Question 1. (25 percent)

Fearon and Laitin (2003) published an article in the American Political Science Review analyzing civil war onset. The regression output above replicates their main findings. Answer the following:

- What kind of model is shown here? **Answer: Linear probability model with standard errors clustered on ccode/country**
- Are oil producing states more likely to see a civil war compared to non-oil producing countries? **Answer: Yes, oil producing states are 1.7% more likely to see a civil war compared to a non-oil producing country, and this is statistically significant at p<.05.**
- What role does ethnic and religious fractionalization play in civil war onset? **Answer: Higher levels of ethnic fractionalization are positively associated with civil war onset; however, this is not statistically significant at p<.05, and so we cannot reject the null hypothesis of no association. Similarly, higher levels of religious fractionalization are positively associated with civil war onset, but this is also not statistically significant at conventional levels (p=.71).**

Which countries are most likely to see a civil war? Justify your answer with reference to the reported effect sizes and hypothesis tests.

Answer:

1) **Countries that experienced war the previous year were 1.5% more likely to experience a civil war, and this association is statistically significant at p<.05;**
2) **Economic growth is significantly negatively associated with civil war onset. To give an illustration: a 10% increase in GDP is associated with a 10% diminution in the probability of having a civil war (p=.001);**
3) **More populous countries were also more civil war prone: a 10% increase in population size increases the probability of a civil war by 4% (p=.002);**
4) **As previously noted, oil producing states are 1.7% more likely to see a civil war compared to a non-oil producing country, and this is statistically significant at p<.05;**
5) **Countries established within the past 2 years were 5.8% more likely to see a civil war compared to more establishes countries, and this is also statistically significant at p<.05;**
6) **Finally, politically unstable countries were 1.4% more likely to experience a civil war compared to politically stable countries, and this is statistically significant at p<.05.**

Question 2. (15 percent)

You have to design a study explaining participation in anti-Trump protests in the United States following the announcement of the Muslim travel ban in January 2017. The data is a sample of 1,219 counties from all 50 states. The unit of analysis is a county located in a state observed on a single day. The dependent variable is the number of anti-Trump protestors in a county. How would you model this? Justify your answer with reference to the estimator's properties and the data structure.

Answer:

1) **Count model answer: the dependent variable is positively skewed and unimodal. This might suggest Poisson regression; however, the variance is substantively larger than the mean, suggesting overdispersion. As this is the case, Poisson would lead to overly confident test statistics and the risk of false positives. Given this, negative binomial regression is more appropriate. As the dependent variable is a count of protestors in a county, the relevant exposure term would be the logged population of a county.**
2) **OLS answer: the dependent variable is positively skewed and contains a large number of zero's. As such, it should be transformed and centered. Due to the large number of zero's, a log transformation is not appropriate. An alternative transformation would be the square root or the inverse hyperbolic sine.**
3) **Standard errors. The unit of analysis is a county nested within a state. To account for any clustering that might bias standard errors downwards, standard errors should be clustered on the state. Alternatively, this can be treated as a multilevel modelling problem, with random intercepts entered at the state level (the number of state is sufficient: 50). Alternatively, between unit variance at the state level can be accounted for using fixed intercepts, thus confining attention to within-state differences.**

Question 3 (25 percent)

Under a number of assumptions, the Ordinary Least Squares model is the Best Linear Unbiased Estimator (BLUE). Choose four of these assumptions and:

1) explain them; **Answer (four from these):**

   Studenmund:

   - **The regression model is linear, is correctly specified, and has an additive error term;**
   - **The error term has a zero population mean;**
   - **All explanatory variables are uncorrelated with the error term;**
   - **Observations of the error term are uncorrelated with each other;**
   - **The error term has a constant variance;**
   - **No explanatory variable is a perfect linear combination of any other explanatory variable(s);**
   - **The error term is normally distributed**

   Stock & Watson:

   - **The conditional distribution of e_i given X_i has a mean of 0**
     - **$E[e_i] = 0$**
     - **$Cov(X_i, e_i) = 0$**
   - **(X_i, Y_i) are independently and identically distributed (i.i.d)**
     - **No autocorrelation**

- **No heteroskedasticity**

- **Large outliers are unlikely**
  - **e is normally distributed**
  - **If N is very large, then large outliers will appear, but this is OK**

Kennedy:

- **The model is linear and correct**
  - **Y = a + bx + e**
- **The expected value of e is 0**
- **The variance of e does not change with any x**
  - **$E(e^2|x) = E(e^2)$**
- **X is fixed**
  - **No measurement error**
- **No X is a linear combination of other Xs**

2) explain the consequences of violating them; **Answer:**

   Model misspecification:

   Omitted variable: if Corr(Z,X) = 0 standard errors are inflated -> smaller t-statistics. If Corr(Z,X) ≠ 0 β_1 becomes biased.

   Inclusion of irrelevant variable: The beta coefficients will remain unbiased if the true effect of irrelevant variable = 0. However, inclusion of irrelevant variable will increase the variance of all beta coefficients = smaller t-statistics and reduction in adjusted r-squared.

   Standard errors:

   Non-normally distributed residuals (heteroskedastic) or serial autocorrelation. Beta coefficients are okay but standard errors are inefficient -> biased test statistics.

   Multicollinearity:

   Beta coefficients will remain unbiased but variance and standard errors will increase -> unexpected signs and risk of false negatives

3) describe how we can test whether they hold or not.

   Model misspecification:

   Visible in small beta and large SE. Link test. Enter quadratic term.

   Standard errors:

   Plot residuals. Breusch-Pagan (1979) and Cook-Weisberg (1983) test for heteroscedasticity. Breusch-Godfrey test for higher-order serial correlation. Durbin's test for serial correlation.

   Multicollinearity:

   Correlation matrix. Variance inflation factor score.

Question 4 (25 percent)

Political scientists often work with hierarchical or nested data.

1) How does clustering affect our hypothesis tests? **Answer: Clustering leads to downwardly biased standard errors and thus upwardly biased test statistics, leading to a greater risk of false positives.**
2) How can we detect clustering? **Answer: Clustering of observations within higher units can be detected using the intraclass correlation. This provides a statistic from 0-100 which tells us how much of the variation on our dependent variable is explained by between-unit differences at higher levels.**
3) What are some possible remedies? **Answer: 1) artificially inflate standard errors by using cluster-adjusted standard errors; 2) random intercepts at higher units; 3) fixed effects at higher units.**

Question 5 (10 percent)

Conflict data is often noisy. What are some possible implications of having measurement error on a key independent variable in a multivariate regression? **Answer: measurement error on a key independent variable affects the beta coefficient, although this depends on the sample size. In large samples, beta will be attenuated, leading to an increased risk of false negatives. However, in small samples, beta can be spuriously large, giving rise to an increased risk of false positives. In a multivariate regression, measurement error on an independent variable can bias the beta coefficients of other correlated variables.**