Homework 6

Applied Machine Learning–CS 498

Mengyu Xie, Adrian Bandolon

March 26, 2018

Problem 1:

- The data set used for this problem are features of music and the latitude and ***Longitude*** from which that music originates.
- This data set was from **UCI Machine Learning Repository.** The data set can be found here: **https://archive.ics.uci.edu/ml/datasets/Geographical+Original+of+Music**

Part 1:

- For this part, we attempt to build a linear regression model for ***Longitude*** and ***Latitude*** against the features.

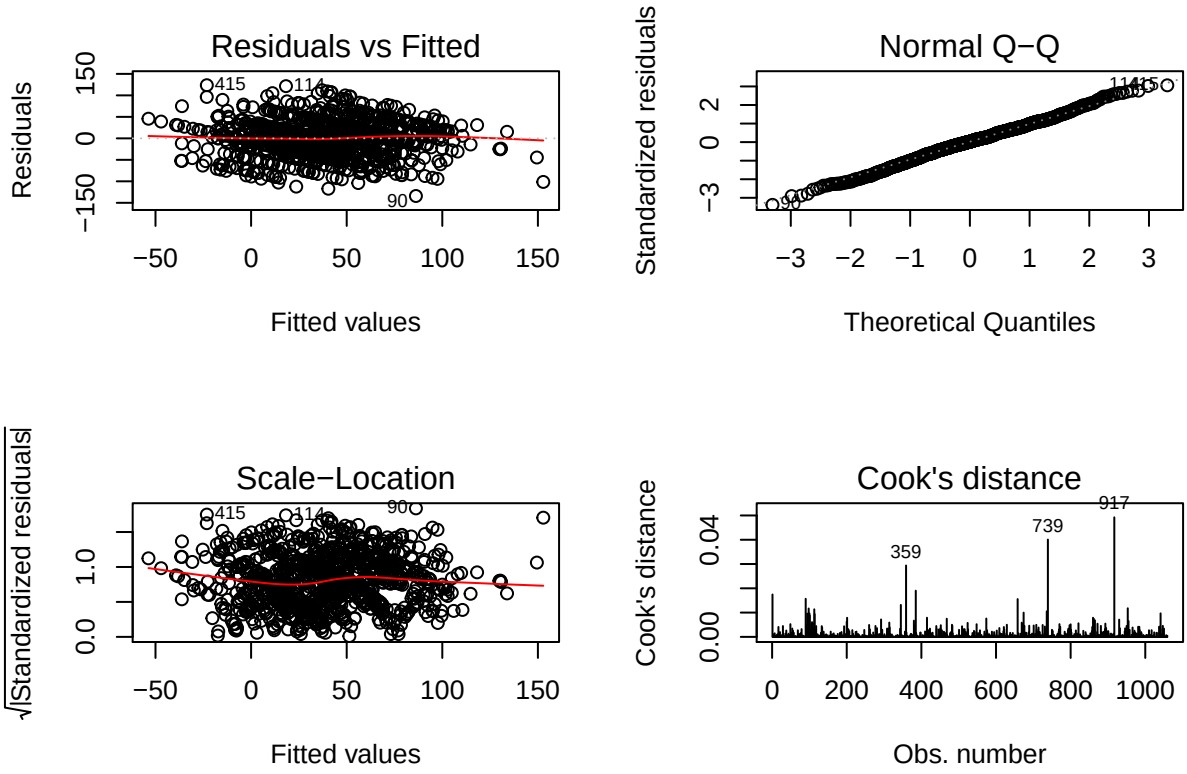

Figure 1. *Diagnostic plots for the unregularized linear regression model of Longitude vs. features*

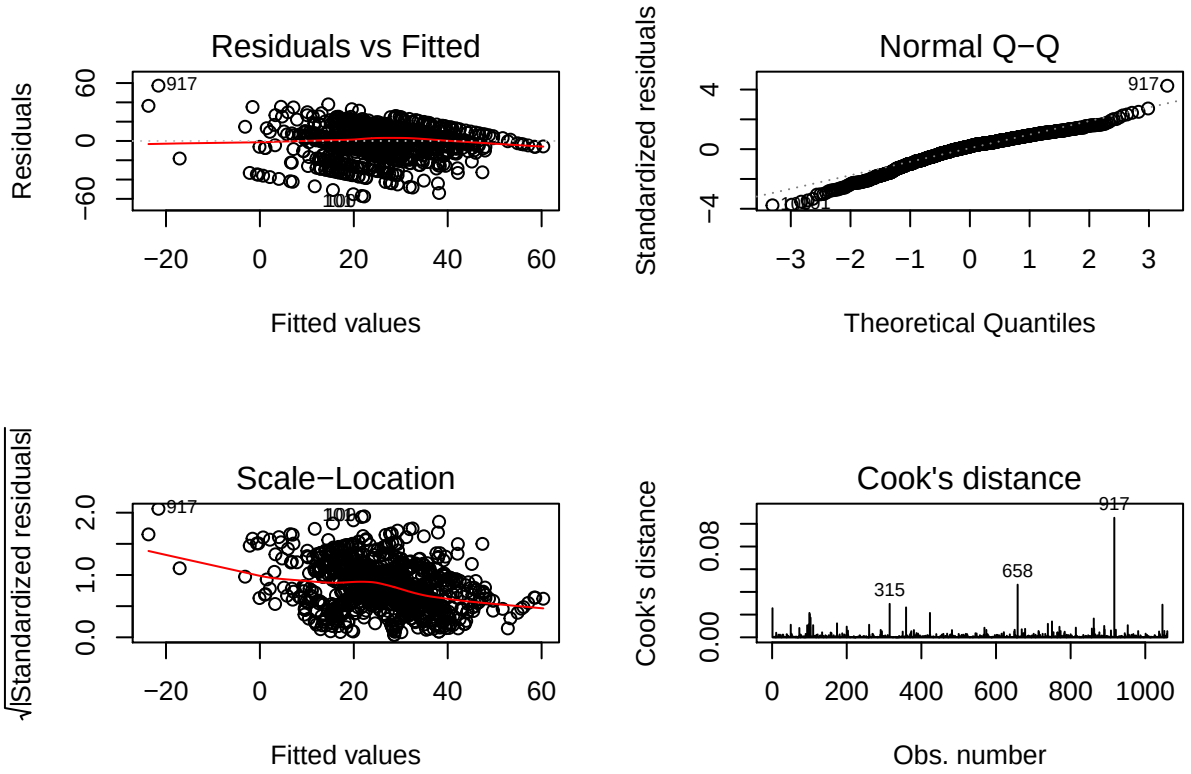

Figure 2. *Diagnostic plots for the unregularized linear regression model of Latitude vs. features*

- From the plots above we can see that there are some outliers that needs to be addressed, but we are ignoring these outliers in this exercise.

Part 2:

- For this part, we used a Box-Cox transformation to model ***Longitude*** and ***Latitude*** against the music features.
- Box-Cox does not work well with negative numbers. Constants were added to ***Longitude*** (180) and ***Latitude*** (90) to remove any negative values.

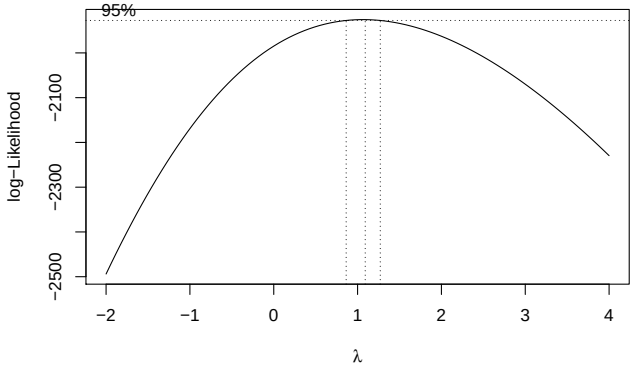

Figure 3. *Box-Cox transformation plot of lambda vs. log-likelihood for the Longitude model*

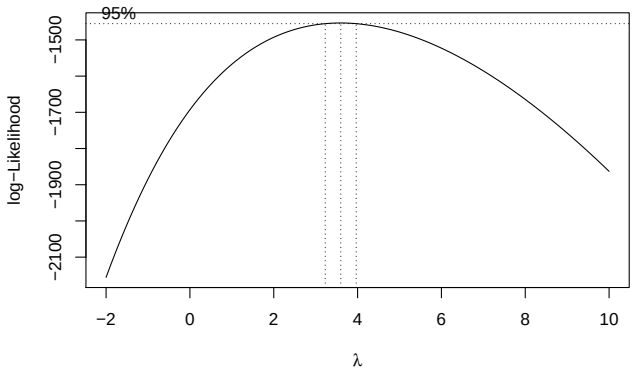

Figure 4. *Box-Cox transformation plot of lambda vs. log-likelihood for the Latitude model*

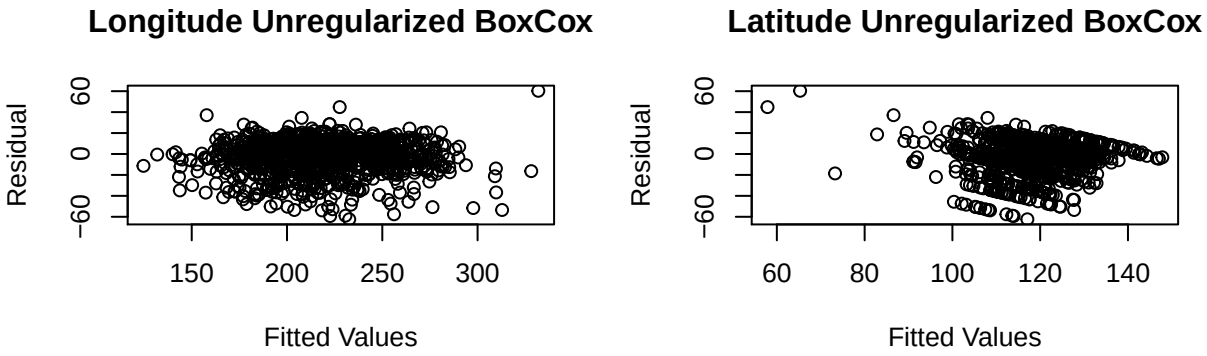

Figure 5. *Box-Cox transformation residual vs. fitted values for the Longitude (left) and Latitude (right) unregularized models.*

	Mean Squared Error	R-Squared
Longitude Linear Model	1613.82	0.36
Latitude Linear Model	240.75	0.29
Box-Cox Transformed Longitude Model	1613.89	0.36
Box-Cox Transformed Latitude Model	257.16	0.26

Table 1. *Mean Squared Error and R^2 values for the Unregularized and Box-Cox transformed Longitude and Latitude models of music features.*

- Based on the results from ***Table 1***, we can see that the Box-Cox transformed models does not improve on the unregularized linear model. Both MSE and R^2 values are only minimally different from each other.

Part 3:

- We used `cv.glmnet` to regularize our ***Longitude*** and ***Latitude*** models of music features and perform 10-fold cross-validation using different regularizers.
- One regularization parameter for `Ridge` ($\alpha = 0$) and `LASSO` ($\alpha = 1$), and three for `ElasticNet` ($\alpha = 0.2, 0.5, 0.7$) was used.
- For comparison, we used `cv.glm` to perform 10-fold cross-validation of the simple unregularized model.
- The regularization coefficient with the minimum error (Lambda Min), number of variables, and mean of mean-squared error from cross-validation (Mean MSE) for each model is provided in two tables: ***Table 2***, and ***Table 3***.

	Simple	Ridge	LASSO	Elastic Net (0.2)	Elastic Net (0.5)	Elastic Net (0.7)
Lambda Min		5.94	0.30	1.35	0.59	0.51
Number of Variables	116.00	116.00	55.00	92.00	85.00	82.00
Mean MSE	1891.79	1880.30	1862.57	1874.23	1871.99	1874.27

Table 2. *10 fold cross-validation of the Longitude simple unregularized model and models using different regularizers.*

	Simple	Ridge	LASSO	Elastic Net (0.2)	Elastic Net (0.5)	Elastic Net (0.7)
Lambda Min		4.37	0.42	2.09	0.92	0.65
Number of Variables	116.00	116.00	23.00	34.00	22.00	22.00
Mean MSE	291.95	278.55	280.60	276.86	279.64	277.22

Table 3. *10 fold cross-validation of the Latitude simple unregularized model and models using different regularizers.*

- The "best" regularization varies with the current random state (i.e. `set.seed(1234)`). `ElasticNet` with $\alpha = 0.2$ is the best for predicting ***Latitude***, while `LASSO` is the best for predicting ***Longitude***.
- Based on the mean-squared error values (***Table 2***, ***Table 3***) LASSO regularization is the best choice for this dataset.

Problem 2:

- We used the dataset that gives whether a Taiwanese credit card user defaults against a variety of features (found here: **http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients**)

Part 1:

- Here we used logistic regression to predict whether the customer defaults. Outliers were ignored.
- We split the dataset into training and test datasets.
- There were 23 variables that could be used to predict if a customer defaults. In an effort to reduce the number of variables used in this model we:

  1. From the training dataset we created a model with all the variables(`glm.full`).
  2. From `glm.full` we measured the ***Variance Inflation Factor (VIF)*** and used to this to remove from the model *collinear* variables (those with VIF < 2). Collinearity refers to the possibility of linear relationships among the explanatory variables. For models built for prediction, collinearity is not really an issue, but we feel that parsimony is important. Some authors suggest that VIF < 5 should be used to judge collinearity, we chose VIF < 2 simply to reduce the number of variables in our model further. This choice can be justified by the results we obtain (see ***Table 5*** below). `glm.vif` is the model produced in this step.

	GVIF	Df	GVIF^(1/(2*Df))
LIMIT_BAL	1.52	1.00	1.23
SEX	1.02	1.00	1.01
EDUCATION	1.23	6.00	1.02
MARRIAGE	1.34	3.00	1.05
AGE	1.39	1.00	1.18
PAY_0	1.51	1.00	1.23
PAY_2	2.73	1.00	1.65
PAY_3	3.31	1.00	1.82
PAY_4	3.84	1.00	1.96
PAY_5	4.29	1.00	2.07
PAY_6	3.09	1.00	1.76
BILL_AMT1	25.74	1.00	5.07
BILL_AMT2	40.06	1.00	6.33
BILL_AMT3	25.30	1.00	5.03
BILL_AMT4	24.33	1.00	4.93
BILL_AMT5	28.28	1.00	5.32
BILL_AMT6	16.16	1.00	4.02
PAY_AMT1	1.45	1.00	1.21
PAY_AMT2	1.41	1.00	1.19
PAY_AMT3	1.41	1.00	1.19
PAY_AMT4	1.46	1.00	1.21
PAY_AMT5	1.52	1.00	1.23
PAY_AMT6	1.12	1.00	1.06

Table 4. *Variance Inflation Factor values for the different variables in the* `glm.full` *model. Variables with* < 2 *GVIF were dropped from the model.*

3. The number of variables were further reduced using `drop1` where the least significant variable based on *likelihood ratio tests* was dropped from the `glm.vif` model. At this point there was a manageable number of variables. More variables could be added or taken away manually from the final model based on the increase or decrease in accuracy. The final model (`mod.fit`) we settled on was:

will_default ~ LIMIT_BAL + EDUCATION + AGE + PAY_0 + PAY_AMT1 + PAY_AMT2

4. To compare the `glm.full`, `glm.vif` and `mod.fit` we used the **test** dataset to run 10-fold crossvalidation. As can be seen from the table below, the differences in accuracy, sensitivity and specificity between the models are minimal. Again, the final model was chosen simply because it is concise and just as accurate.

	Accuracy	Specificity	Sensitivity
Full Model	0.8102	0.9728	0.2374
No Collinear Variables	0.8092	0.9704	0.2419
Final Model	0.8102	0.9695	0.2491

Table 5. *Accuracy, sensitivity and specificity of the full glm model (with all the variables), vif model (with variables that have a VIF< 2 removed) and final model using the test dataset.*

Part 2:

- Here we try to improve upon the final model we built in the previous section by using different regularizers. As aparent from the previous section, we only need to use some of the variables (6 in this case) to achieve similar predictive accuracy as the model that uses all the variables.
- For this part we only used the variables:

  LIMIT_BAL + EDUCATION + AGE + PAY_0 + PAY_AMT1 + PAY_AMT2

- As can be seen in the table below, there is a minimal (+0.0039) improvement in accuracy from the final model and the best regularized model (lasso).

	Accuracy	Specificity	Sensitivity
Full Model	0.8102	0.9728	0.2374
No Collinear Variables	0.8092	0.9704	0.2419
Final Model	0.8102	0.9695	0.2491
Ridge Regularized Model	0.7850	0.9953	0.0445
Lasso Regularized Model	0.8141	0.9708	0.2622

Table 6. *Accuracy, sensitivity and specificity of the full glm model (with all the variables), vif model (with variables that have a VIF< 2 removed), final model, ridge regularized model and lasso regularized model using the test dataset.*

	alpha	lambda
Ridge	0.00	0.10
Lasso	1.00	0.00

Table 7. *Parameters used in ridge and lasso regularization.*

References:

1. https://stackoverflow.com/questions/33999512/how-to-use-the-box-cox-power-transformation-in-r
2. https://stackoverflow.com/questions/40901445/function-to-calculate-r2-r-squared-in-r
3. https://stackoverflow.com/questions/26237688/rmse-root-mean-square-deviation-calculation-in-r
4. https://www4.stat.ncsu.edu/~post/josh/LASSO_Ridge_Elastic_Net_-_Examples.html
5. Zuur,A., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., ***Mixed Effects Models and Extensions in Ecology with R.*** pg. 246-252; 327-339. Springer, 2011