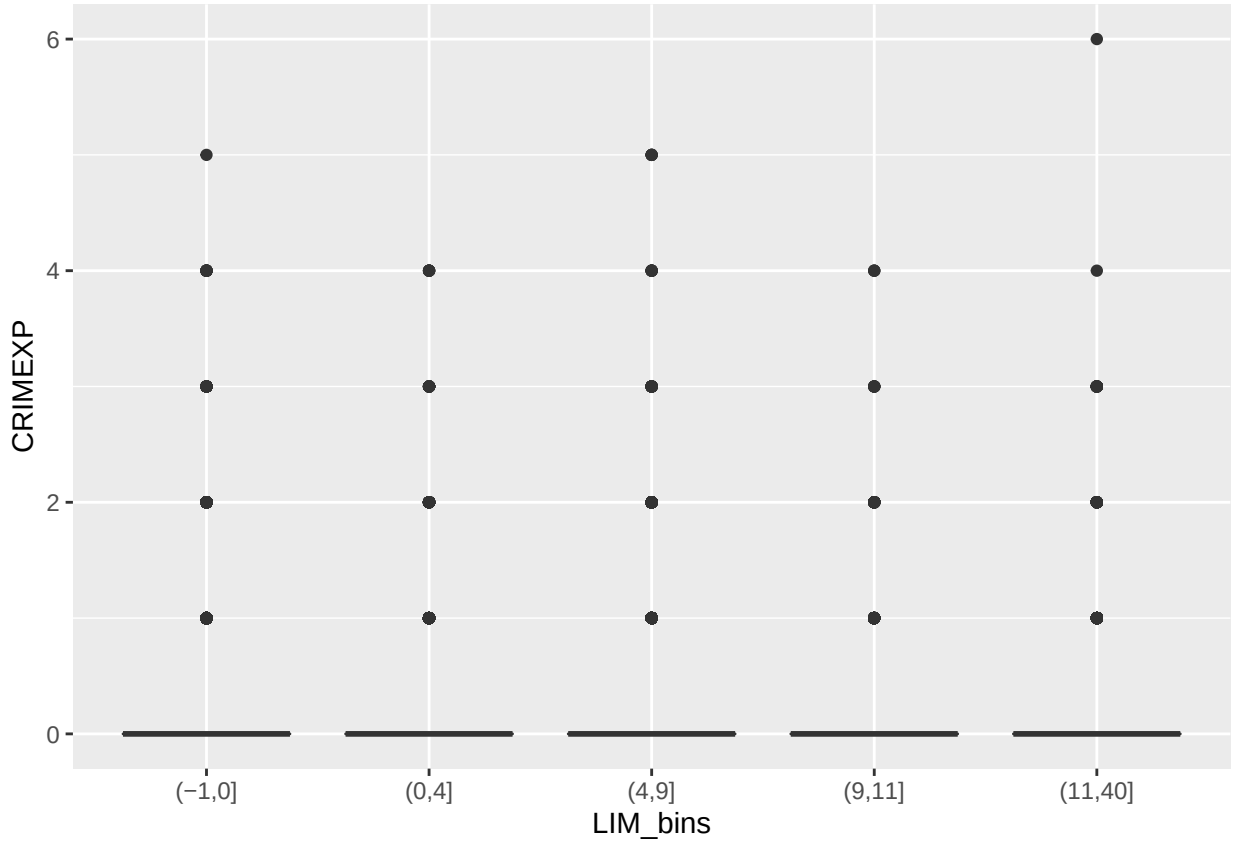
```
t.test(final_data$CRIMEXP ~final_data$W2SEX)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  final_data$CRIMEXP by final_data$W2SEX
## t = 5.7066, df = 18324, p-value = 1.171e-08
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  0.02511019 0.05138469
## sample estimates:
## mean in group 1 mean in group 2
##       0.1777122       0.1394647
```

Limitation Severity

```
ggplot(final_data, aes(x=LIM_bins,y=CRIMEXP))+
  geom_boxplot()
```



It's interesting the highest limitation group skips five crimes experienced but, given the smaller number of observations within each group above zero, the presence of high outliers is encouraging.

```
print(xtable(group_by(final_data, LIM_bins) %>%
    summarise(mean = mean(CRIMEXP, na.rm = T), sd = sd(CRIMEXP,
        na.rm = T))), comment = FALSE, type = "latex")
```

Unfortunately, group sizes for limitations are unequal due to a large portion (45.4%) of zeroes. Mean crime does increase as limitations do. Variance is not equal, so a Welch test was conducted, which returned a p-value of 1.14e-14.

	LIM_bins	mean	sd
1	(-1,0]	0.13	0.41
2	(0,4]	0.17	0.48
3	(4,9]	0.17	0.49
4	(9,11]	0.19	0.53
5	(11,40]	0.21	0.54

```
oneway.test(CRIMEXP ~ LIM_bins, data=final_data, var.equal=FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  CRIMEXP and LIM_bins
## F = 17.989, num df = 4.0, denom df = 4623.6, p-value = 1.14e-14
```

All variables tested were significant at a 0.05 alpha level, thus the null hypothesis is rejected. Ultimately, all variables studied here appear to be significant predictors for crimes experienced.

Odds Ratio and Chi-Squared

Odds Ratio

The odds ratio compares the odds of success of an event with odds of failure (Agresti, 2018). The odds ratios presented here go a step further and comparing odds found for different values of a given binary variable. When the odds ratio is greater than 1 odds of experiencing crime given the first group are higher than those given the second group (Agresti, 2018). An odds ratio of one indicates there is no difference between groups.

	Variable	Odds
1	Age	0.59793132977317
2	Ethnicity	1.23143463682717
3	Sex	0.780445640574121
4	Limitation	1.41744061121275

All odds ratios display the odds of the group suspected to be more vulnerable compared to the other group. Respondents with unknown or non-white ethnicities have higher odds of experiencing crime. Female respondents face lower odds, contradicting von Hentig's hypothesis that being female increased the odds of being victimised (Fisher et al., 2013). Older respondents face lower odds of experiencing crime and respondents with limitations face higher odds of experiencing crime.

Chi-Squared

A χ^2 test tests for statistical independence (Agresti, 2018). This is done by comparing the observed frequency of conditional probabilities with what would be expected if the variables had no effect on one another (Agresti, 2018). The larger this number, the less likely the variables are independent.

	Variable	Statistic	df	pvalue
1	Age	136.40	1.00	1.63277257796695e-31
2	Ethnicity	8.17	1.00	0.00426012621873636
3	Sex	32.90	1.00	9.70091053358868e-09
4	Limitations	62.34	1.00	2.88665655425596e-15

All p-values are far below 0.05, thus, the null hypothesis is rejected. In line with the odds ratios, ethnicity appears to have the weakest relationship with crime experienced.

Multiple Linear Regression

Linear regression is a fairly straightforward concept, even with multiple variables. The interact and estimates for a linear equation Inputs are the various variables included and output is predicted crime experienced.

Complete Data Model

```
regression_r<-data.frame(Model=character(),rSquared=double())
#Complete
regression_model<-lm(CRIMEXP~W2DVAGE_bins+W2SEX+LIM_bins+ETHREP_bins,data=final_data)
print(xtable(summary(regression_model)$coef),comment = FALSE,type = "latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.22	0.01	17.56	0.00
W2DVAGE_bins(34,48]	0.01	0.01	1.16	0.24
W2DVAGE_bins(48,60]	-0.04	0.01	-3.99	0.00
W2DVAGE_bins(60,71]	-0.09	0.01	-8.67	0.00
W2DVAGE_bins(71,80]	-0.15	0.01	-13.42	0.00
W2SEX	-0.04	0.01	-6.22	0.00
LIM_bins(0,4]	0.06	0.01	6.61	0.00
LIM_bins(4,9]	0.08	0.01	9.13	0.00
LIM_bins(9,11]	0.11	0.02	7.39	0.00
LIM_bins(11,40]	0.14	0.01	11.59	0.00
ETHREP_bins[-0.5,0.5]	-0.12	0.02	-4.93	0.00
ETHREP_bins[3.5,4.5]	0.09	0.03	3.25	0.00
ETHREP_bins[1.5,2.5]	0.05	0.04	1.20	0.23
ETHREP_bins[4.5,5.5]	0.05	0.03	1.50	0.13
ETHREP_bins[2.5,3.5]	0.06	0.02	3.29	0.00

```
r<-summary(regression_model)$adj.r.squared
regression_r[nrow(regression_r)+1,]=c("Full",r)
```

The first model was constructed using all variables and data. This model's intercept implies that the base amount of crime experienced in the last twelve months is 0.29. In contrast to what the odds ratio suggested, non-white ethnicities are associated with an increase in crime experienced. The disparity may be due to the unknown ethnicity values, as an unknown ethnicity is associated with a decrease in crime experienced. All p-values are significant at 0.05 so we can reject the null hypothesis.

White Respondent Model

```
regression_model_w <- lm(CRIMEXP ~ W2DVAGE_bins + W2SEX + LIM_bins +
    ETHREP, data = filter(final_data, ETHREP_binary == 0))
print(xtable(summary(regression_model_w)$coef), comment = FALSE,
    type = "latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.23	0.01	17.19	0.00
W2DVAGE_bins(34,48]	0.00	0.01	0.09	0.93
W2DVAGE_bins(48,60]	-0.05	0.01	-4.61	0.00
W2DVAGE_bins(60,71]	-0.10	0.01	-9.10	0.00
W2DVAGE_bins(71,80]	-0.16	0.01	-14.01	0.00
W2SEX	-0.04	0.01	-5.74	0.00
LIM_bins(0,4]	0.06	0.01	6.33	0.00
LIM_bins(4,9]	0.08	0.01	9.47	0.00
LIM_bins(9,11]	0.11	0.02	7.10	0.00
LIM_bins(11,40]	0.14	0.01	11.74	0.00

```
r <- summary(regression_model_w)$adj.r.squared
regression_r[nrow(regression_r) + 1, ] = c("White", r)
```

With the complete data, a good portion of change in the estimate seemed to be influenced by ethnicity, to investigate this further, the data was divided into two models using the ethnicity binary variable. The model using data from white respondents starts with a slightly higher intercept than from the full data set. The trends found in the complete model are all still found in the model using White respondents, however, the magnitudes changed.

Respondents of Unknown and Non-White Ethnicities Model

```
regression_model_poc <- lm(CRIMEXP ~ W2DVAGE_bins + W2SEX + LIM_bins +
    ETHREP_bins, data = filter(final_data, ETHREP_binary == 1))
print(xtable(summary(regression_model_poc)$coef), comment = FALSE,
    type = "latex")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.09	0.05	2.08	0.04
W2DVAGE_bins(34,48]	0.08	0.03	2.61	0.01
W2DVAGE_bins(48,60]	0.01	0.04	0.37	0.71
W2DVAGE_bins(60,71]	-0.05	0.05	-1.07	0.29
W2DVAGE_bins(71,80]	-0.00	0.06	-0.08	0.94
W2SEX	-0.06	0.03	-2.23	0.03
LIM_bins(0,4]	0.08	0.04	2.14	0.03
LIM_bins(4,9]	0.01	0.04	0.18	0.86
LIM_bins(9,11]	0.21	0.08	2.60	0.01
LIM_bins(11,40]	0.07	0.06	1.21	0.23
ETHREP_bins[3.5,4.5]	0.21	0.04	4.96	0.00
ETHREP_bins[1.5,2.5]	0.19	0.05	3.55	0.00
ETHREP_bins[4.5,5.5]	0.16	0.05	3.60	0.00
ETHREP_bins[2.5,3.5]	0.18	0.03	5.48	0.00

```
r <- summary(regression_model_poc)$adj.r.squared
regression_r[nrow(regression_r) + 1, ] = c("POC", r)
```

The intercept for the model using respondents of colour is notably lower than the complete model. Due the large majority of survey respondents being White, the overall model appears to differ significantly from the model for respondents of colour. In a departure from the trend exhibited by the complete model, the estimate continues to increase for ages between 48 and 60. After that the estimate decreases, but by less than the complete model does. Being female has a larger negative effect. Limitations do not have a consistent trend and overall, affect the estimate by less. Changes in ethnic group have a considerably larger impact than in the complete model. However, considering unknown ethnicity has a negative impact in the original model and serves as the base here, ethnicities likely still have the same impact, the baseline has simply shifted.

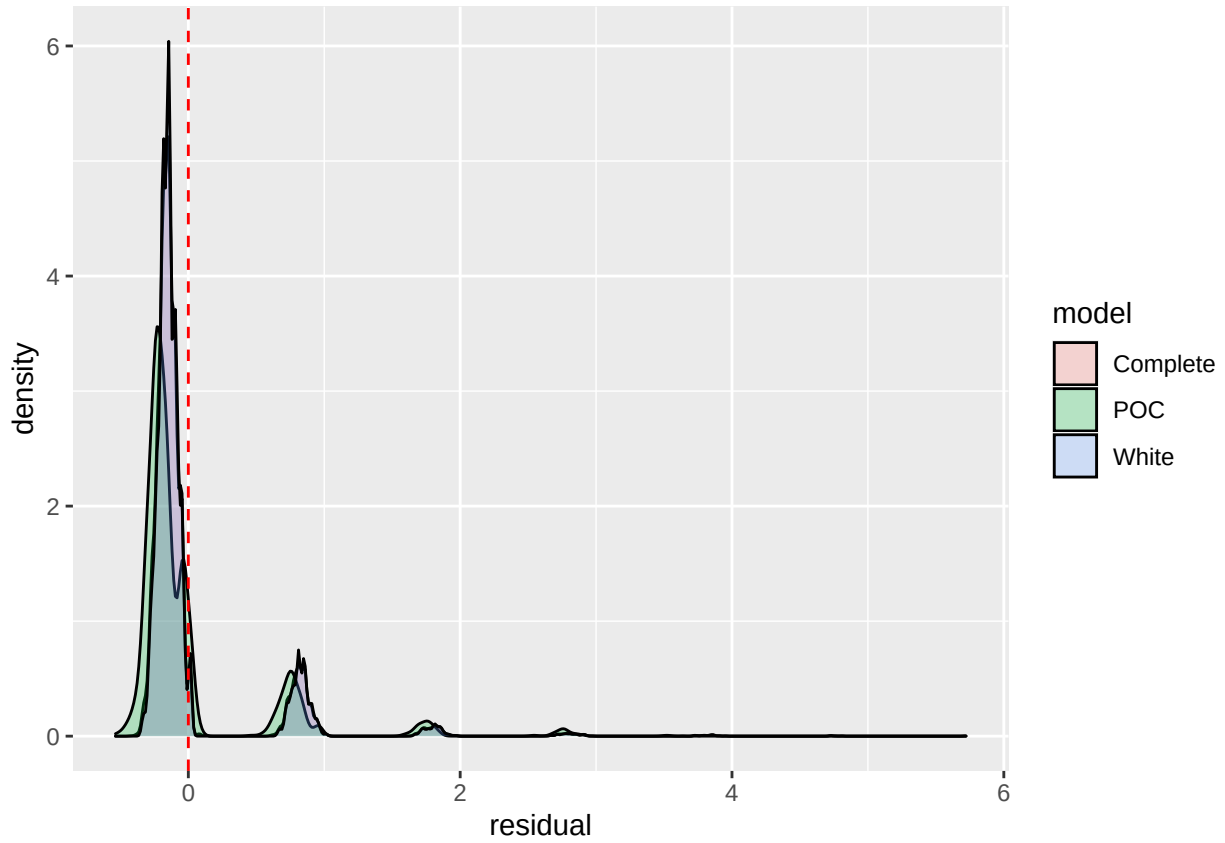
Evaluating the Models

```
print(xtable(regression_r),comment=FALSE,type="latex")
```

	Model	rSquared
1	Full	0.0232181780187883
2	White	0.0218847781538727
3	POC	0.0332267310670332

For the linear models, adjusted r^2 was used to evaluate effectiveness for all three models. Adjusted r^2 was selected as for evaluating models with multiple variables, it is less biased (Agresti, 2018). All models exhibit an extremely weak relationship, with the higher r^2 value being 0.033. Despite the low r^2 values, with all p-values for all models falling far under 0.05, the results are statistically significant, and the null hypothesis is rejected.

```
model_residuals<-cbind(model=character(),residual=double())
comp_residuals<-data.frame(model="Complete",residual=resid(regression_model))
w_residuals<-data.frame(model="White",residual=resid(regression_model_w))
poc_residuals<-data.frame(model="POC",residual=resid(regression_model_poc))
model_residuals<-rbind(comp_residuals,w_residuals,poc_residuals)
ggplot(model_residuals)+
  geom_density(mapping=aes(x=residual,fill=model),alpha=0.25)+
  geom_vline(aes(xintercept=mean(residual)),color="red",linetype="dashed")
```



The residual density graph is an excellent way to visualise model accuracy. The figure once again illustrates the great similarity between the model for the complete data and the model using data from White respondents, the two curves are indistinguishable. All three models have a rightward skew, with an equal number of observations on both sides of zero, this indicates the models appear to under predict by more on average than they over predict by. This skew shows a violation of the assumption of equal variance and a slight violation of normality for all models. Additionally, with the clear presence of residuals with an absolute value greater than 3, there are clearly potential outliers (Agresti, 2018).

GLM

A generalised linear model (GLM) is advantageous for situations where the predicted outcome is not normally distributed, such as for crime experienced (Agresti, 2018). The model predicts a binary result, in this case, whether or not a person has experienced crime in the last twelve months.

```
print(xtable(AIC(glm_model, glm_model_poc, glm_model_w, glm_model_m,
    glm_model_f, glm_model_wf, glm_model_wm, glm_model_pocm,
    glm_model_pocf, glm_model_interaction)), comment = FALSE,
    type = "latex")
```

To evaluate and select models, the AIC of several models was tested, as it is applicable to models assuming non-normal distributions (Agresti, 2018). Lower AICs indicate a better model, also rewarding less variables included (Agresti, 2018). While the models separated by both the ethnic group binary variable and sex have the lowest AIC values, separating by only sex is the next set of most effective models. This is interesting as it didn't appear to have a large impact in the linear regressions conducted above.

	df	AIC
glm_model	12.00	14431.21
glm_model_poc	11.00	1369.14
glm_model_w	7.00	13053.13
glm_model_m	11.00	7350.34
glm_model_f	11.00	7076.21
glm_model_wf	6.00	6437.55
glm_model_wm	6.00	6612.36
glm_model_pocm	7.00	735.68
glm_model_pocf	7.00	654.34
glm_model_interaction	17.00	14433.96

```
print(xtable(summary(glm_model), caption = "Complete Model"),
    comment = FALSE, type = "latex", caption.placement = "top")
```

Table 1: Complete Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9273	0.0904	-10.25	0.0000
W2DVAGE	-0.0184	0.0013	-14.69	0.0000
W2SEX	-0.2712	0.0436	-6.22	0.0000
LIM_bins(0,4]	0.4466	0.0643	6.94	0.0000
LIM_bins(4,9]	0.5414	0.0583	9.28	0.0000
LIM_bins(9,11]	0.6984	0.0973	7.17	0.0000
LIM_bins(11,40]	0.8347	0.0737	11.33	0.0000
ETHREP_bins[-0.5,0.5]	-1.2634	0.2580	-4.90	0.0000
ETHREP_bins[3.5,4.5]	0.4852	0.1596	3.04	0.0024
ETHREP_bins[1.5,2.5]	0.2887	0.2274	1.27	0.2042
ETHREP_bins[4.5,5.5]	0.2573	0.1937	1.33	0.1842
ETHREP_bins[2.5,3.5]	0.2859	0.1024	2.79	0.0052

```
print(xtable(summary(glm_model_m), caption = "Model for Male Respondents"),
    comment = FALSE, type = "latex", caption.placement = "top")
```

```
print(xtable(summary(glm_model_f), caption = "Model for Female Respondents"),
    comment = FALSE, type = "latex", caption.placement = "top")
```

Table 2: Model for Male Respondents

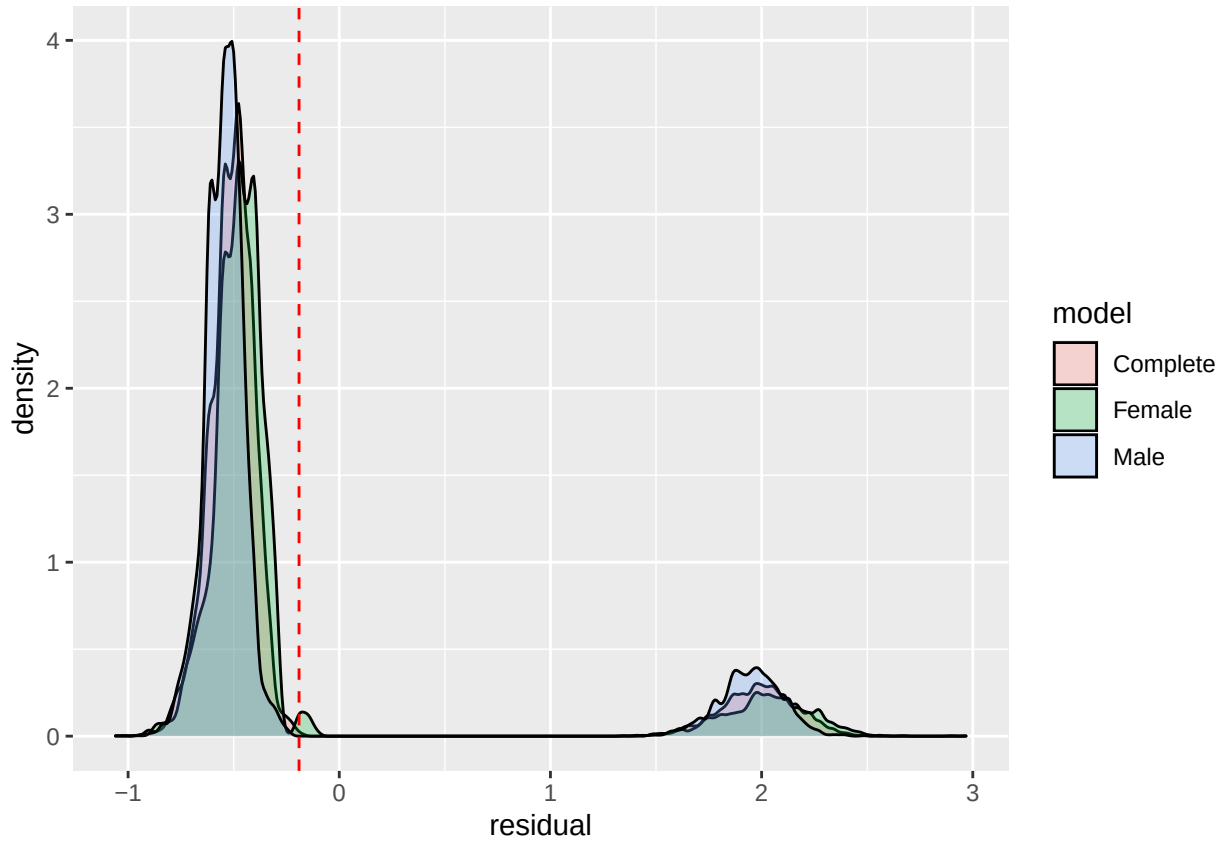
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2722	0.0887	-14.34	0.0000
W2DVAGE	-0.0155	0.0017	-8.90	0.0000
LIM_bins(0,4]	0.3347	0.0894	3.74	0.0002
LIM_bins(4,9]	0.4668	0.0805	5.80	0.0000
LIM_bins(9,11]	0.3456	0.1444	2.39	0.0166
LIM_bins(11,40]	0.6759	0.1051	6.43	0.0000
ETHREP_bins[-0.5,0.5]	-1.0494	0.2790	-3.76	0.0002
ETHREP_bins[3.5,4.5]	0.5338	0.2204	2.42	0.0154
ETHREP_bins[1.5,2.5]	0.0743	0.3537	0.21	0.8336
ETHREP_bins[4.5,5.5]	0.5609	0.2526	2.22	0.0264
ETHREP_bins[2.5,3.5]	0.2192	0.1447	1.51	0.1298

Table 3: Model for Female Respondents

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.3994	0.0934	-14.98	0.0000
W2DVAGE	-0.0216	0.0018	-11.90	0.0000
LIM_bins(0,4]	0.5700	0.0930	6.13	0.0000
LIM_bins(4,9]	0.6304	0.0849	7.43	0.0000
LIM_bins(9,11]	1.0496	0.1330	7.89	0.0000
LIM_bins(11,40]	1.0036	0.1040	9.65	0.0000
ETHREP_bins[-0.5,0.5]	-2.1273	0.7145	-2.98	0.0029
ETHREP_bins[3.5,4.5]	0.4395	0.2318	1.90	0.0580
ETHREP_bins[1.5,2.5]	0.4450	0.2978	1.49	0.1352
ETHREP_bins[4.5,5.5]	-0.1335	0.3106	-0.43	0.6672
ETHREP_bins[2.5,3.5]	0.3540	0.1449	2.44	0.0145

In these models, all variables are statistically significant at 5% with the exception of some ethic groups. The precise ones that are not statistically significant change between the three models. This change in statistical significance could indicate an interaction between ethnicity and sex. However, according to the AIC values, adding an interaction term to the complete model actually creates a worse model. This is likely due to the significance increase in degrees of freedom without a large improvement to model prediction.

```
glm_model_residuals<-cbind(model=character(),residual=double())
comp_residuals<-data.frame(model="Complete",residual=resid(glm_model))
m_residuals<-data.frame(model="Male",residual=resid(glm_model_m))
f_residuals<-data.frame(model="Female",residual=resid(glm_model_f))
model_residuals<-rbind(comp_residuals,m_residuals,f_residuals)
ggplot(model_residuals)+
  geom_density(mapping=aes(x=residual,fill=model),alpha=0.25)+
  geom_vline(aes(xintercept=mean(residual)),color="red",linetype="dashed")
```



Unfortunately, it is clear that the trend observed in the linear regression has remained in this model as well. With the skew of the residual density graph, it is clear that the assumptions of equal variance, as well as normality have been violated.

Conclusions

In conclusion, it is clear that all variables studied influence crime experienced in a statistically significant way. Sex and ethnicity influence the way other variables impact crime experienced. In the linear models for example, being a woman decreased predicted crime experienced by more for respondents of colour than it did for white respondents. These interactions certainly highlight the intersectionality of issues like crime. Limitations themselves, while statistically significant, did not appear to be particularly practically significant. However, as most studies of disability and crime focus on developmental disabilities, separating the limitations studied may give different results. The violation of equal variance and normality assumptions are concerning for the models produced and they make it difficult to discuss the real relationship of the explanatory variables and crime experienced. But the p-values found in every statistical test conducted, make it clear that it exists.

References

Agresti, A. (2018). Statistical methods for the social sciences. Harlow, United Kingdom: Pearson Education Limited.

Fisher, M. H., Baird, J. V., Currey, A. D. & Hodapp, R. M. (2016). 'Victimisation and social vulnerability of adults with intellectual disability: A review of research extending beyond wilson and brewer' Australian Psychologist, 51 (2). DOI: 10.1111/ap.12180.

Office for National Statistics, S. S. D. (2016). 'Life opportunities survey: Waves 1-3, 2009-2014' (Version 5th Edition). Available at: http://doi.org/10.5255/UKDA-SN-6653-4.

Office of National Statistics (2013). Life opportunities survey qmi. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/disability/bulletins/disabilityandcrimeuk/2019 (Accessed: 12 May).

Petersilia, J. R. (2001). 'Crime victims with developmental disabilities' Criminal Justice and Behaviour, 28 pp. 655-694. DOI: https://doi.org/10.1177/009385480102800601.

Appendix

Data Cleaning

```
# Loads full STATA dataset
alldata <- haven::read_dta(file_path_personal)
# Creates consistent column name capitalization
colnames(alldata) <- toupper(names(alldata))
# Creates a function subtracting a given value from 6, used
# to recode limitation vars
minus <- function(x) 6 - x
larger_data <- select(alldata, c("W2SEELIM", "W2HEARLIM", "W2SPKLIM",
    "W2MOBLIM", "W2PAINLIM", "W2CONDLIM", "W2BRETHLIM", "W2LRNLIM",
    "W2INTELLIM", "W2BEVLIM", "W2MEMLIM", "W2MENLIM", "W2DEXLIM",
    "W2OTHLIM", "W2CRIM01", "W2CRIM02", "W2CRIM03", "W2CRIM04",
    "W2CRIM05", "W2CRIM06", "W2HCRIM", "W2SEX", "W2DVAGE", "W1ETHREP",
    "W2ETHREP", "W2HHLDDV", "W2CASE"))
# Removes W1 rows from the data set
larger_data = larger_data[larger_data$W2CASE != -9, ]
# These three mutations replace negative values, indicating
# non answers, with NA values
final_data <- larger_data %>%
    mutate(across(everything(), ~na_if(., -9))) %>%
    mutate(across(everything(), ~na_if(., -8))) %>%
    mutate(across(everything(), ~na_if(., -7))) %>%
    # Recoding limitation vars so 5 is the higher amount of
    # limitation, makes sums later make sense
mutate_at(vars(W2SEELIM, W2HEARLIM, W2SPKLIM, W2MOBLIM, W2PAINLIM,
    W2CONDLIM, W2BRETHLIM, W2LRNLIM, W2INTELLIM, W2BEVLIM, W2MEMLIM,
    W2MENLIM, W2DEXLIM, W2OTHLIM), funs(. = minus(.))) %>%
    # Crime questions were only asked for respondents older
    # than 15, filtering for that
filter(W2DVAGE > 15) %>%
    # Forming bins for age
mutate(W2DVAGE_bins = cut(W2DVAGE, breaks = c(0, 34, 48, 60,
    71, 80))) %>%
    # Forms groups for each individual row, enables row
    # sums
rowwise() %>%
    # Ethnicity was only recorded once, either during wave
    # one or wave two, they were combined into a single
    # variable
mutate(ETHREP = sum(W1ETHREP, W2ETHREP, na.rm = TRUE)) %>%
    # Forms new columns CRIMEXP, which sums crime columns,
    # and LIM, which sums limitation columns.
mutate(CRIMEXP = sum(c(W2CRIM01, W2CRIM02, W2CRIM03, W2CRIM04,
    W2CRIM05, W2CRIM06, W2HCRIM), na.rm = TRUE), LIM = sum(c(W2SEELIM,
    W2HEARLIM, W2SPKLIM, W2MOBLIM, W2PAINLIM, W2CONDLIM, W2BRETHLIM,
    W2LRNLIM, W2INTELLIM, W2BEVLIM, W2MEMLIM, W2MENLIM, W2DEXLIM,
    W2OTHLIM), na.rm = TRUE)) %>%
    # Forming bins for limitations
mutate(LIM_bins = cut(LIM, breaks = c(-1, 0, 4, 9, 11, 40))) %>%
    # Creating binary variables for all variables (except
```

```
    # sex) Divided into experienced crime Y(1)/N(0)
mutate(CRIMEXP_binary = ifelse(CRIMEXP == 0, 0, 1)) %>%
    # White(0) or Unknown/POC (1)
mutate(ETHREP_binary = ifelse(ETHREP == 1, 0, 1)) %>%
    # Limitation Y(1)/N(0)
mutate(LIM_binary = ifelse(LIM == 0, 0, 1)) %>%
    # Divided by median value up to and including 55(0),
    # older than 55(1)
mutate(AGE_binary = ifelse(W2DVAGE <= 55, 0, 1)) %>%
    select(W2SEX, W2DVAGE, ETHREP, W2DVAGE_bins, W2HHLDDV, LIM,
        LIM_bins, CRIMEXP, W2CASE, CRIMEXP_binary, LIM_binary,
        ETHREP_binary, AGE_binary)
```

Odds Ratio

```
# Creates blank data frame to track odds ratios
Odds_df <- data.frame(Variable = character(), Odds = double())
# Age
a <- length(final_data$AGE_binary[final_data$AGE_binary == 1 &
    final_data$CRIMEXP_binary == 1])
b <- length(final_data$AGE_binary[final_data$AGE_binary == 1 &
    f, inal_data$CRIMEXP_binary == 0])
c <- length(final_data$AGE_binary[final_data$AGE_binary == 0 &
    final_data$CRIMEXP_binary == 1])
d <- length(final_data$AGE_binary[final_data$AGE_binary == 0 &
    final_data$CRIMEXP_binary == 0])
Odds_df[nrow(Odds_df) + 1, ] = c("Age", ((a/b)/(c/d)))

# Ethnicity
a <- length(final_data$ETHREP_binary[final_data$ETHREP_binary ==
    1 & final_data$CRIMEXP_binary == 1])
b <- length(final_data$ETHREP_binary[final_data$ETHREP_binary ==
    1 & final_data$CRIMEXP_binary == 0])
c <- length(final_data$ETHREP_binary[final_data$ETHREP_binary ==
    0 & final_data$CRIMEXP_binary == 1])
d <- length(final_data$ETHREP_binary[final_data$ETHREP_binary ==
    0 & final_data$CRIMEXP_binary == 0])
Odds_df[nrow(Odds_df) + 1, ] = c("Ethnicity", ((a/b)/(c/d)))

# Sex
a <- length(final_data$W2SEX[final_data$W2SEX == 2 & final_data$CRIMEXP_binary ==
    1])
b <- length(final_data$W2SEX[final_data$W2SEX == 2 & final_data$CRIMEXP_binary ==
    0])
c <- length(final_data$W2SEX[final_data$W2SEX == 1 & final_data$CRIMEXP_binary ==
    1])
d <- length(final_data$W2SEX[final_data$W2SEX == 1 & final_data$CRIMEXP_binary ==
    0])
Odds_df[nrow(Odds_df) + 1, ] = c("Sex", ((a/b)/(c/d)))

# Lim
a <- length(final_data$LIM_binary[final_data$LIM_binary == 1 &
```

```
    final_data$CRIMEXP_binary == 1])
b <- length(final_data$LIM_binary[final_data$LIM_binary == 1 &
    final_data$CRIMEXP_binary == 0])
c <- length(final_data$LIM_binary[final_data$LIM_binary == 0 &
    final_data$CRIMEXP_binary == 1])
d <- length(final_data$LIM_binary[final_data$LIM_binary == 0 &
    final_data$CRIMEXP_binary == 0])
Odds_df[nrow(Odds_df) + 1, ] = c("Limitation", ((a/b)/(c/d)))

print(xtable(Odds_df), comment = FALSE, type = "latex")
```

Chi Squared

```
# Create Blank Dataframe
chi_test <- data.frame(Variables = character(), Statistic = double(),
    df = double(), pvalue = character())

# Age
a <- chisq.test(final_data$CRIMEXP_binary, final_data$AGE_binary)
chi_test[nrow(chi_test) + 1, ] = c("Age", a[1], a[2], a[3])

# Ethnicity
a <- chisq.test(final_data$CRIMEXP_binary, final_data$ETHREP_binary)
chi_test[nrow(chi_test) + 1, ] = c("Ethnicity", a[1], a[2], a[3])

# Sex
a <- chisq.test(final_data$CRIMEXP_binary, final_data$W2SEX)
chi_test[nrow(chi_test) + 1, ] = c("Sex", a[1], a[2], a[3])

# Limitations
a <- chisq.test(final_data$CRIMEXP_binary, final_data$LIM_binary)
chi_test[nrow(chi_test) + 1, ] = c("Limitations", a[1], a[2],
    a[3])

print(xtable(chi_test), comment = FALSE, type = "latex")
```

GLM

```
glm_model <- glm(CRIMEXP_binary ~ W2DVAGE + W2SEX + LIM_bins +
    ETHREP_bins, data = final_data, family = "binomial")

glm_model_w <- glm(CRIMEXP_binary ~ W2DVAGE + W2SEX + LIM_bins,
    data = filter(final_data, ETHREP_binary == 0), family = "binomial")
glm_model_poc <- glm(CRIMEXP_binary ~ W2DVAGE + W2SEX + LIM_bins +
    ETHREP_bins, data = filter(final_data, ETHREP_binary == 1),
    family = "binomial")

glm_model_m <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins + ETHREP_bins,
    data = filter(final_data, W2SEX == 1), family = "binomial")
```

```
glm_model_f <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins + ETHREP_bins,
    data = filter(final_data, W2SEX == 2), family = "binomial")

glm_model_wf <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins + ETHREP,
    data = filter(final_data, W2SEX == 2 & ETHREP_binary == 0),
    family = "binomial")
glm_model_wm <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins + ETHREP,
    data = filter(final_data, W2SEX == 1 & ETHREP_binary == 0),
    family = "binomial")
glm_model_pocf <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins + ETHREP,
    data = filter(final_data, W2SEX == 2 & ETHREP_binary == 1),
    family = "binomial")
glm_model_pocm <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins + ETHREP,
    data = filter(final_data, W2SEX == 1 & ETHREP_binary == 1),
    family = "binomial")

glm_model_interaction <- glm(CRIMEXP_binary ~ W2DVAGE + LIM_bins +
    W2SEX * ETHREP_bins, data = final_data, family = "binomial")
```