Contents

[Background 1](#_Toc42423350)

[U6 from U3 2](#_Toc42423351)

[U3 from ICSA 5](#_Toc42423352)

[U3 from CCSA 8](#_Toc42423353)

[Summary 11](#_Toc42423354)

[Appendix A – Data cleaning and tidying 12](#_Toc42423355)

[Appendix B – Labor force estimation 13](#_Toc42423356)

[Appendix C – Excerpts from BLS report 14](#_Toc42423357)

[March Report, dated 3 April 2020 14](#_Toc42423358)

[April Report, dated 8 May. 2020 15](#_Toc42423359)

[May Report, dated 5 Jun 2020 15](#_Toc42423360)

[Appendix D – Residual Histograms 16](#_Toc42423361)

# Background

The headline unemployment number, U3 or UNRATE on FRED, trends do not agree with other unemployment data series. In particular the change in relationship between U3 and ICSA and CCSA is striking.

Starting in March, 2020 the Bureau of Labor Statistics (BLS) included a notation that the Covid19 absences should have been included in the unemployment number, see Appendix B – Excerpts from BLS report. This error persists in the April and May reports. Had this error been corrected, the U3 numbers would have 1%, 5%, and 3% higher for March, April and May respectively.

Another problem with the BLS data is that it shows a decrease in civilian labor force rates as the Covid19 virus hit. The LNU01000000 series shows a drop from 164MM to 155MM. This is because of the way the BLS calculates the number (employed + unemployed) when it should be the number of people eligible to work. To get a constant value of eligible workers the LNU01000000 was fit to the overall US population (POPTHM) to get an average percentage of the population that is eligible to work. This value came out to XX% and it is POPTHM times this value that is used to normalized ICSA and CCSA numbers.

The notations also state that the survey has experience some disruptions due to Covid 19 lockdowns. These notations make it likely that the BLS unemployment numbers are not as representative of the actual conditions as past surveys have been.

This document discusses the changes in relationship across the some of the employment data series. Data was partitioned as explained in Appendix A – Data cleaning and tidying to avoid overfitting. Claims and payroll counts have been normalized by an average labor force value as described in Appendix B – Labor force estimation.

# U6 from U3

Since the U3/U6 are surveys by BLS I thought the two should roughly track each other. To test this, I estimate a linear between U3 and U6 where U6 is estimated from U3. Sure enough, the U3 data could be used to make a reasonable estimate of the U6 data, which makes sense the numbers tend to track. The beige line below is the U6RATE series and the other line is the linear model prediction based on U3. This dataset begins in 1994 since that is the first time U6 is published in FRED.

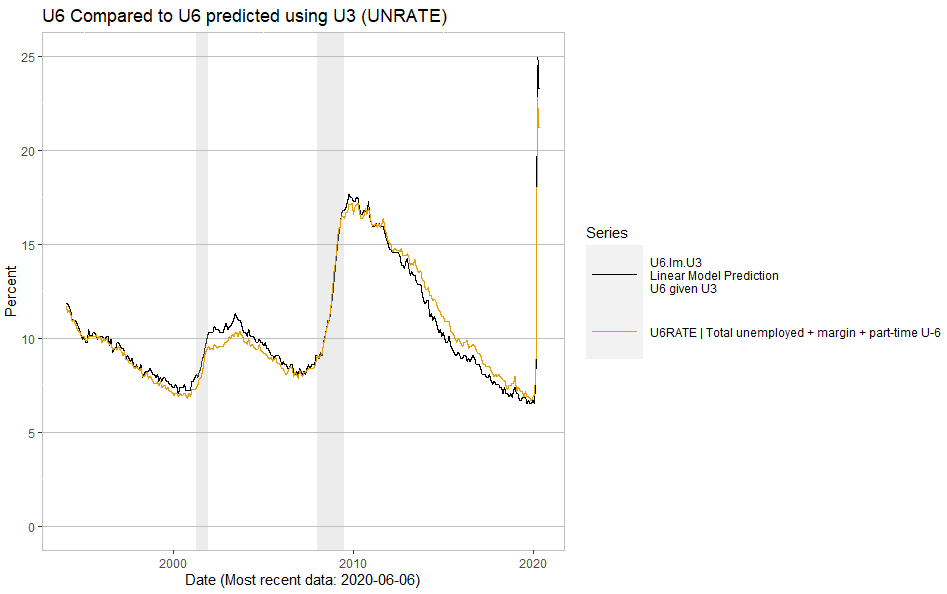


Figure 1 - U6 from FRED and U6 predicted by a linear model from U3.

It is interesting to see how the linear model behaves as the Covid19 outbreak spreads. The difference between the model and FRED U6 data is small in February and March, then widens in April and May. The linear model is predicting 3% higher values for U6 in this time frame.

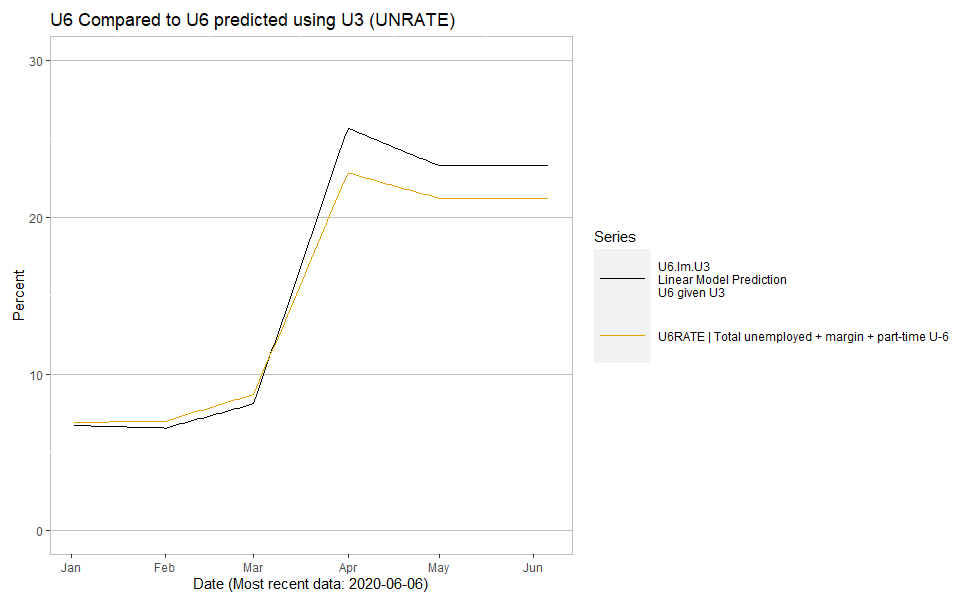


Figure 2-- U6 from FRED and U6 predicted by a linear model from U3 from January 1st, 2020 until present day.

The real reason to put together a model is to see how the residuals look, shown in Figure 3. The horizontal dashed lines are the 5-95% confidence intervals. The horizontal red dashed lines are the +/- 6-sigma values derived from the standard deviation of the test set residual values shown in Figure 9.

The residual values nearly exceed the lower 6-sigma level and are far out of the historical norms. The two surveys do not seem hold to the same relationship that existed for the previous 26 years.

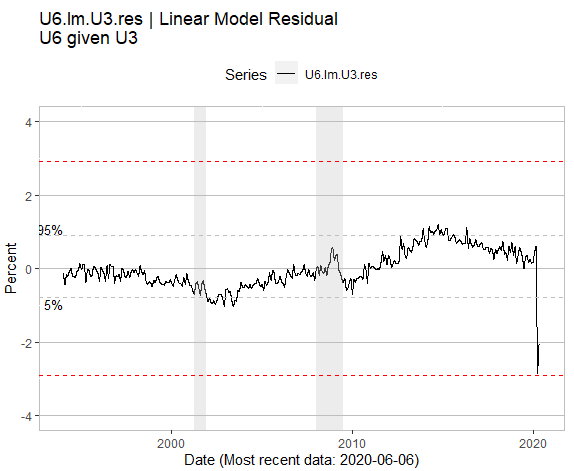


Figure 3 - Residual (U6 - estimated U6).

# U3 from ICSA

Initial claims data represents claim numbers tabulated from administrative sources, rather than surveys. The problem with the ICSA is that it only represents claims, not claims that resulted in benefits. More people make claims than receive benefits so ICSA is not a great indicator of unemployment. Even so, there is enough similarity that a linear fit can be used to estimate U3, shown Figure 3. The shape of the two curves are similar, but the U3 derived from ICSA has higher peaks and more noise. The ICSA data series begins in 1967 so this relationship has more data points than did the U6 from U3 discussed above.

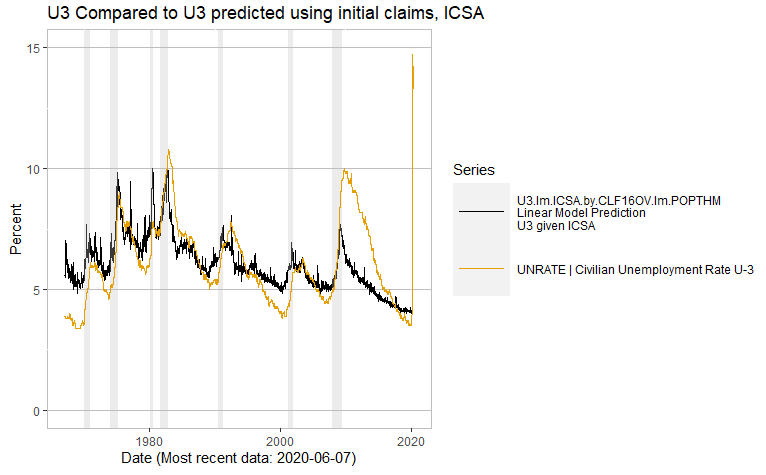


Figure 4 - U3 (UNRATE) from FRED and estimated U3 from ICSA

The divergence of FRED’s U3 from the predicted U3, Figure 5, over the last few months is striking. The ICSA predicts the U3 should be well above 90%. Obviously, this prediction is unrealistic. This radical change is almost certainly due to the desire by some folks to claim enhanced unemployment benefits. After all, if you could make more money on unemployment, why not give it a try and file? Most the claims are either denied or still be process, but have not turned into benefits.

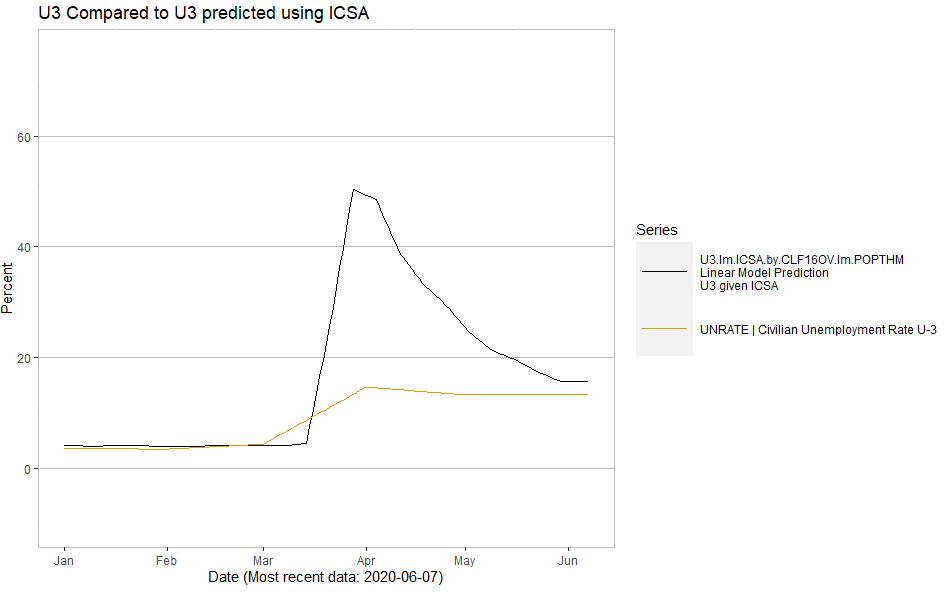


Figure 5 - U3 (UNRATE) from FRED and estimated U3 from ICSA since January 1st, 2020.

The residuals are the interesting story here too. The difference between the linear model prediction and the actual U3 is off the charts. Figure 10 shows the histogram for the residual values. The gray horizontal lines are the 5-95% confidence intervals and the red lines are the +/- 6-sigma values calculated from the histogram in Figure 10.

Over the 53-year history of the two data series, never have there been more than a +/- 3% difference between the linear model and the actual U3 values. In the last two months that difference has jumped to over 90%.

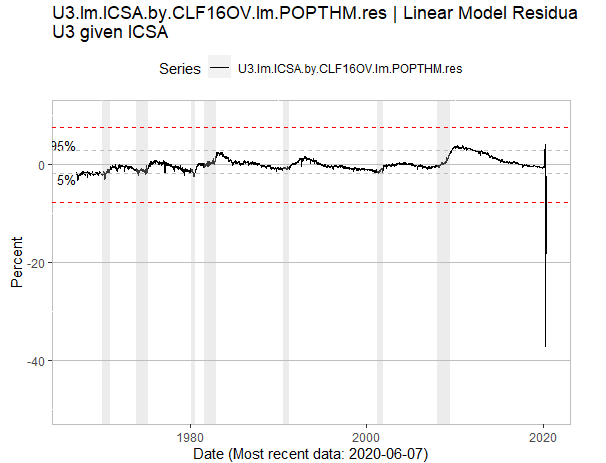


Figure 6 - Residual (U3 - U3 estimated from linear fit).

# U3 from CCSA

The continued claims series (CCSA from FRED) shows the number of people who have already filed an initial claim, experienced a week of unemployment, and gone on to file another claim. These are people receiving unemployment benefits and should be more representative of the number of unemployed.

Figure 7 shows the U3 data as the black line and U3 predicted from the CCSA data series as the beige line. The shape of the series is similar; however, there are differences between the series when unemployment peaks. Especially in the last two recessions the predicted U3 has overshot the actual U3 by 2-3% percentage points.

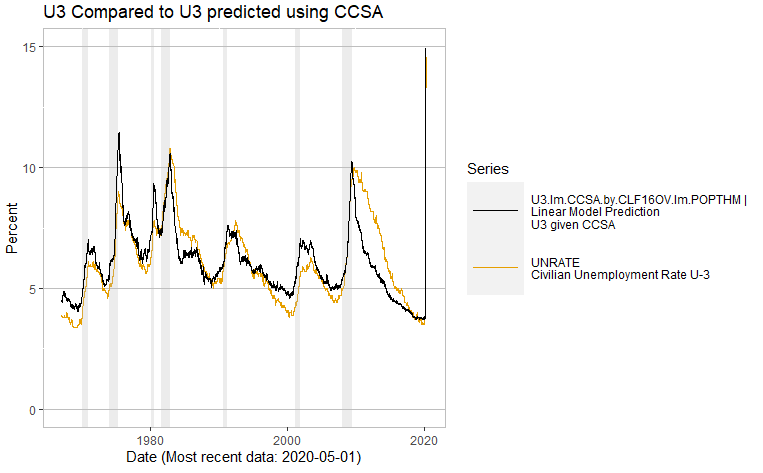


Figure 7 - U3 (UNRATE) from FRED and estimated U3 from CCSA.

The difference between the predicted and actual data has increase in the last few months. Figure 8 shows the predicted U3 rate rising above 40%, compared to the BLS U3 rate of 13%.

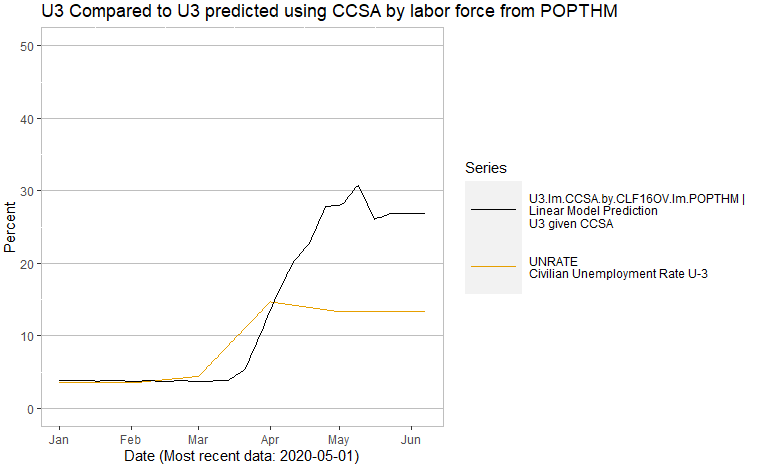


Figure 8 - U3 (UNRATE) from FRED and estimated U3 from CCSA since January 1st, 2020.

The residuals over the last few months have gone well beyond the statically expected bounds. The gray horizontal lines are the 5-95% confidence intervals and the red lines are the +/- 6-sigma values calculated from the histogram in Figure 11.

Over the 53-year history of these two series the difference between the linear model and the measured U3 has been with the +/-3% bounds shown in the histogram. The difference between predicted and measured increased to 26% over the last two months, well outside the lower 6-sigma value.

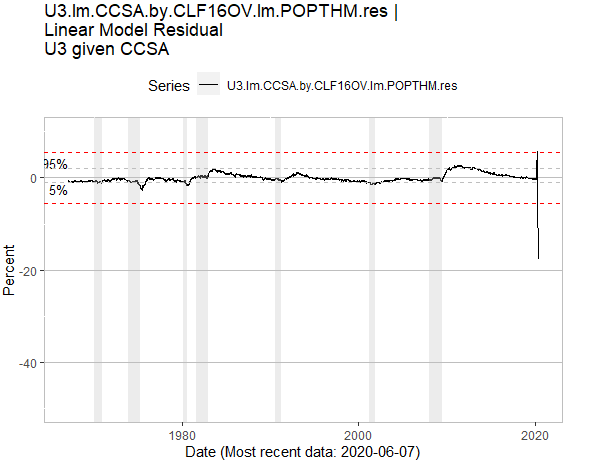


Figure - Residual (U3 - U3 estimated from CCSA linear fit).

# Summary

Data series reporting to measure different aspects of the same underlying economic variable should have some relationship to each other. The U3 and U6 data originate from BLS surveys where as the ICSA and CCSA numbers originate from administrative tabulations. Historically, a linear relationship as sufficed to describe the relationship between the two data sources. That has not held true for the last 3 months since Covid19 lockdowns came into effect.

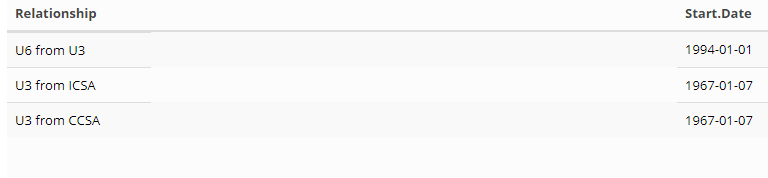
It is interesting that using either the ICSA or CCSA data series the predicted U3 is always higher than BLS U3 numbers. This is consistent with the BLS notes about misclassified responses in which they say the numbers would have higher had the problems been corrected.

The breakdown in the linear relationship could be an anomaly due to the way the survey results and processing were adapted to Covid19 as described in the BLS notes. It could also be due to the way the lockdowns were put into place. There are almost certainly other factors in play as well given the magnitude of the change in employment. Whatever the reason, the survey numbers seem to be underreporting unemployment levels and to have a higher degree of uncertainty than at anytime in the last 3-5 decades.

# Appendix A – Data cleaning and tidying

I followed the data science principles of partitioning and validation and used data up through January 1st of 2020. From that point on, I thought Covid19 would start to influence the data so I excluded from the fit estimation.

Data is broken into 3 separate and randomly selected partitions: training (50%), validation (25%), and testing (25%). The training set is used to build the model and validation is used to confirm there is no over-fitting. There is some back and forth with these data sets. Once the model is locked down the testing set is used to establish expected performance values. This is the data set that was used to establish 6-sigma values for the residuals. Residual Histograms were also plotted and reviewed, but are not included here.



# Appendix B – Labor force estimation

The BLS estimation of labor force is defined as the sum of employed and unemployed; however, it shows transients during events like Covid19 that are not representative of the labor force changes. For this project the labor force was estimated from the U.S. population data by performing a linear fit of the BLS estimated of labor force (CLF16OV) to the U.S. population to estimate the average labor force as a percentage of population. This should eliminate the spikes and transients that appear in the BLS data. Figure 9 shows the BLS values compared to this estimate.

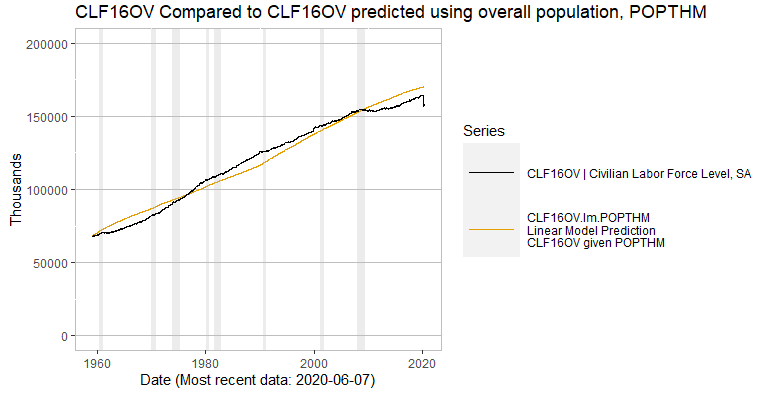


Figure - Labor force estimated from U.S. population.

# Appendix C – Excerpts from BLS report

## March Report, dated 3 April 2020

However, there was also a large increase in the number of workers who were classified as employed but

absent from work. Special instructions sent to household survey interviewers just before data collection

started for March called for all employed persons absent from work due to coronavirus-related business

closures to be classified as unemployed on temporary layoff. However, it is apparent that not all such

workers were so classified. Such a misclassification is an example of nonsampling error and can occur

when respondents misunderstand questions or interviewers record answers incorrectly.

If the workers who were recorded as employed but absent from work due to “other reasons” (over and

above the number absent for other reasons in a typical March) had been classified as unemployed on

temporary layoff, the overall unemployment rate would have been almost 1 percentage point higher than reported. However, according to usual practice, the data from the household survey are accepted as recorded. To maintain data integrity, no ad hoc actions are taken to reclassify survey responses.

In the establishment survey, workers who are paid by their employer for all or any part of the pay period including the 12th of the month are counted as employed, even if they were not actually at their jobs. Workers who are temporarily or permanently absent from their jobs, but are not being paid, are not counted as employed, even if they are continuing to receive benefits. The length of the reference period does vary across the respondents in the establishment survey; one-third of respondents have a weekly pay period, slightly over 40 percent a bi-weekly, about 20 percent semi-monthly, and a small amount monthly. Data collection for both surveys was affected by the coronavirus. The household survey is generally collected through in-person and telephone interviews, but personal interviews were suspended during the collection period for the safety of interviewers and respondents. The household survey response rate, at 73 percent, was about 10 percentage points lower than in recent months. In the establishment survey, about one-fifth of the data is generally collected by telephone at four regional data collection centers. Although these centers were closed during the collection period, efforts were made to collect data electronically. The collection rate for the establishment survey, at 66 percent, was about 9 percentage points lower than average.

## April Report, dated 8 May. 2020

However, there was also a large increase in the number of workers who were classified as employed but

absent from work. As was the case in March, special instructions sent to household survey interviewers

called for all employed persons absent from work due to coronavirus-related business closures to be

classified as unemployed on temporary layoff. However, it is apparent that not all such workers were so

classified.

If the workers who were recorded as employed but absent from work due to “other reasons” (over and

above the number absent for other reasons in a typical April) had been classified as unemployed on

temporary layoff, the overall unemployment rate would have been almost 5 percentage points higher

than reported (on a not seasonally adjusted basis). However, according to usual practice, the data from the household survey are accepted as recorded. To maintain data integrity, no ad hoc actions are taken to reclassify survey responses.

## May Report, dated 5 Jun 2020

However, there was also a large number of workers who were classified as employed but absent from

work. As was the case in March and April, household survey interviewers were instructed to classify

employed persons absent from work due to coronavirus-related business closures as unemployed on

temporary layoff. However, it is apparent that not all such workers were so classified. BLS and the

Census Bureau are investigating why this misclassification error continues to occur and are taking

additional steps to address the issue.

If the workers who were recorded as employed but absent from work due to “other reasons” (over and

above the number absent for other reasons in a typical May) had been classified as unemployed on

temporary layoff, the overall unemployment rate would have been about 3 percentage points higher than reported (on a not seasonally adjusted basis). However, according to usual practice, the data from the household survey are accepted as recorded. To maintain data integrity, no ad hoc actions are taken to reclassify survey responses.

# Appendix D – Residual Histograms

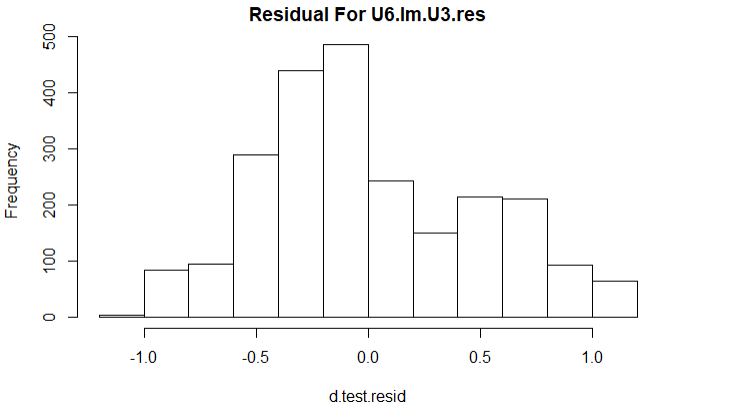


Figure 11 - Residual histogram for U6 predicted from U3 using a linear model.

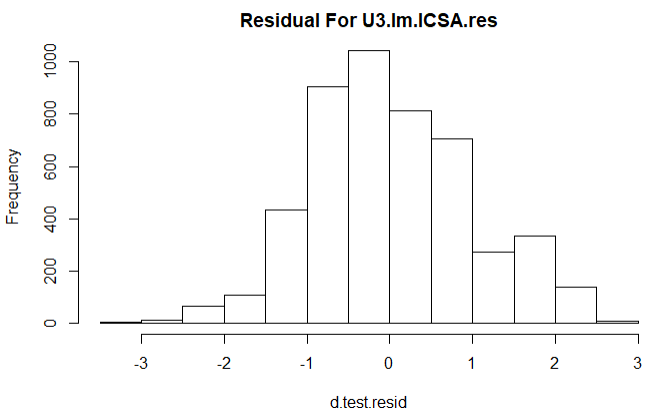


Figure 12 - Residual histogram for U3 predicted from ICSA using a linear model.

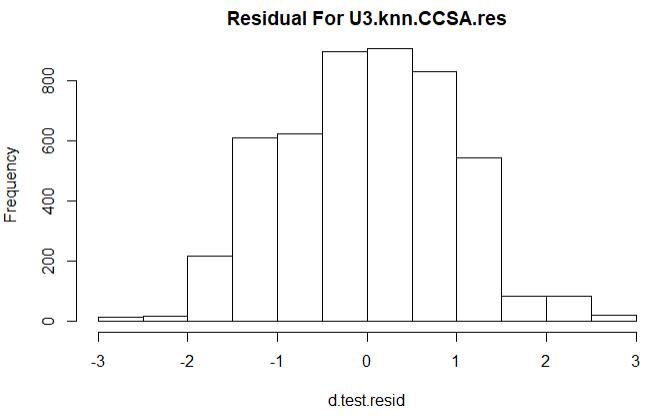


Figure 13 - Residual histogram for U3 predicted from CCSA using a linear model.