Final Exam

Please read all of the questions and tables carefully and follow all instructions. Each question has an allotted point value. Be as thorough and specific as possible; extra calculations and incorrect information, even in the presence of correct information, will result in point dedu ctions. Be sure to show all formulas and all necessary work to answer the questions. You may upload your completed exam to the Elms course site.

> **Note**
>
> While this is an open-note exam, you are not to receive assistance from anyone else. As usual, the Honor Code applies.

Total points: 50

Conceptual Questions

Please include all work (and computations) necessary to answer the questions.

Total points: 20 (2 points each)

Question 1

Suppose you estimate a bivariate regression model with a total sample size of 50 and obtain a standard error for β_1 of 2.50. What is the smallest regression coefficient (β_1) that you could have and still be able to reject the null hypothesis (of no relationship between X_1 and Y) at the 0.05 level with a one-tailed test?

Ans 1:

To do this we can use the T-Statistic value:

We know that:

$$T\_Statisic = BetaCoef/StandardError$$

Degree of freedom = 50-2 = 48

Critical value for 48 degrees of freedom for 0.05 and a single tailed test = 1.68 approx

Standard Error = 2.50

Hence plugging in the values, the smallest value will be: 4.19 as a coefficient.

Question 2

I regress Y on X_1 (i.e., $Y = \beta_0 + \beta_1 X_1$) and find that $\beta_1 = 4.20$, $SE_{\beta_1} = 1.90$, $t = 2.21$, $p = 0.01$, and a 95% confidence interval is [0.48, 7.92]. What is my best estimate of the effect of a two-unit change in X on the mean of Y in the population?

Ans 2:

To estimate the effect of a two-unit change in X1 on the mean of Y in the population, we can use the estimated coefficient Beta1 from the regression model. The regression coefficient Beta1 represents the change in the dependent variable Y for a one-unit change in the independent variable X1.

Given that Beta = 4.20, a two-unit change in X1 would result in a change of 2×Beta1 in Y.

Plugging in values:

The best estimate of the effect of a two-unit change in X1 on the mean of Y in the population is 8.48.4. This means that, according to our regression model, a two-unit increase in X1 is associated with an average increase of 8.4 units in Y.

Question 3

I conduct an OLS regression with a sample size of 90 and 5 independent variables. To determine a p-value for each coefficient, I would examine a t-distribution with how many degrees of freedom?

Ans 3:

The usual convention is to take the number of independent variables and subtract it from the total sample size. So if there are 5 independent variables along with an intercept, the total number of degrees of freedom would be:

df = 90 - (5+1)

df = 84

So we would examine a t-distribution with 84 degrees of freedom totally.

Question 4

Suppose I regress Y on X and compute the mean response for Y at some specified value of X. When determining the confidence interval around this mean response, which of the following will **NOT** have any effect on the width (or, size) of that confidence interval?

A. Total sample size,

B. Mean-squared error,

C. The specified value of X,

D. None of the above: all of these (above) will affect the confidence interval,

E. There is not enough information to answer this question.

Ans 4:

1. Total sample size can affect confidence intervals as the larger the number of data points the narrower the confidence intervals becomes, i.e. the more accurate readings you can get
2. Larger Mean-Square error usually translates to wider confidence intervals indicating lower accuracy. There are some exceptions to this but it largely holds good.
3. Value of X depending on how far it is from the mean of the series will have an effect on the interval. Broadly speaking the further it is from the mean the larger the confidence interval will be.

So the final answer will be:

D. None of the above: all of these (above) will affect the confidence interval

Question 5

I regress Y on X and find that β_1 has a two-tailed p-value of 0.04. Which of the following statements is most accurate?

A. The lower and upper bound of a 95% confidence interval around β_1 will have the same sign,

B. The absolute value for the t-statistic for β_1 will be greater than 1.96,

C. A 90% confidence interval around β_1 will not contain zero,

D. All of the above,

E. None of the above and/or there is not enough information.

Ans 5:

1. For very small sample sizes it could be possible that the confidence bounds straddle zero and hence the values on the lower and upper bounds will have different signs
2. The absolute value for large values at a 95% confidence interval will be 1.96 , so it will be a little more than 1.96
3. A value will not contain 0

Hence the final answer will be:

D: All of the above

Question 6

I regress Y on three independent variables – X_1, X_2, and X_3 – and I find the following 95% confidence intervals – β_1: [0.12, 1.45], β_2: [-0.01, 0.15], and β_3: [-0.64, -0.01]. Which of the following statements is most accurate?

A. β_2 and β_3 will have negative coefficients, and β_1 is statistically significant at the 0.05 level (two-tailed),

B. β_1 will have a positive coefficient, and β_1 is the only statistically significant coefficient (of the three coefficients) at the 0.05 level (two-tailed),

C. β_1 will have a positive coefficient, β_2 and β_3 will have negative coefficients, and only β_1 and β_3 are statistically significant at the 0.05 level (two-tailed),

D. All of the above,

E. None of the above and/or there is not enough information.

Ans 6:

Looking at the three Betas, we have Beta 1 not containing zero and positive hence it is statistically significant at 0.05 for a double tailed test. Beta 2 contains a zero and is hence probably not statistically significant at the 0.05 level. We also cannot tell if it is positive or negative. Beta 3 does not contain a zero and is negative, hence can be viewed as statically significant.

Based on this understanding, the most accurate statement will be:

C: β_1 will have a positive coefficient, β_2 and β_3 will have negative coefficients, and only β_1 and β_3 are statistically significant at the 0.05 level (two-tailed)

Question 7

Suppose I estimate a regression with two independent variables and obtain a R^2 of 0.40 where the Residual Sum of Squares is equal to 5,150. What does the Total Sum of Squares equal in this regression model?

Ans 7:

We know that:

$$R^2 = 1 - RSS/TSS$$

Hence,

$$0.40 = 1 - 5,150/TSS$$

Or TSS = 8,583.33 approx

Question 8

I regress Y on X_1 and X_2 (i.e., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$). I find that $\beta_1 = -0.87$ with a two-tailed p value of 0.001 and $\beta_2 = 3.52$ with a two-tailed p-value of 0.04. Which of the following statements is most accurate?

A. X_2 has a larger substantive effect on Y than X_1, but the effect of X_1 is more statistically significant,

B. X_1 has a larger substantive effect on Y than X_2 and X_1 is more statistically significant than X_2,

C. X_1 has a larger substantive effect on Y than X_2, but the effect of X_2 is more statistically significant,

D. None of the above and/or there is not enough information.

Ans 8:

- **Substantive Effect Size**: The coefficients (-0.87 for X1 and 3.52 for X2) indicate the magnitude of the effect each variable has on Y. However, without knowing the units or scale of X1 and X2, we cannot conclusively say which has a larger substantive effect. For instance, a small change in a variable with a small coefficient might have a more substantial effect than a large change in a variable with a large coefficient, depending on the scale of the variables.

- **Statistical Significance**: The p-values (0.001 for X1 and 0.04 for X2) indicate the statistical significance of the effects. While X1's effect is more statistically significant (lower p-value), this does not directly relate to the substantive effect size.

Hence the final answer is:

D. None of the above and/or there is not enough information.

Question 9

I regress Y on X_1 (i.e., $Y = \beta_0 + \beta_1 X_1$) and obtain a R^2 of 0.45. Then, I regress Y on both X_1 and X_2 (i.e., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$) and obtain a R^2 of 0.53. Which of the following statements must be true?

A. The correlation between X_1 and Y is stronger than the correlation between X_2 and Y,

B. The coefficient on β_1 is statistically significant in both models,

C. The model with X_1 and X_2 explains more of the variation in Y than a model with just X_1,

D. All of the above: all of these statements are true.

Ans 9:

1. Correlation cannot be directly to the Variance that is explained by X1 or X2. You can have a variable that is highly correlated and yet does not explain a lot of the variance
2. Statistical significance can also not be explained solely through R squared values.
3. X2 and X1 together has a larger R squared compared to simply X1. This means that the varience of Y that is explained by X1 and X2 together is more than that what is explained solely by X1,

Hence C: The model with X_1 and X_2 explains more of the variation in Y than a model with just X_1,

Question 10

Based on the following regression equation, might the negative coefficient on the South variable be the result of southern states having a lower proportion of high school graduates than non-southern states? In no more than one sentence, explain your answer.

$$Turnout = 34 + 0.5(Percent\ High\ School\ Graduates) – 5.9(South)$$

Ans 10 :

Yes, the negative coefficient on the South variable might indicate that southern states have a lower voter turnout, potentially due to a lower proportion of high school graduates, as the positive coefficient on Percent High School Graduates suggests a link between higher education levels and increased voter turnout.

Applied Questions

All data sets referenced below are posted on the ELMs course site.

Question 10

Points: 15

Use the `world` dataset to answer the following questions. The relevant variables and their coding information are as follows: `literacy` indicates a country's literacy rate; `dem_score14` represents a country's level of democratization (higher values indicate greater democratization); `spendeduc` reflects the amount of public expenditures on education as a percentage of GDP; `gdp_10_thou` represents GDP per capita (one unit is $10,000 USD); `educ_quality` indicates the average quality rating of a country's educational system; and `ungr9095` represents the percent average annual population growth.

Part A

Points: 4

Evaluate the bivariate relationship (using R) between a country's support for public education and its literacy rate. Report the results in a professionally formatted table. Do countries that spend more money on education appear to have a higher literacy rate?

Ans 10 Part A:

```
library(ggplot2)
world_df <- read.csv("world.csv")

#Here we want to indicate the relationship between spendeduc and literacy

#Lets drop na values
world_literacy_df <- na.omit(world_df[, c("literacy", "spendeduc")])
```

```
#Some Descriptive stats before we do modelling
summary(world_literacy_df$literacy)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.80   68.15   88.85   80.96   98.35  100.00
```

```r
summary(world_literacy_df$spendeduc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.600   3.225   4.350   4.468   5.300  13.600
```

So literacy goes all the way to 100% but the spend only hits 13.6% at max. It is on the same scale however so we can continue with the analysis.

```r
#Correlation analysis
cor(world_literacy_df$spendeduc, world_literacy_df$literacy)
```

```
[1] 0.1644235
```

Not a very strong correlation , but it is positive

Let's try a regression and see if we can explain variance in one due to the other:

```r
#Explaining the varience of literacy due to education spending. Or is an increase in liter

#Showing the results in the form of a professional table
model <- lm(literacy ~ spendeduc, data = world_literacy_df)

library(gtsummary)
table <- gtsummary::tbl_regression(model)

table <- tbl_regression(model,
                        label = list(spendeduc = "Education Spending"))

table
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

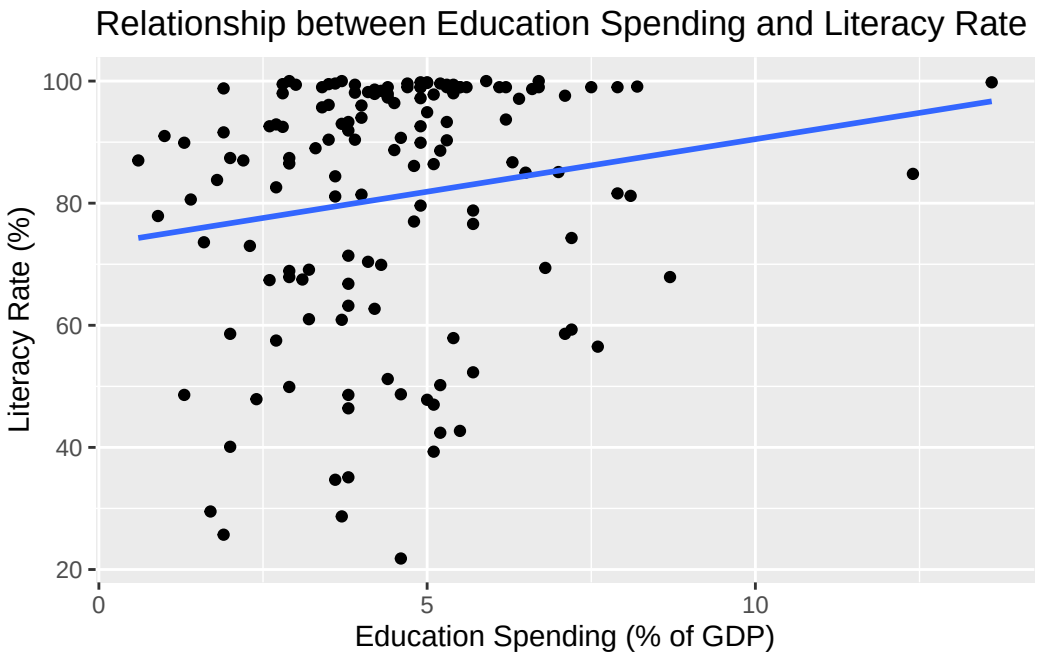
Characteristic	Beta	95% CI	p-value
Education Spending	1.7	0.00, 3.4	0.051

So we have the estimate as a positive value with a p_value of near 0.05. This means that there is a weak positive link between the two variables. The P-value would probably increase if there are other variables introduced that explained more of the variance of the literacy variable.

Let's look at this graphically:

```
ggplot(world_literacy_df, aes(x = spendeduc, y = literacy)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relationship between Education Spending and Literacy Rate",
       x = "Education Spending (% of GDP)",
       y = "Literacy Rate (%)")
```

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



Based on the movement of the line we can see that there is a slight positive relationship between the two variables. This is consistent with the other methods we have employed to check the relationship.

Based on this, we can conclude that there is a slight but positive increase of literacy based on education spend. We can also conclude that yes, countries that spend more money on education appear to have a higher literacy rate.

Part B

Points: 5

Estimate a second regression model (using R) that includes both a country's public expenditures on education and its level of democratization as independent variables. When controlling for democratization, does education expenditures exhibit a significant impact on literacy? Report the results in a professionally formatted table. Be sure to discuss the extent to which the results change, and if so, why they changed.

Ans 10 Part B:

Running a regression with the mentioned variables:

```
#Dropping na values
world_df_spend <- na.omit(world_df[, c("literacy", "spendeduc", "dem_score14")])

#Running a regression
model_2 <- lm(literacy ~ spendeduc + dem_score14, data = world_df_spend)

#table_2 values
table_2 <- tbl_regression(model_2,
                          label = list(spendeduc = "Education Spending", dem_score14 = "Leve

# Print the table
table_2
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	Beta	95% CI	p-value
Education Spending	0.26	-1.3, 1.8	0.7
Level of Democratization	4.7	3.2, 6.1	<0.001

Based on table results we can see that the P_value for the spending of education drops significantly such that it is not statically significant anymore. The estimate is still positive

though. This indicates that the variable democratization is a confounding variable whose effects need to be taken into account if there is analysis to be done on the relationship between literacy and education spending. The beta for education spending has decreased from 1.7 to 0.26.

```
cor.test(world_df_spend$spendeduc, world_df_spend$dem_score14)
```

```
    Pearson's product-moment correlation

data:  world_df_spend$spendeduc and world_df_spend$dem_score14
t = 3.4994, df = 140, p-value = 0.0006257
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1247075 0.4283254
sample estimates:
      cor
0.2836087
```

Just to test, the correlation between spend and democratization is positive and significant. This indicates that there is a significant relationship between them and the variables need to be taken into account before any further analysis can be taken into account.

Part C

Points: 6

Now estimate a third model (using R) that includes all of the variables listed above. Report the regression results in a professionally formatted table and interpret each coefficient. When interpreting the impact of democratization using this regression model, present/utilize a visual representation (using R) of the expected change in literacy as a function of democratization.

Ans 10 Part C:

Creating a regression equation with all the variables included such as:

`dem_score14, spendeduc, gdp_10_thou, educ_quality and ungr9095`

```
#Dropping na values
world_df_final <- na.omit(world_df[,c("literacy", "dem_score14", "spendeduc", "gdp_10_thou

#Running a regression
model_3 <- lm(literacy ~ dem_score14 + spendeduc + gdp_10_thou + educ_quality + ungr9095,
```

```r
#table_2 values
table_3 <- tbl_regression(model_3,
                          label = list(spendeduc = "Education Spending", dem_score14 = "Leve

# Print the table
table_3
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

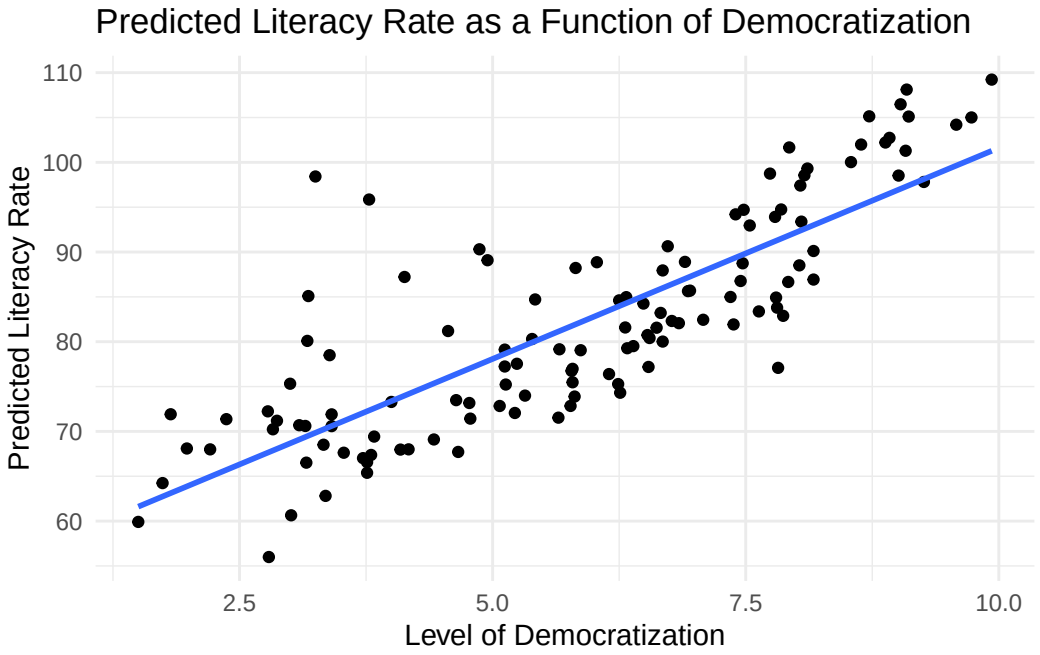
Characteristic	**Beta**	**95% CI**	**p-value**
Level of Democratization	2.5	0.73, 4.4	0.006
Education Spending	0.11	-1.7, 1.9	>0.9
GDP per capita	2.4	-1.6, 6.4	0.2
Average quality of a country's educational system	3.8	-0.56, 8.2	0.087
Percent Average Annual Population Growth	-3.3	-5.2, -1.5	<0.001

Prediction and Visualization of the variable:

```r
# Predict literacy rates using the model across a range of democratization scores
world_df_final$predicted_literacy <- predict(model_3, newdata = world_df_final)

# Create the plot
ggplot(world_df_final, aes(x = dem_score14, y = predicted_literacy)) +
  geom_point() +  # Plot the actual points
  geom_smooth(method = "lm", se = FALSE) +  # Add a linear regression line
  labs(x = "Level of Democratization", y = "Predicted Literacy Rate",
       title = "Predicted Literacy Rate as a Function of Democratization") +
  theme_minimal()
```

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



Based on the graph we can see that the predicted literacy level is positively related with the level of Democratization. This is backed by the regression results that give the beta of democratization as a positive value.

Question 12

Points: 15

Using the `states` dataset, use R to regress the variable `obama08` (the percentage of a state's vote that President Obama received in the 2008 U.S. presidential election) on the following independent variables: `cig_tax` represents the amount of a state's cigarette tax (in dollars); `college` is the percentage of a state's population that graduated college (0 to 100); `union07` is the percentage of a state's workers that are union members (0 to 100); and `south` is a dichotomous variable coded as `1` if the state is located in the South (`0` otherwise). Report the results in a professionally formatted table. Answer the following questions.

Part A

Points: 6

Interpret the substantive results from this regression model. Be sure to discuss the direction, magnitude, and statistical significance of each slope coefficient. Does the `union07` variably exhibit a substantively significant impact on voting behavior in the 2008 election?

Ans 12 A:

```
#Reading data
df_states <- read.csv("states.csv")

#Running the regression with the given variables

#Dropping na values
states_df_final <- na.omit(df_states[,c("obama08", "cig_tax", "college", "union07", "south

#Running a regression
model_states <- lm(obama08 ~ cig_tax + college + union07 + south, data = states_df_final)

states_table <- tbl_regression(model_states,
                        label = list(cig_tax = "State Cigarette Tax(Dollars)", college = "

#Note add the standard error to this:
states_table
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	Beta	95% CI	p-value
State Cigarette Tax(Dollars)	0.52	-4.0, 5.0	0.8
Percentage of a state that graduated college	0.96	0.47, 1.5	<0.001
Percentage of a State's Workers that are Union Members	0.75	0.26, 1.2	0.004
Dicthomous variable indicating if the state is South or not	1.4	-3.8, 6.6	0.6

Analysis:

The regression analysis conducted models the percentage of a state's vote that President Obama received in the 2008 U.S. presidential election (`obama08`) as a function of several independent variables: the amount of a state's cigarette tax (`cig_tax`), the percentage of a state's population that graduated college (`college`), the percentage of a state's workers that are union members (`union07`), and a dichotomous variable indicating whether the state is located in the South (`south`).

1. **Coefficients**

- **(Intercept)**: The estimated baseline percentage of votes for Obama (when all other variables are 0) is 16.13%, and it is statistically significant (p = 0.018374).
- **cig_tax**: The coefficient of 0.523 suggests that for each dollar increase in the cigarette tax, the percentage of votes for Obama is expected to increase by 0.523%. However, this effect is not statistically significant (p = 0.815676).
- **college**: The coefficient of 0.965 indicates that for each 1% increase in the college graduation rate, the percentage of votes for Obama is expected to increase by 0.965%. This effect is highly significant (p = 0.000299).
- **union07**: The coefficient of 0.751 implies that for each 1% increase in the proportion of union workers, the percentage of votes for Obama is expected to increase by 0.751%. This effect is also statistically significant (p = 0.003511).
- **south**: The coefficient of 1.393 suggests that being in the South is associated with a 1.393% increase in the votes for Obama, but this effect is not statistically significant (p = 0.593179).

2. **Model Fit**:
   - **Residual Standard Error**: The residual standard error is 6.869, indicating the average distance of the data points from the fitted line.
   - **Multiple R-squared**: 0.5201. This value indicates that approximately 52.01% of the variability in the percentage of votes for Obama is explained by the model.
   - **Adjusted R-squared**: 0.4775. This is a modification of the R-squared that adjusts for the number of predictors in the model. It's a more accurate measure of the goodness of fit, especially when dealing with multiple predictors.
   - **F-statistic**: The F-statistic is 12.19 with a very small p-value (8.504e-07), suggesting that the model as a whole is statistically significant.
3. **Interpretation**:
   - The results indicate that the percentage of a state's population that graduated college and the percentage of union workers are significant predictors of Obama's vote share in the 2008 election, with both showing positive relationships.
   - The cigarette tax and whether the state is in the South do not appear to be significant predictors in this model.
   - The high p-values for cig_tax and south indicate that their coefficients are not statistically different from zero in the context of this model, suggesting little to no linear relationship with Obama's vote share when controlling for the other variables.

- The model explains a significant portion of the variation in the vote share, but there are other factors not included in the model that also affect the outcome. Having an Adjusted R-squared of only 47% is fairly low when making any sort of high level prediction or analysis.

From the interpretation we can say that **union07** is is highly significant variable based on it's p_value and estimate. Hence it has significantly affect voting behavior in the 2008 election.

Part B

Points: 2

Interpret the intercept coefficient. What does this represent and is it meaningful to interpret this coefficient on its own? Why or why not?

Ans 12 B:

Taking a part of the answer from the above question,

The intercept coefficient in a regression model represents the expected value of the dependent variable when all the independent variables are set to zero. In the context of your model, the intercept coefficient (16.1280) represents the expected percentage of votes for Obama in a hypothetical state where:

- The cigarette tax (`cig_tax`) is \$0.
- The percentage of the population that graduated college (`college`) is 0%.
- The percentage of workers that are union members (`union07`) is 0%.
- The state is not in the South (`south` is 0).

Interpreting the Intercept:

1. **Contextual Meaningfulness**: In many regression models, especially those involving real-world data, the intercept can be a theoretical construct rather than a practically meaningful value. For example, it's unlikely that any state would have a 0% college graduation rate, 0% union membership, and no cigarette tax, all simultaneously. Hence, while the intercept has statistical significance in the model ($p = 0.018374$), its real-world interpretability might be limited.
2. **Intercept in Isolation**: Interpreting the intercept on its own can be misleading in many cases, particularly if the values of zero for all independent variables are not realistic or meaningful within the context of the data. The intercept is more useful as a part of the overall equation that predicts the dependent variable based on a combination of the independent variables.

3. **Role in the Model**: The primary role of the intercept in a regression model is to ensure that the regression line correctly models the relationship between the dependent and independent variables over the range in which the data exist. It adjusts the height of the regression line and is essential for the model's accuracy, but its value should be interpreted in the context of the specific variables and the data being analyzed.

In summary, while the intercept in our model is statistically significant, its practical interpretation should be approached with caution. The value of the intercept is more informative when considered in conjunction with the coefficients of the independent variables.

Part C

Points: 2

What is the expected mean percentage of Obama's vote total in a Southern state with a cigarette tax of $2.00, a state population where 30% of people are college graduates, and where 15% of workers in a state are union members?

Ans 12 C:

Plugging values into the regression equation we have derived ,

We have:

- The cigarette tax (`cig_tax`) is $2.
- The percentage of the population that graduated college (`college`) is 30%.
- The percentage of workers that are union members (`union07`) is 15%.
- The state is in the South (`south` is 1).

Hence the total equation would be:

1. $$ObamaVoteExpected = intercept + (coef\_cig\_tax * cig\_tax) + (coef\_college\_tax * college\_tax) + (coef\_$$

Which would equal to about 58.7%

Hence Obama would get 58.7% of the total vote based on the given values.

Part D

Points: 3

What is a 90% confidence interval around the coefficient for the `south` variable.

Ans 12 D:

To calculate the confidence interval , we first need to derive the t value around the interval first:

We know that,

1. $$Confidence\_interval = Coefficent + -Standard\_Error * t - value$$

```
#Checking the dimensions of the data for degrees of freedom:
dim(states_df_final)
```

```
[1] 50  5
```

We have:

Coefficient = 1.393

Standard Error = 2.58

Degrees of Freedom = 50 - 5 [Number of Variables - Variable Parameters]

Hence the t_value for this would be:

```
qt(0.95, 45)
```

```
[1] 1.679427
```

Hence the upper and lower bounds of the confidence interval would be:

```
lower_bound <- 1.393 - 1.679427 * 2.58
upper_bound <- 1.393 + 1.679427 * 2.58

lower_bound
```

```
[1] -2.939922
```

```
upper_bound
```

```
[1] 5.725922
```

Hence the confidence interval runs from [-2.93, 5.725]

Part E

Points: 3

What is the null hypothesis that the F-test in this regression output is testing?

Ans 12 E:

The F-test in a multiple regression analysis is used to test the null hypothesis that all of the regression coefficients in the model are equal to zero. This hypothesis essentially assumes that the model, as a whole, has no explanatory power in predicting the dependent variable.

In the context of this regression model, which includes the independent variables `cig_tax`, `college`, `union07`, and `south` to predict `obama08` , the null hypothesis for the F-test can be stated as follows:

Null Hypothesis (H0): The coefficients for `cig_tax`, `college`, `union07`, and `south` are all equal to zero, implying that none of these variables have a statistically significant effect on the percentage of votes for Obama.

In other words, the F-test is checking whether there is any evidence to suggest that at least one of these independent variables is useful in predicting the dependent variable `obama08`. If the F-test's p-value is small (typically less than 0.05), it suggests that we can reject the null hypothesis and conclude that at least one of the coefficients is significantly different from zero, indicating that the model provides some explanatory power. If the p-value is not small, we would not reject the null hypothesis.