Homework 6

CS498df

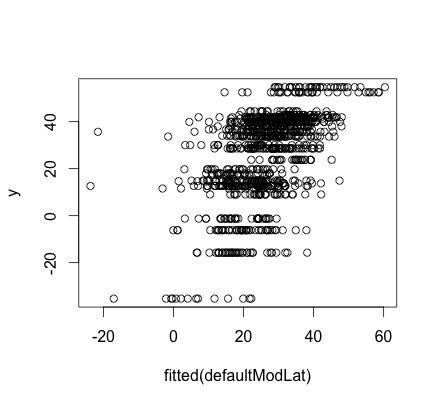
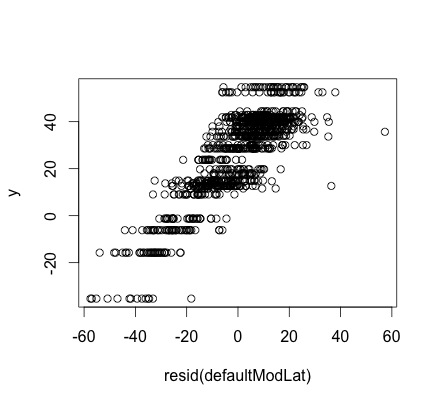
David Young, Varun Somani and Cybelle Smith

* 1. R^2 of latitude and longitude linear regressions against features:

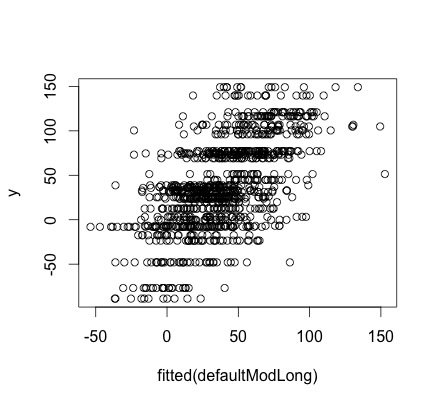
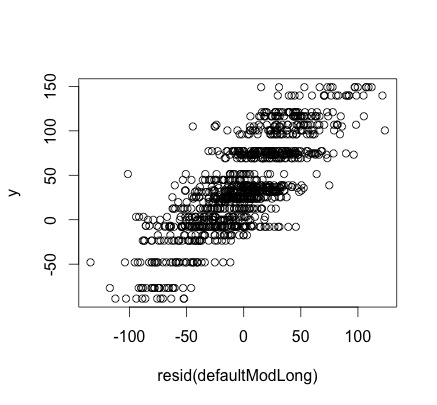
latitude: R^2 = 0.2412

longitude: R^2 = 0.3182

a. Residuals by Actual Latitude b. Predicted by Actual Latitude



c. Residuals by Actual Longitude d. Predicted by Actual Longitude



* 1. We performed a boxcox transformation and found that applying it did not substantially improve performance of the model, either for longitude or latitude. The R^2 values for the models did not improve much; in fact for longitude it went down slightly:

boxcox transformed latitude: R^2 = 0.2546 (previously, 0.2412)

boxcox transformed longitude: R^2 = 0.3159 (previously, 0.3182)

* 1. For latitude and longitude, we tried 10 different values of alpha, ranging from 0 to 1 in increments of .1, where alpha = 0 indicates ridge regression and alpha=1 indicates lasso regression (and alpha values >0 but <1 indicate an ‘elastic net’ regression was applied). We also obtained cross validated MSE for the unregularized regression for comparison. All cross validation used 10 folds. All regularized regressions improved a lot on the unregularized model that contained all variables as predictors, approximately halving the MSE, and they were quite close to each other in terms of performance. lambda.min below indicates the lambda value (i.e. the regularization coefficient) that showed the best performance, ncoeffs indicates the number of coefficients that were kept in each model.

**unregularized:**

latitude – cross-validated MSE: 550

longitude – cross-validated MSE: 3934

**regularized:**

**alpha: 0 (ridge regression)**

latitude -- lambda.min: 7.6381632495138

longitude -- lambda.min: 3.73325959734027

latitude -- ncoeffs: 116

longitude -- ncoeffs: 116

latitude -- cross-validated MSE: 281.91133440168

longitude -- cross-validated MSE: 1890.58057894255

**alpha: 0.1**

latitude -- lambda.min: 3.80153184810457

longitude -- lambda.min: 2.69571545684487

latitude -- ncoeffs: 38

longitude -- ncoeffs: 85

latitude -- cross-validated MSE: 276.364087251967

longitude -- cross-validated MSE: 1863.1853181416

**alpha: 0.2**

latitude -- lambda.min: 2.28947928472423

longitude -- lambda.min: 1.01960494231266

latitude -- ncoeffs: 29

longitude -- ncoeffs: 81

latitude -- cross-validated MSE: 279.845357989028

longitude -- cross-validated MSE: 1873.82631311344

**alpha: 0.3**

latitude -- lambda.min: 1.52631952314949

longitude -- lambda.min: 0.898571818948289

latitude -- ncoeffs: 22

longitude -- ncoeffs: 78

latitude -- cross-validated MSE: 279.024481961714

longitude -- cross-validated MSE: 1894.33365633273

**alpha: 0.4**

latitude -- lambda.min: 1.14473964236212

longitude -- lambda.min: 0.55950758273938

latitude -- ncoeffs: 22

longitude -- ncoeffs: 70

latitude -- cross-validated MSE: 280.267024735496

longitude -- cross-validated MSE: 1868.31606921484

**alpha: 0.5**

latitude -- lambda.min: 1.00508027544271

longitude -- lambda.min: 1.03402806744638

latitude -- ncoeffs: 22

longitude -- ncoeffs: 70

latitude -- cross-validated MSE: 277.16046491693

longitude -- cross-validated MSE: 1871.33352379455

**alpha: 0.6**

latitude -- lambda.min: 1.10721414468274

longitude -- lambda.min: 0.493090731001299

latitude -- ncoeffs: 19

longitude -- ncoeffs: 70

latitude -- cross-validated MSE: 288.050325580771

longitude -- cross-validated MSE: 1877.18456749995

**alpha: 0.7**

latitude -- lambda.min: 0.717914482459077

longitude -- lambda.min: 0.385102208120696

latitude -- ncoeffs: 21

longitude -- ncoeffs: 69

latitude -- cross-validated MSE: 279.277663109394

longitude -- cross-validated MSE: 1873.07213224524

**alpha: 0.8**

latitude -- lambda.min: 0.572369821181058

longitude -- lambda.min: 0.336964432105608

latitude -- ncoeffs: 21

longitude -- ncoeffs: 67

latitude -- cross-validated MSE: 275.150264102725

longitude -- cross-validated MSE: 1882.24272958121

**alpha: 0.9**

latitude -- lambda.min: 0.558377930801505

longitude -- lambda.min: 0.29952393964943

latitude -- ncoeffs: 21

longitude -- ncoeffs: 70

latitude -- cross-validated MSE: 279.397689255809

longitude -- cross-validated MSE: 1911.47712791714

**alpha: 1 (lasso)**

latitude -- lambda.min: 0.502540137721354

longitude -- lambda.min: 0.245623552536189

latitude -- ncoeffs: 21

longitude -- ncoeffs: 39

latitude -- cross-validated MSE: 280.020558327408

longitude -- cross-validated MSE: 1882.16665665947

2. We used different regularization schemes and found that an elastic net with alpha = .3 worked the best. Our optimal model achieved ~81% accuracy using 10-fold cross validation and on an 80/20 train-test split.

10-fold cross validated models used to select optimal alpha:

**alpha: 0**

lambda.min: 0.0147950762551908

ncoeffs: 30

cross-validated MSE: 0.193666666666667

**alpha: 0.1**

lambda.min: 0.000951038319441683

ncoeffs: 30

cross-validated MSE: 0.1891

**alpha: 0.2**

lambda.min: 0.000757160975927511

ncoeffs: 29

cross-validated MSE: 0.189233333333333

**alpha: 0.3**

lambda.min: 0.000968111203646219

ncoeffs: 27

cross-validated MSE: 0.188966666666667

**alpha: 0.4**

lambda.min: 0.000726083402734664

ncoeffs: 27

cross-validated MSE: 0.189133333333333

**alpha: 0.5**

lambda.min: 0.00101508109364243

ncoeffs: 26

cross-validated MSE: 0.189533333333333

**alpha: 0.6**

lambda.min: 0.000639893018195456

ncoeffs: 28

cross-validated MSE: 0.189266666666667

**alpha: 0.7**

lambda.min: 0.000601955826443263

ncoeffs: 27

cross-validated MSE: 0.189233333333333

**alpha: 0.8**

lambda.min: 0.000526711348137854

ncoeffs: 27

cross-validated MSE: 0.189233333333333

**alpha: 0.9**

lambda.min: 0.00051383560386887

ncoeffs: 27

cross-validated MSE: 0.189233333333333

**alpha: 1**

lambda.min: 0.000383935810917274

ncoeffs: 27

cross-validated MSE: 0.189333333333333

Accuracy on model retrained with 80/20 train-test split at alpha = .3: 0.809.

Here are the estimated betas for the optimal model, indicating which variables were excluded. mar1, mar2 and mar3 are indicator variables for the original MARRIAGE variable, edu1-edu6 are indicator variables for the original EDUCATION variable, and SEX has been converted to an indicator variable as well:

(Intercept) -1.19e+00

LIMIT\_BAL -7.36e-07

SEX -1.27e-01

AGE 2.71e-03

PAY\_0 5.83e-01

PAY\_2 7.53e-02

PAY\_3 5.63e-02

PAY\_4 2.77e-02

PAY\_5 5.50e-02

PAY\_6 .

BILL\_AMT1 -2.03e-06

BILL\_AMT2 .

BILL\_AMT3 1.12e-08

BILL\_AMT4 .

BILL\_AMT5 8.67e-07

BILL\_AMT6 6.79e-09

PAY\_AMT1 -1.08e-05

PAY\_AMT2 -7.78e-06

PAY\_AMT3 -2.57e-06

PAY\_AMT4 -3.94e-06

PAY\_AMT5 -2.77e-06

PAY\_AMT6 -3.42e-06

edu1 7.61e-02

edu2 .

edu3 .

edu4 -9.49e-01

edu5 -9.67e-01

edu6 -2.54e-01

mar1 2.27e-01

mar2 .

mar3 1.53e-01