Homework 3

PA 541

Alexis Kwan

Part 1

Variable name (Description)

- code (School)
- county (County name)
- name (School name)
- type (School type (public or private))
- district (School district name)
- city (City location)
- enrollment (Total kindergarten enrollment)
- pbe_pct (Rounded percent of personal belief exemptions)
- exempt (Percent of students exempt from providing vaccination records)
- med_exempt (Percent exempt due to medical reasons)
- rel_exempt (Percent exempt due to religious reasons)

Question 1

1a

In how many schools is the percentage of students exempt for medical reasons (med_exempt) greater than the percentage exempt for religious reasons (rel_exempt)? [2 pts] Of this set of schools, what percent are public schools? [2 pts]

```
all_schools <- cavax %>%
  filter(med_exempt > rel_exempt) %>%
  count()
all_schools
```

```
##     n
## 1 518
```

There are 518 schools where the percentage of medically exempt students is greater than the percentage of religiously exempt students.

```
public_schools <- cavax %>%
  filter(med_exempt > rel_exempt & type == "PUBLIC") %>%
  count()
public_schools / all_schools
```

```
##           n
## 1 0.8648649
```

86% of the schools are public schools.

1b

Which county, when averaging across all the schools in that county, has the highest average percentage of exempt students (exempt)? Note, we are using the variable exempt here. [2 pts]

```
cavax %>%
  group_by(county) %>%
  summarise(mean_exempt = mean(exempt)) %>%
  arrange(desc(mean_exempt))
```

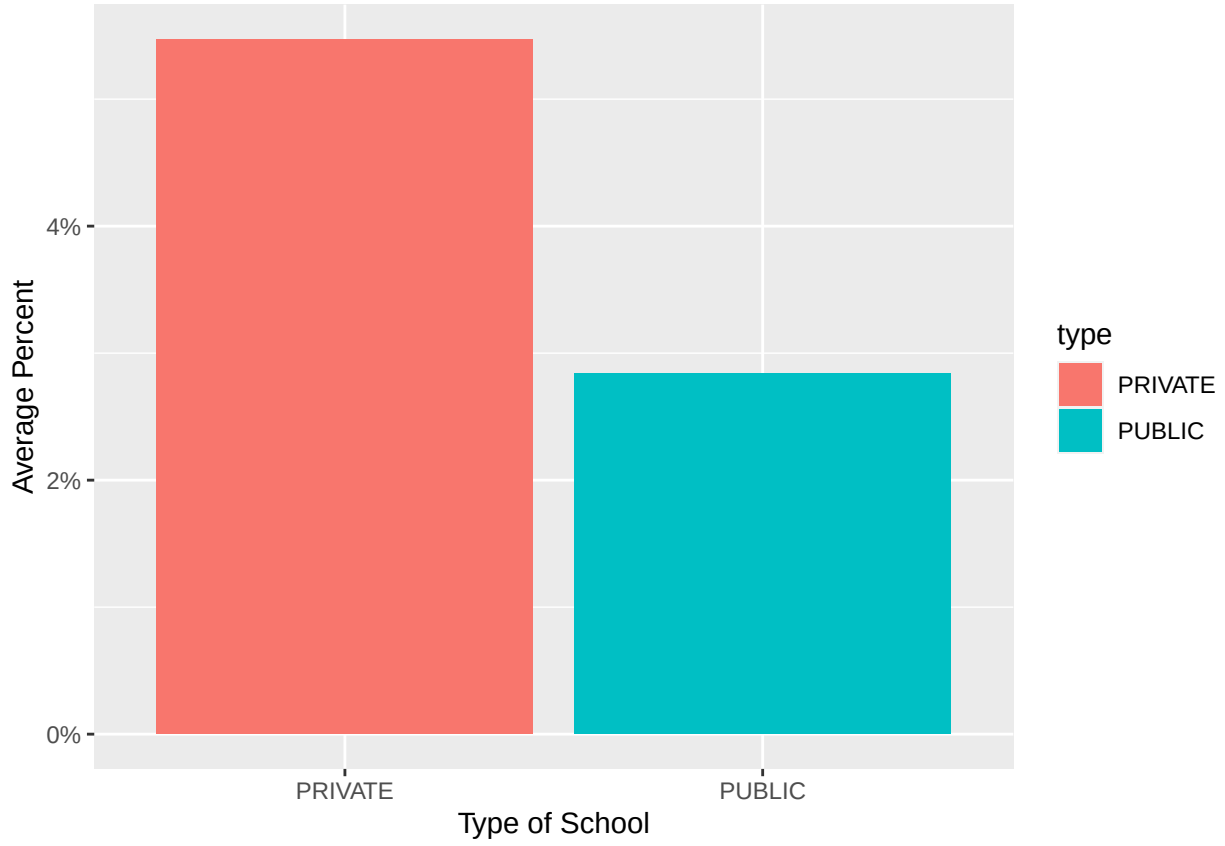
```
## # A tibble: 57 x 2
##    county     mean_exempt
##    <chr>            <dbl>
##  1 NEVADA            24.0
##  2 MARIPOSA          16.5
##  3 HUMBOLDT          14.2
##  4 DEL NORTE         13.2
##  5 LASSEN            12.5
##  6 SANTA CRUZ        11.8
##  7 EL DORADO         10.6
##  8 CALAVERAS          9.80
##  9 SHASTA             9.78
## 10 MENDOCINO          9.37
## # ... with 47 more rows
```

Nevada county seems to have the highest average percent of students exempt at 24%.

1c

Create a bar chart that shows for private and public schools (type) the percent of students exempt from providing vaccination records (exempt). [2 pts]

```
cavax %>%
  group_by(type) %>%
  summarise(mean_exempt = mean(exempt)) %>%
  ggplot(aes(x = type, y = mean_exempt, fill = type)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1, scale = 1)) +
  ylab("Average Percent") +
  xlab("Type of School")
```



Question 2

- Estimate a model predicting exempt (exempt) by district type (type) and enrollment (enrollment). [2 pts]
- Treat exempt as a continuous variable (and thus you can use standard OLS). Interpret the intercept and the coefficients. [4 pts]
- What is the predicted exempt percentage for a public school with 100 students in kindergarten? [2 pts]
- What is the predicted exempt percentage for a private school with 80 students in kindergarten? [2pts]

```
cavax_lm <- lm(exempt ~ type + enrollment, data = cavax)
summary(cavax_lm)
```

```
##
## Call:
## lm(formula = exempt ~ type + enrollment, data = cavax)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.993 -3.296 -1.727  0.588 86.986
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.288602   0.209140  30.069  < 2e-16 ***
## typePUBLIC  -0.860822   0.263483  -3.267  0.00109 **
## enrollment  -0.029606   0.002383 -12.425  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7.396 on 7029 degrees of freedom
## Multiple R-squared:  0.04034,    Adjusted R-squared:  0.04007
## F-statistic: 147.7 on 2 and 7029 DF,  p-value: < 2.2e-16
```

When the school is a private school and there are no enrolled students, there are on average 6.29% of students exempted from providing vaccination records. The average difference in percent of exemptions between public and private schools is 0.86%, with private schools having the larger percentage of exemptions. Every additional enrollment predicts on average a decrease of 0.02 percentage points of exemptions. Also every coefficient is statistically significant with p-values below a critical value of 0.01, meaning we can reject the null hypothesis they are zero.

Question 3

- Test whether the assumption of homoskedasticy has been met. [2 pts]
- Discuss results. [2 pts]
- Calculate the VIF for each variable. [2 pts]
- Should we be concerned with multicollinearity. [2 pts]
  - (Note, the vif() command is in the 'car' package. You can also calculate the VIF yourself as we did in class)

Testing homoskedasticy with a squared residual term:

```
cavax$residsq <- resid(cavax_lm)**2
cavax_resid_lm <- lm(residsq ~ type + enrollment, data = cavax)
summary(cavax_resid_lm)
```

```
##
## Call:
## lm(formula = residsq ~ type + enrollment, data = cavax)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -125.9  -57.9  -37.1  -17.8 7439.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 136.5528     9.1287  14.959  < 2e-16 ***
## typePUBLIC  -39.2252    11.5007  -3.411 0.000652 ***
## enrollment   -0.6662     0.1040  -6.406 1.59e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322.8 on 7029 degrees of freedom
## Multiple R-squared:  0.01512,    Adjusted R-squared:  0.01484
## F-statistic: 53.95 on 2 and 7029 DF,  p-value: < 2.2e-16
```

The F-statistic on the regression of squared residuals is clearly significant here with a very small p-value that is less than 2.2e-16 and we fail to reject the null hypothesis that the model is homoskedastic. This means that the independent variables have a correlation with the residual and therefore violates the homoskedacity assumption for the regression model.

Testing multicollinearity with VIF:

```
vif(cavax_lm)
```

```
##       type enrollment
##   1.413965   1.413965
```

```r
sqrt(vif(cavax_lm))
```

```
##       type enrollment
##   1.189103   1.189103
```

We see that with the square root of the VIF, about $\sqrt{1.41} \approx 1.19$ for each variable, that there is a slight inflation of standard errors due to multicollinearity. Multicollinearity does not seem to be issue since the square root of the VIF value is relatively small.

Question 4

- Recenter the variable enrollment at its mean. [2 pts]
- Create an interaction between type (type) and student enrollment (enrollment) recentered and rerun the model predicting exempt (exempt). [2 pts]
- Assume that type moderates the effect of enrollment in your interpretation of the interaction.
  - (i) Interpret the results on each coefficient.
  - (ii) Create a plot to visualize the interaction. [4 pts]

```r
cavax$enrollment_ctr <- cavax$enrollment - mean(cavax$enrollment)
cavax_lm4 <- lm(exempt ~ type + enrollment_ctr + type*enrollment_ctr, data = cavax)
summary(cavax_lm4)
```

```
##
## Call:
## lm(formula = exempt ~ type + enrollment_ctr + type * enrollment_ctr,
##     data = cavax)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.070 -3.293 -1.726  0.587 86.926
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.837202   0.600435   6.391 1.76e-10 ***
## typePUBLIC                 -0.651213   0.609131  -1.069  0.28507
## enrollment_ctr             -0.034017   0.011800  -2.883  0.00395 **
## typePUBLIC:enrollment_ctr   0.004599   0.012048   0.382  0.70272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.396 on 7028 degrees of freedom
## Multiple R-squared:  0.04036,    Adjusted R-squared:  0.03995
## F-statistic: 98.53 on 3 and 7028 DF,  p-value: < 2.2e-16
```

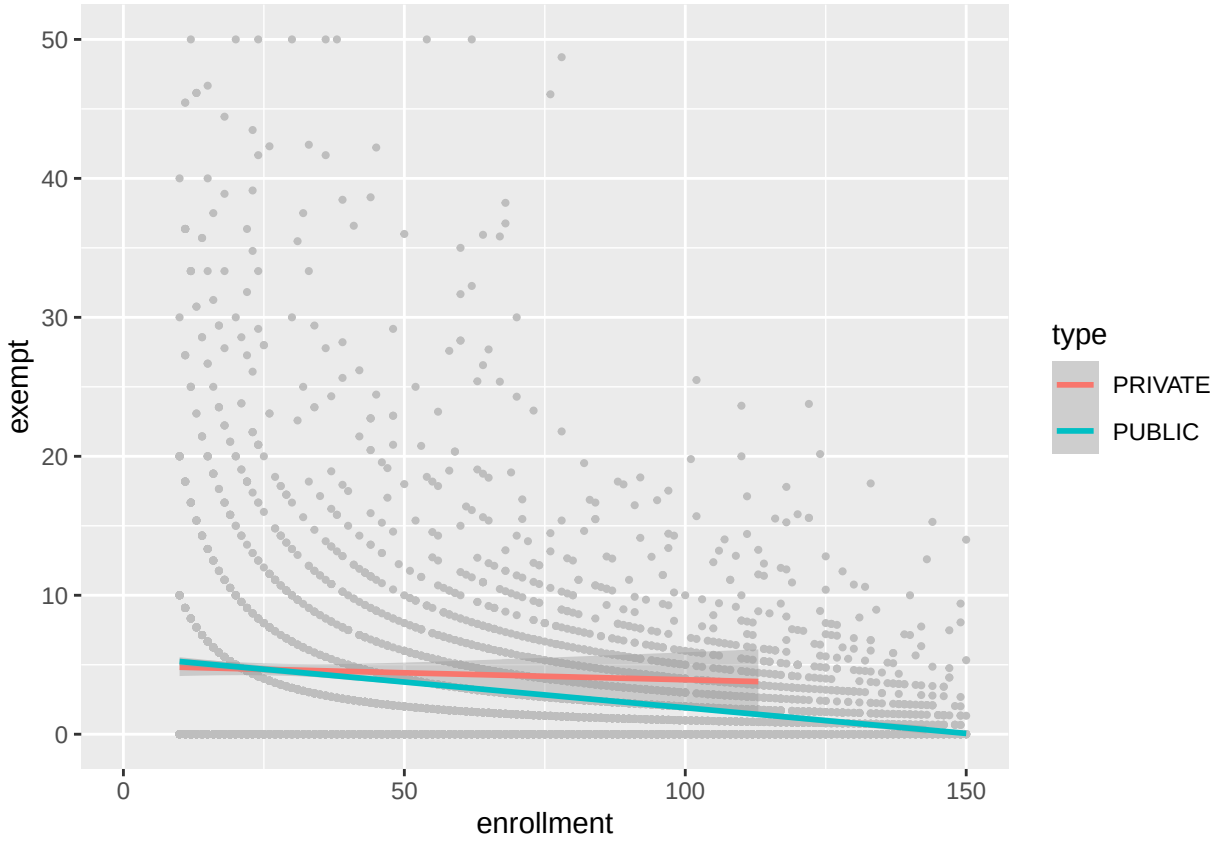
With average total enrollment, a private school has about 3.84% exemptions from providing vaccine records. Private and public schools differ by -0.65% in exemptions conditional on enrollment being average. However this difference is not statistically significant by any good signifcance level. An additional unit of enrollment contributes a decrease of 0.03% in exemptions conditional on the school being private or a simple main effect. Public schools adds an additional 0.005% increase on top of -0.03% on every additional unit of enrollment such that for every additional enrollment for public schools there is a 0.02% decrease in exemptions.

```r
cavax %>%
  ggplot(aes(x = enrollment, y = exempt, color = type)) + geom_point(size = 0.75, color = "grey") + geo
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 405 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 405 rows containing missing values (geom_point).
```



Question 5

- Let's log transform (using the natural log) the variable enrollment and call the new variable log_enroll. [2pts]

- Estimate a model predicting exempt (exempt) by district type (type) and log of enrollment (log_enrollment). [2pts]
- Interpret the coefficient on the log of enrollment. [2 pts]
- Does it make more sense to use enrollment or the log of enrollment as the predictor variable? Why? [4 pts]

```
cavax$log_enrollment <- log(cavax$enrollment)
cavax_lm5 <- lm(exempt ~ type + log_enrollment, data = cavax)
summary(cavax_lm5)
```

```
##
## Call:
## lm(formula = exempt ~ type + log_enrollment, data = cavax)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.417 -2.982 -1.461  0.867 85.934
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     14.1408     0.5410  26.137   <2e-16 ***
## typePUBLIC       0.5707     0.2878   1.983   0.0474 *
## log_enrollment  -2.7338     0.1589 -17.201   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.324 on 7029 degrees of freedom
## Multiple R-squared:  0.05888,    Adjusted R-squared:  0.05861
## F-statistic: 219.9 on 2 and 7029 DF,  p-value: < 2.2e-16
```

When a school is private and there is no enrollment, there is on average an exemption rate of 14%. Public schools differ from private schools by about 0.57% in exemptions of students. Every one percent increase in enrollment results in a 0.027 unit decrease in exemptions. All of the coefficients are statistically significant at significance levels of 0.05 and lower. Log of enrollment seems to make more sense since the shape of the distribution looks more exponential-like and the R-square value increased for this model compared to the non-log model, while still maintaining statistical significance.

Question 6

- Create a binary variable to indicate high versus low exempt rates. For schools with exempt percentages equal to or greater than 33 percent, indicate them as "high", for all other schools indicate them as "low". [2 pts]
- Run a logistic regression predicting whether a school is high versus low, in other words, we want our model to predict schools falling into the high category. In your model use the predictors of school type (type) and enrollment (enrollment) (note: do not use log_enroll in this model). [2 pts]
- Interpret the coefficients on type and enrollment in terms of both log odds and odds. [6 pts]
- What is the probability of being a high exempt school if the school is private and has 100 students enrolled? [2 pts]

```
cavax$binary_exempt <- ifelse(cavax$exempt >= 33, 1, 0)
head(cavax$binary_exempt)
```

```
## [1] 0 0 0 0 0 0
```

```
cavax_glm <- glm(binary_exempt ~ type + enrollment, family=binomial(link="logit"), data = cavax)
summary(cavax_glm)
```

```
##
## Call:
## glm(formula = binary_exempt ~ type + enrollment, family = binomial(link = "logit"),
##     data = cavax)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3496  -0.2117  -0.1293  -0.0801   4.5249
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.569319   0.174010 -14.765  < 2e-16 ***
## typePUBLIC   0.115309   0.232708   0.496     0.62
## enrollment  -0.031009   0.004197  -7.388 1.49e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1108.0  on 7031  degrees of freedom
## Residual deviance: 1005.3  on 7029  degrees of freedom
## AIC: 1011.3
##
## Number of Fisher Scoring iterations: 8
```

```
exp(coef(cavax_glm))
```

```
## (Intercept)  typePUBLIC  enrollment
##  0.07658769  1.12222035  0.96946696
```

Public status of schools decreases the log odds of high exemption by about 0.12 units as opposed to private status. Every additional unit of enrollment decreases the log odds of there being high percentage of exemptions by about 0.03 units. At the intercept or base level, a private school with 0 enrollment has logit odds of -2.57 versus nonzero enrollment and nonprivate school status. Every coefficient is statistically significant except for for the coefficient for public vs. private, which is not significant under any reasonable threshold.

Public status of schools increases the odds of high exemption percentage by 0.89 times versus private schools. For every additional unit of enrollment, the odds of high exemptions increases by 1.03 times. In the base case, private schools with no enrollment will have 13.06 times greater odds of high exemptions, versus schools with nonzero enrollment or public school.

```
# school is private and has 100 students
s1=as.matrix(c(1,0,100))
# probability
1/(1+exp(-crossprod(s1, coef(cavax_glm))))
```

```
##             [,1]
## [1,] 0.003435308
```

The probability that a private school with 100 students has a high percentage of exemptions is about 0.3%.

Part 2

Question 7

Variable name (Description)

- stateid (State id)
- statename (Name of state)
- shall (Equals 1 if state has concealed carry on that year and 0 if not)
- year (Year)
- vio (Violent crime rate per 100,000 people)
- mur (Murder rate per 100,000 people)

7a

Let's begin by exploring the data. How many years are there in concealed_carry data? How many observations per state? [2 pts]

```
concealed_carry <- read.csv("concealed_carry.csv")
length(unique(concealed_carry$year))
```

```
## [1] 23
```

23 different years are represented in the concealed_carry data set.

```
concealed_carry %>%
  group_by(statename) %>%
  count()
```

```
## # A tibble: 51 x 2
## # Groups:   statename [51]
##    statename                n
##    <chr>                <int>
##  1 Alabama                 23
##  2 Alaska                  23
##  3 Arizona                 23
##  4 Arkansas                23
##  5 California              23
##  6 Colorado                23
##  7 Connecticut             23
##  8 Delaware                23
##  9 District of Columbia    23
## 10 Florida                 23
## # ... with 41 more rows
```

Each state also has 23 observations, one for each year.

7b

How many states had concealed carry laws (shall) in 1977 and how many had concealed carry laws in 1999? [4pts]

```
concealed_carry %>%
  filter(shall == 1 & year == 1977) %>%
  nrow()
```

```
## [1] 4
```

```
concealed_carry %>%
  filter(shall == 1 & year == 1999) %>%
  nrow()
```

```
## [1] 29
```

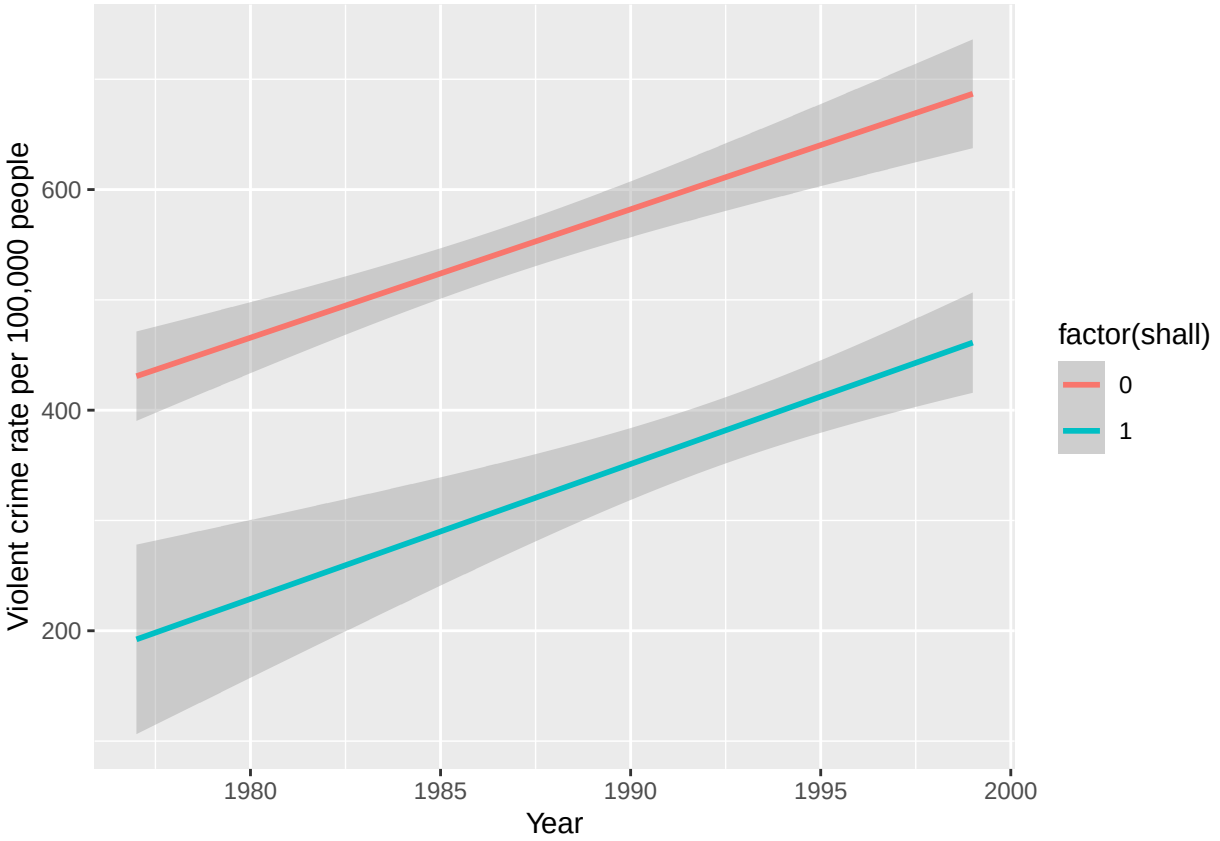
4 states in 1977 and 29 in 1999.

7c

Create a plot tracking the violent crime rate (vio) over time for states that have ever adopted conceal carry laws (shall) and those that have never adopted the law. [4 pts]

```
ggplot(data=concealed_carry, aes(x = year, y = vio, color = factor(shall))) +
  geom_smooth(method = "lm") +
  xlab("Year") +
  ylab("Violent crime rate per 100,000 people")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Question 8

- Convert the violent crime rate (vio) into a logged variable (using the natural log), call it log_vio. [2 pts]
- This will be our dependent variable. Run a pooled regression of the data (i.e., standard OLS model as if this was cross-sectional data) predicting the log of violent crimes (log_vio) as a function of the presence of concealed carry laws (shall) and a set of dummy variables for year. [2 pts]
- Interpret the effect of shall. [2 pts]
- In general terms, what do the year dummy variables tell us about crime trends? [2 pts]
- In our current specification of the model, is the effect of shall the same for all years? Why or why not? [2 pts]

```
concealed_carry$log_vio <- log(concealed_carry$vio)
concealed_carry$year <- as.factor(concealed_carry$year)
concealed_lm <- lm(log_vio ~ shall + year, data = concealed_carry)
summary(concealed_lm)
```

```
##
## Call:
## lm(formula = log_vio ~ shall + year, data = concealed_carry)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14538 -0.43698  0.05355  0.40556  1.68492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.86727    0.08372  70.086  < 2e-16 ***
```

```
## shall       -0.59812    0.04446 -13.452  < 2e-16 ***
## year1978     0.04838    0.11829   0.409 0.682589
## year1979     0.14372    0.11829   1.215 0.224631
## year1980     0.18781    0.11829   1.588 0.112634
## year1981     0.17746    0.11829   1.500 0.133838
## year1982     0.14767    0.11829   1.248 0.212166
## year1983     0.08999    0.11829   0.761 0.446945
## year1984     0.10077    0.11829   0.852 0.394470
## year1985     0.12826    0.11829   1.084 0.278473
## year1986     0.20558    0.11832   1.738 0.082559 .
## year1987     0.19364    0.11834   1.636 0.102056
## year1988     0.24561    0.11837   2.075 0.038211 *
## year1989     0.28168    0.11837   2.380 0.017490 *
## year1990     0.41694    0.11849   3.519 0.000451 ***
## year1991     0.48647    0.11868   4.099 4.44e-05 ***
## year1992     0.52834    0.11883   4.446 9.59e-06 ***
## year1993     0.53923    0.11883   4.538 6.28e-06 ***
## year1994     0.51493    0.11883   4.333 1.60e-05 ***
## year1995     0.54706    0.11921   4.589 4.94e-06 ***
## year1996     0.54408    0.11983   4.540 6.21e-06 ***
## year1997     0.55553    0.12028   4.619 4.30e-06 ***
## year1998     0.49975    0.12028   4.155 3.50e-05 ***
## year1999     0.44010    0.12028   3.659 0.000265 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5973 on 1149 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.1442
## F-statistic: 9.585 on 23 and 1149 DF,  p-value: < 2.2e-16
```

Those states that allow concealed carry have on average 59.8% lower violent crime rate than those without concealed carry. This coefficient is highly statistically significant. The dummy variables tell us that crime is generally tending upwards as time goes on, every year after 1977 has a coefficient great than or about equal to the coefficient for the previous year. Also only the coefficients for the later years seem to be significant. Since there is no interaction term, it also means that we assume that the effect of "shall" is the same for all years.