DNSC 6290
Summer 2025
Assignment 1

Due: 28 May 2025, 11:59:59 PM ET
Email to: jphall@gwu.edu

1. (1 pt.) The linear regression model below in Equation 1 is used to estimate medical insurance premiums in dollars. What is the exact interpretation of the coefficient for age in years?

$$premium = 400 + 2.1 * blood_pressure + 18 * age + 2.3 * blood_sugar$$

Equation 1

2. (1 pt.) The logistic regression model below in Equation 2 is used to estimate the probability of a medical insurance claim. What is the exact interpretation of the coefficient for age in years?

$$claim = -1.7 + 3.1 * blood_pressure + 1.5 * age + 2.4 * blood_sugar$$

Equation 2

3. (1 pt.) You are managing a machine learning project. You see that the data scientists on the project are using traditional OLS regression on a wide, dirty data set. You would like to suggest a more contemporary, and potentially more appropriate, approach. You remember learning about elastic net regression in machine learning class, but you'll need to convince your team it's worth evaluating. Draw a line between the two columns below to match the advantages of elastic net regression to the contemporary technique that elastic net incorporates.

Technique Name	Advantage
Iteratively reweighted least squares	Automatic variable selection for wide data
L1/LASSO regularization	More stable model in the presence of correlation
L2/Ridge regularization	More stable model in the presence of outliers

4. (1 pt.) After fitting a logistic regression model, you notice a numeric coefficient with a value of -20.2368.

What is the odds ratio associated with this coefficient?

Should you use this model?

DNSC 6290
Summer 2025
Assignment 1

5. (1 pt.) Circle the correct answers: For a dataset with *P* >> *N*, **(elastic net / LASSO)** regression can select only *N* features, whereas **(elastic net / LASSO)** can select greater than *N* features.

6. (1 pt.) Figure 1 below displays the trace plot for a LASSO model that was trained for 220 iterations. The best performance in validation data was observed at 100 iterations.

How many features are selected for the model at 100 iterations?

At 100 iterations, how many features have positive standardized coefficients?

At 100 iterations, how many features have negative standardized coefficients?

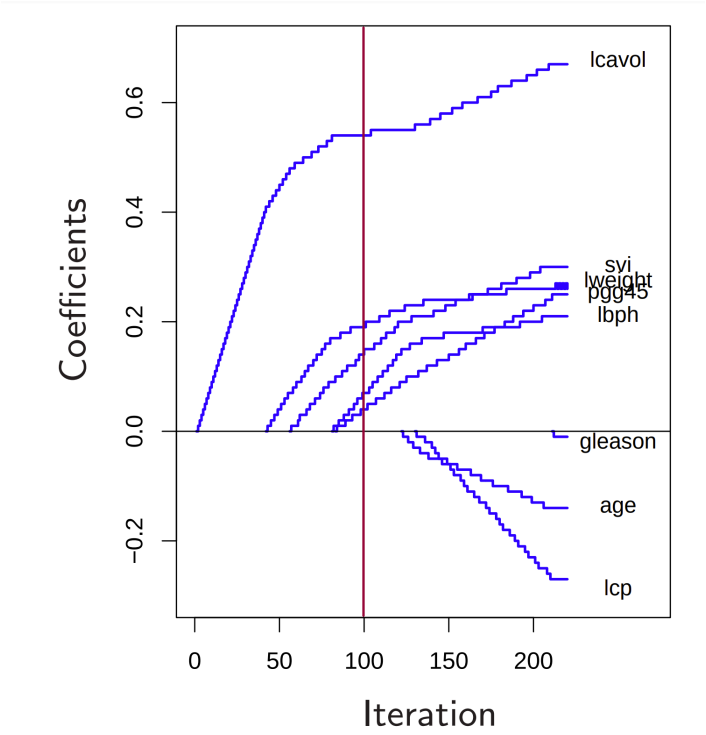


Figure 1: A LASSO trace plot with a reference line for the best performing model in validation data. Adapted from *Elements of Statistical Learning*, Figure 3.19.

DNSC 6290
Summer 2025
Assignment 1

7. (1 pt.) Circle the term of the Normal Equation below (Equation 3) that is compromised by strong correlation between input features.

$$\boxdot = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\,\boldsymbol{y}$$

Equation 3

Rewrite the equation and include the term that equips the formula for L2/ridge regression and addresses the common correlated inputs problem.

For questions 8 - 10 refer to the colab notebook that accompanies the assignment. Copy the notebook to your own Google drive. Then …

- Upload loan_clean.csv used in the class example. (Cell 5-7)
- Set the X/input features to GRP_REP_home_ownership, GRP_addr_state, GRP_purpose, GRP_verification_status, STD_IMP_REP_annual_inc, STD_IMP_REP_delinq_2yrs, STD_IMP_REP_dti, STD_IMP_REP_emp_length, STD_IMP_REP_int_rate, STD_IMP_REP_loan_amnt, STD_IMP_REP_longest_credit_lengt, STD_IMP_REP_revol_util, STD_IMP_REP_term_length, STD_IMP_REP_total_acc. (Cell 11)
- Set the y/target feature to bad_loan. (Cell 11)
- Set the glm_grid() function to consider the following values for α: 0.01, 0.25, 0.5, 0.99. (Cell 13)
- Use the correct function call for the glm_grid() function. (Cell 15)
- After training, set the most important feature after training and viewing results. (Cell 18)
- After training, input the correct data and use the correct function call to generate a prediction. (Cell 22)

The notebook will perform many regressions in a grid search to find the best values of α and λ for the data. It will generate various results and plots. Use the notebook to fill in the answers below.

DNSC 6290
Summer 2025
Assignment 1

8. (1 pt.) What are the selected features and their coefficient values?

Coefficient Name	Value

9. (1 pt.) Which is the most important feature?

DNSC 6290
Summer 2025
Assignment 1

10. (1 pt.) What is the prediction for the following new customer? If the decision threshold is 0.15, should you lend to this customer?

Feature	Value
GRP_REP_home_ownership	3
GRP_addr_state	20
GRP_purpose	5
GRP_verification_status	3
STD_IMP_REP_annual_inc	1
STD_IMP_REP_delinq_2yrs	5
STD_IMP_REP_dti	3
STD_IMP_REP_emp_length	-2
STD_IMP_REP_int_rate	4
STD_IMP_REP_loan_amnt	3
STD_IMP_REP_longest_credit_lengt	4
STD_IMP_REP_revol_util	3
STD_IMP_REP_term_length	3
STD_IMP_REP_total_acc	-2