Name and Github user ID

No more than 2-3 pages

Describe research question

Discuss approach I took and the coding involved

Including discussing any weaknesses or difficulties encountered

Finish with brief discussion of results and how it could be fleshed out in future research

Primary purpose of writeup is inform them of what I’m reading before we look at code

I intend to look at disabled veteran earnings and labor force participation compared to civilian nondisabled cohort earnings and labor force participation by region of the United States. “Disabled” veteran is a wide definition that can include anything as mild as hearing loss or pre-existing conditions prior to joining the military; to as extreme as losing a limb in combat, with associated monthly payments to scale based on a “rating” (percentage disabled) given when exiting the military. Unfortunately, there is a culture in the military of trying to maximize your disability benefits when exiting service (“make sure you say you have sleep apnea, that’s an automatic 20%”) on Facebook groups, Reddit threads, word of mouth, and even specialty consultants or lawyers to help you appeal your rating post hoc. These efforts may result in an equal disability rating through a sum of minor injuries, to that of a serious wound or injury, that truly impacts an individual’s ability to serve in the labor force and thus lifetime earnings. The Department of Veterans Affairs continues to pay out increasing disability compensation as years pass – in FY22, disability payouts totaled $125 billion. The conversation around veteran disability payments, especially when evaluating government accountability and ballooning budgets, mirrors the common conversations around the social safety net – do these “benefits” change willingness to work? Does a rating change a person’s perceptions of their prospects, and their intrinsic motivation? Disability benefits have another layer of analysis beyond the social safety net, as they are not means tested, and if a conversation ever burgeons regarding means testing there is a vicious outcry. They are also benefits for life.

My hypothesis is generally that lower disability ratings translate to comparable earnings and LFP to nondisabled counterparts in similar areas, if not greater; while high disability ratings would then affect ability to work and earn all else being equal, so earnings may be comparable but LFP will be much lower. Regionally, I hypothesize there will be significant differences in earnings and LFP in the American South, and will be interested to see other differences. Earnings will be interesting to evaluate due to the scale of monetary benefits available as disability ratings increase, and academics have looked at whether this compensation “makes up” for the loss of earnings due to true disability and impacted employment prospects. However, the difficult thing to pull out is who is truly disabled and unable to work; versus who is not incentivized to work but could, because of their high ratings or higher ratings then they should have been evaluated for. I plan to use IPUMS CPS data to look at veteran earnings and LFP, from 2011 to 2015 (five years’ worth of data) and compare to cohort non-veteran earnings and LFP. I then plan to use the annual IPUMS Veterans Supplement, again for 2011 to 2015, to expand the analysis to then look at these cohorts when now including different levels of their service-connected disability status (low to medium to high), corresponding compensation for the disability, and the resulting effect of the disability on the person's labor force status, and participation in veterans' programs. For a third dataset I intend to look at this data by state, or pull data from data.gov to add some sort of state-based analysis – there is a component to this discussion where some states or areas of the United States contribute disproportionately more servicemembers on average, so it would be interesting to see a by-state/region breakdown.

The output you've provided is from an R linear regression model summary, where the dependent variable is `INCTOT` (which I assume represents total income), and the independent variables include various categories of `DISABILITY\_SIMPLE` and `AGE`. Here's what the output tells us:

1. \*\*Coefficients\*\*: The table lists the estimated impact of each independent variable on the dependent variable (`INCTOT`). Each row corresponds to a predictor in the model.

- `(Intercept)`: The intercept term represents the average expected value of `INCTOT` when all other predictors are held at zero. The value is extremely high, suggesting the data or the model may not be centered around zero or appropriately scaled.

- `DISABILITY\_SIMPLE` categories: These represent the difference in average `INCTOT` associated with each category of disability, compared to the reference category (which is not shown and is implicitly set to a coefficient of 0).

- `AGE`: The coefficient for `AGE` indicates the average change in `INCTOT` for each additional year of age, holding other variables constant.

2. \*\*Standard Error\*\*: This column shows the standard error of each coefficient estimate, which measures the variability or uncertainty in the coefficient estimates.

3. \*\*t value\*\*: The t value is the coefficient divided by its standard error. It's used to determine the statistical significance of each coefficient.

4. \*\*Pr(>|t|)\*\*: This is the p-value associated with the hypothesis test for each coefficient. A low p-value (typically < 0.05) suggests that the corresponding predictor is statistically significantly associated with the dependent variable.

- `\*`: Indicates a p-value < 0.05, which is commonly considered statistically significant.

- `\*\*`: Indicates a p-value < 0.01, which is considered highly statistically significant.

- `\*\*\*`: Indicates a p-value < 0.001, which is considered extremely statistically significant.

5. \*\*Significance Codes\*\*: The stars next to the coefficient estimates indicate their level of statistical significance based on the p-values.

6. \*\*Residual Standard Error\*\*: This is the estimated standard deviation of the error term (the residuals). It gives you an idea of how much the responses deviate from the fitted regression line.

7. \*\*Multiple R-squared\*\*: This value tells you the proportion of variance in the dependent variable (`INCTOT`) that can be explained by the model. Here, approximately 42.17% of the variation in `INCTOT` can be explained by the model.

8. \*\*Adjusted R-squared\*\*: This is a modified version of the R-squared that adjusts for the number of predictors in the model. It provides a more accurate measure of the goodness of fit when you have multiple predictors. An adjusted R-squared of 42.17% suggests a good fit.

9. \*\*F-statistic\*\*: This is a test statistic for the overall significance of the model. A very low p-value here suggests that the model is statistically significantly better than an empty model (a model with no predictors).

10. \*\*Degrees of Freedom\*\*: This refers to the number of independent pieces of information used to calculate the estimate. Here, the model has 6 predictors (`DISABILITY\_SIMPLE` categories and `AGE`), and there are 6268497 degrees of freedom for the residuals, suggesting a very large dataset.

### Interpretation of Specific Coefficients:

- `DISABILITY\_SIMPLE1-60%`: Being in the 1-60% disability category is associated with a decrease in `INCTOT` by approximately $9,950, compared to the reference disability category.

- `DISABILITY\_SIMPLE70%+`: Being in the 70%+ disability category does not show a statistically significant difference in `INCTOT` compared to the reference category (p-value > 0.05).

- `DISABILITY\_SIMPLEN/A`: This category has a significant negative association with `INCTOT`, with an average decrease of about $434,972 compared to the reference category.

- `AGE`: Each additional year of age is associated with a decrease in `INCTOT` by about $105,161 on average.

The negative coefficients for `AGE` and some disability categories may indicate a need to further investigate the data and potentially transform or scale the variables for more meaningful interpretations. It's also possible that there are other confounding variables or that the relationships are not linear, which might require a more complex model to capture accurately.