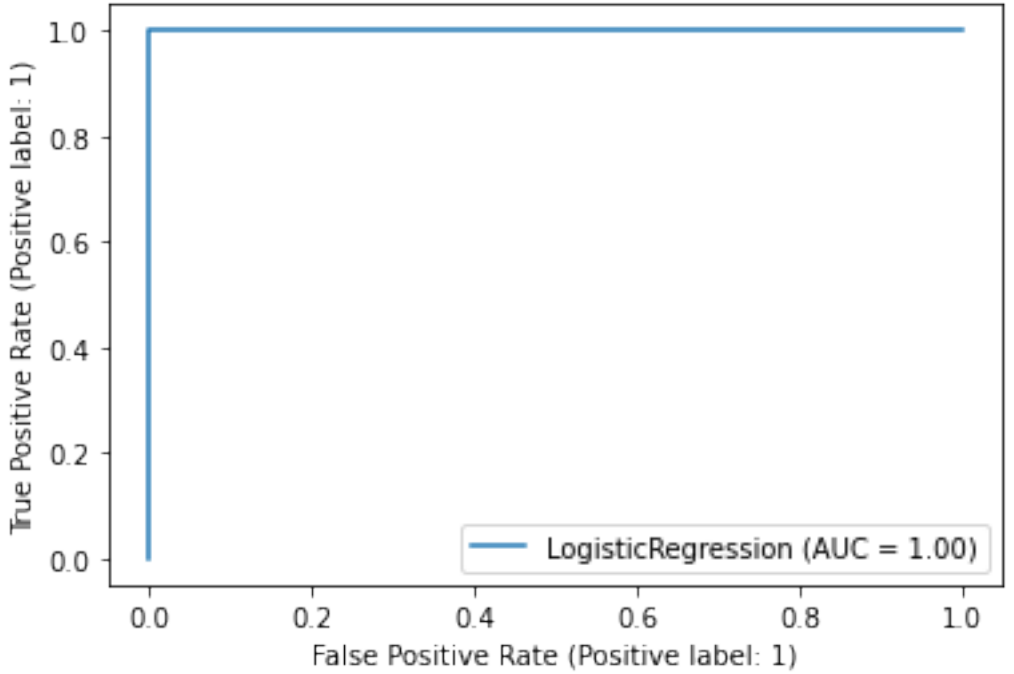
Week 04 Report

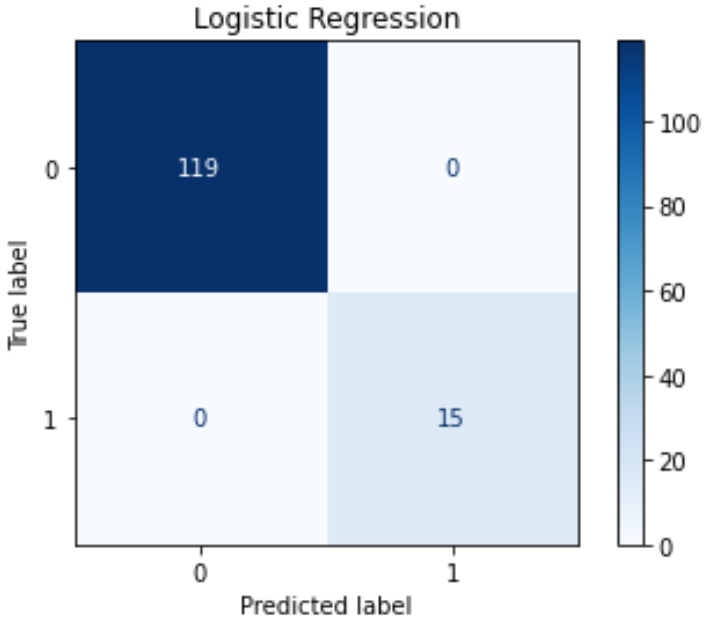
Jeffrey Li

Previous Week

- Kept the new feature ‘diff’ but **removed Alph1 and Alph2 as features**, since they may be redundant.
  - Total of 3 features.
- Kept features as continuous but maintained binary labels for target.
- Perfect AUC score.
  - The model perfectly predicts small chi and large chi with no errors.
- Potential Issue:
  - Model may benefit from being further trained/tested with more data or noise.

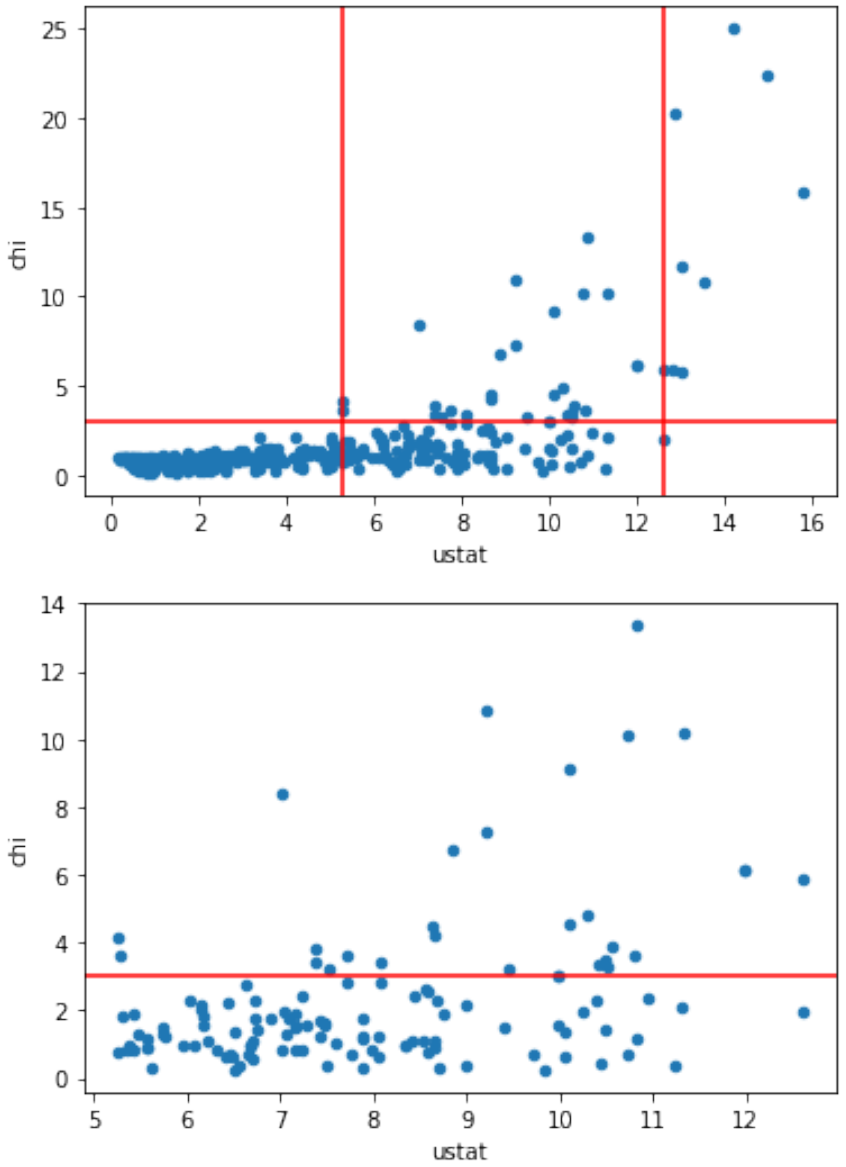
```
Logistic Regression train set score:  0.9925925925925926
Logistic Regression test set score:  1.0
```





Previous Week

- Process only data instances that lie between the overlapping scatter area seen on the top right.
  - Min ustat value for large chi: 5.26.
  - Max ustat value for small chi: 12.6.
- Only data instances containing ustat value **between 5.26 and 12.6** are considered.
  - See Bottom Right.
  - 128 data instances: 98 small chi and 30 large chi.

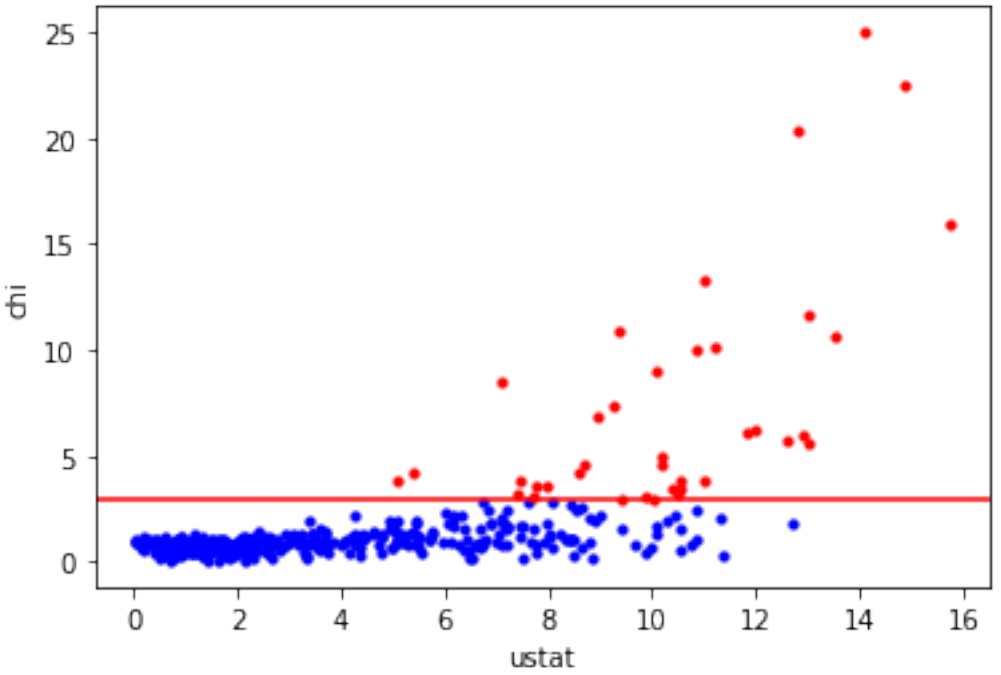


Updates

- Introduced Artificial Noise Data
- Examined Overlap Data

Artificial Data

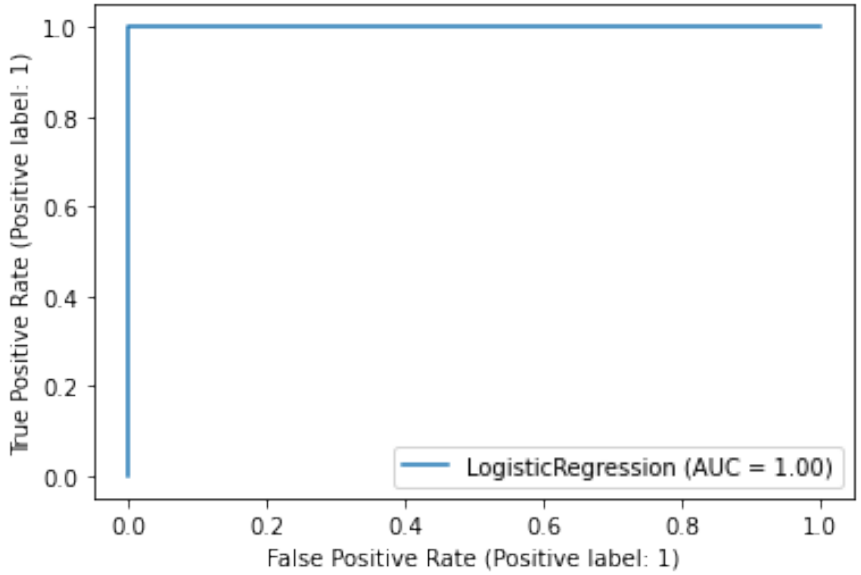
- Artificial data was generated by taking the original data and **adding noise**.
- Noise
  - Created with the same dimension as the original dataset.
  - **numpy.random.normal:** draws sample from normal distribution
    - Mean: 0, Standard Deviation: 0.1
- Artificial data plotted on right.

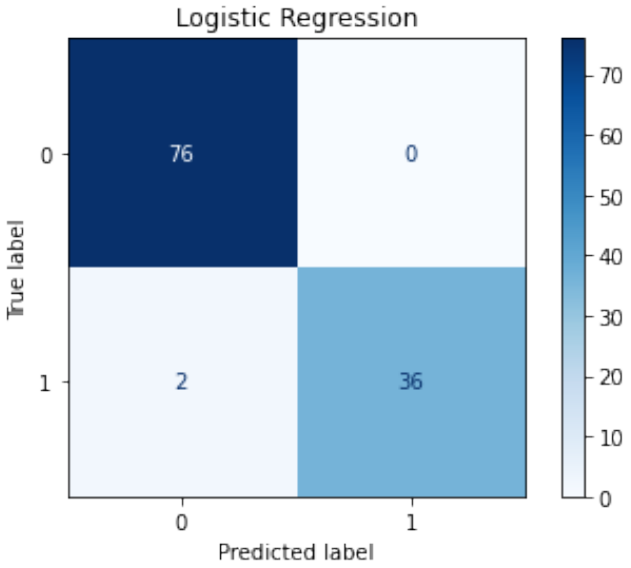


Idea 1: Using Artificial Data in Testing

- Use artificial data as testing data to see how it performs on more unseen data.
  - If the model performs poorly, it may be suffering from overfitting.
- Testing data: 114 entries
  - 38 noisy large chi.
  - 76 randomly selected noisy small chi.

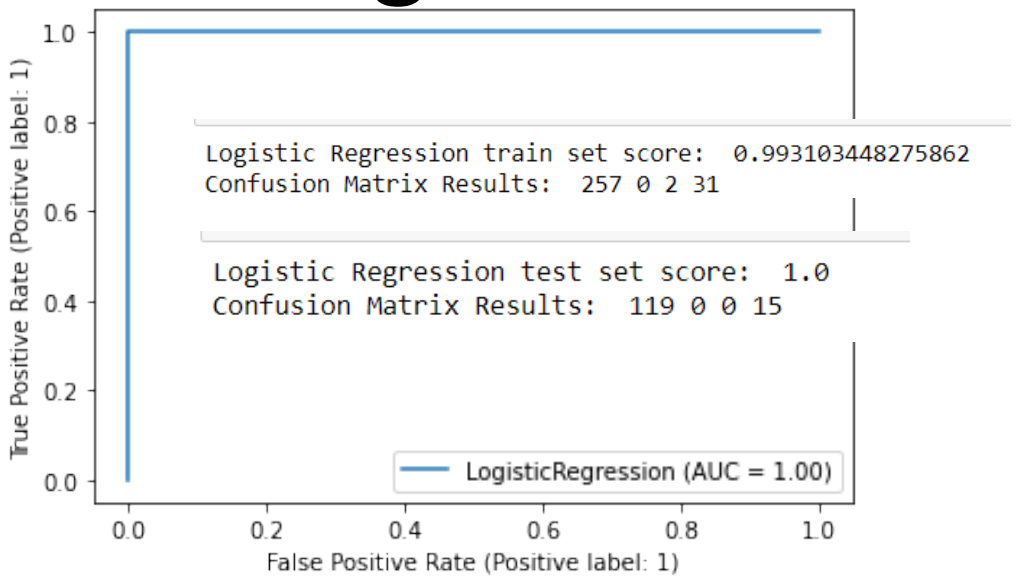
```
Logistic Regression test set score:  0.9824561403508771
Confusion Matrix Results:  76 0 2 36
```

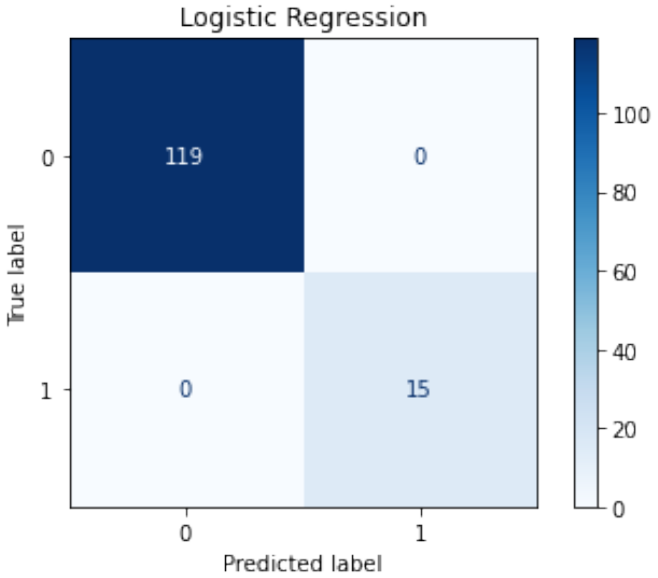




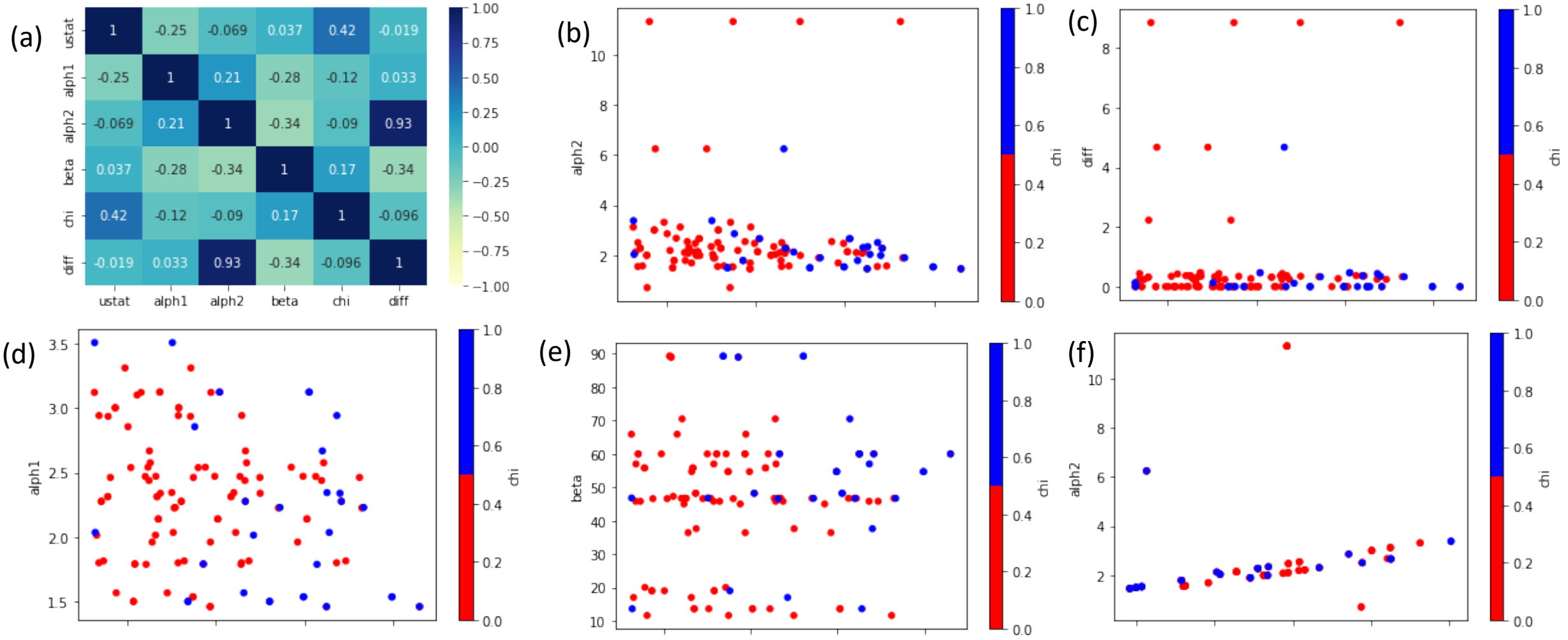
Idea 2: Using Artificial Data in Training

- Use artificial data into the training data to regularize the model.
  - Adding noise during training can make the training process more robust and reduce overfitting.
- Training data: 290 entries
  - Training data had 270 entries after train_test_split (67:33 ratio).
  - Added 10 random noisy large chi.
  - Added 10 random noisy small chi.



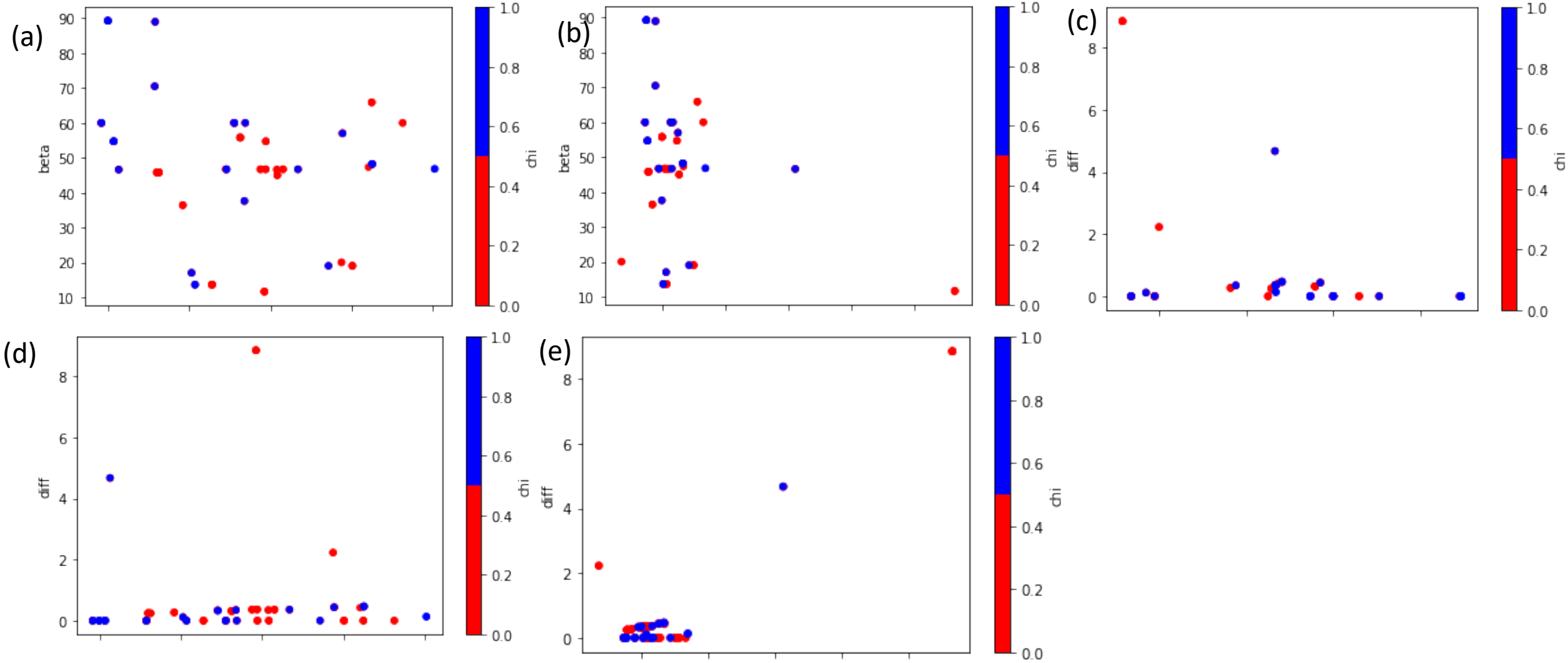


Examination of Overlap Data



(a) Correlation Matrix: Note high correlation between alph2 and diff (b) ustat v alph2 (c) ustat v diff (d) ustat v alph1 (e) ustat v beta (f) alph1 v alph2

Examination of Overlap Data (cont.)



(a) alph1 v beta (b) alph2 v beta (c) beta v diff (d) alph1 v diff (e) alph2 v diff