7390–Machine Learning and Data Sciences

Northeastern University, Spring 2018

Midterm, due on Sunday July 15, 2018

- The exam is a group assignment.
- Discussing the exam with other teams is *not* permitted.
- Justifications are only needed for those problems for which we ask it. For all other problems, justifications need not be given, nor will they result in partial credit.
- Write all the team member full names and IDs on the submission.
- Late submissions will *not* be accepted. No exceptions.
- The solutions should be submitted on the website on the day the exam is due.
- Each team should submit one submission. If a team has multiple submissions, either by one person or by different members, the first submission will be graded.

1. (Total: 15 points) This problem involves the sales data set for Toyota Corolla, which can be found in the file ToyotaCorolla.csv. The data set contains 1436 observations on the following 10 variables.

   Price (in Dollars)

   Age (in months)

   Mileage

   FuelType Fuel Type (diesel, petrol, CNG)

   MetColor Metallic color (1=yes, 0=no)

   Automatic Automatic transmission (1=yes, 0=no)

   Displacement Engine displacement (in cu. inches)

   Doors Number of doors

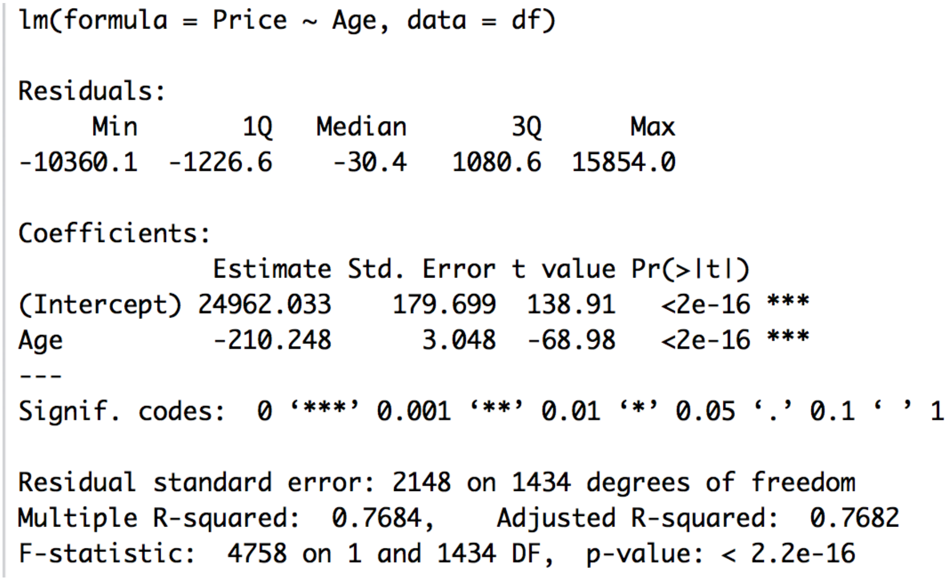
   Weight (in pounds)

   Horsepower Engine horsepower

   (a) (4 points) Fit a simple linear regression with Price as the response and Age as the predictor.

   (i) Is there a relationship between the predictor and the response?

   Solution: The p-value corresponding to the T-statistic in the summary output is very low, indicating clear evidence of a relationship between Age and Price.



```
lm(formula = Price ~ Age, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-10360.1  -1226.6    -30.4   1080.6  15854.0 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24962.033    179.699  138.91   <2e-16 ***
Age          -210.248      3.048  -68.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2148 on 1434 degrees of freedom
Multiple R-squared:  0.7684,	Adjusted R-squared:  0.7682 
F-statistic:  4758 on 1 and 1434 DF,  p-value: < 2.2e-16
```

(ii) How strong is the relationship between the predictor and the response?

Solution: The predictor explains almost 76.8% of the variance in price indicating a strong relationship between price and age.

(iii) What is the predicted price associated for a car with an age of 30 months? What are the associated 90% confidence intervals? What are the associated 95% confidence intervals?

Solution:
From the linear regression we have

$$\text{Price} \simeq 24962.033 - 210.248 \times \text{Age}$$

Then the estimated price would be

$$\text{Price} \simeq 24962.033 - 210.248 \times 30 \simeq \$18654.59$$

The standard error is 2147.62474. Hence,

$$\textbf{Price} \sim N(24962.033 - 210.248 \times \text{Age}, (2147.62474)^2)$$

Then, 95% confidence interval is given by

$$\begin{aligned}&(24962.033 - 210.248 \times \text{Age} - 1.96 * 2147.62474,\\ &\ 24962.033 - 210.248 \times \text{Age} + 1.96 * 2147.62474)\\ &\ = (14445.25, 22863.94)\end{aligned}$$

and 90% confidence interval by

$$\begin{aligned}&(24962.033 - 210.248 \times \text{Age} - 1.645 * 2147.62474,\\ &\ 24962.033 - 210.248 \times \text{Age} + 1.645 * 2147.62474)\\ &\ = (15121.75, 22187.44)\end{aligned}$$

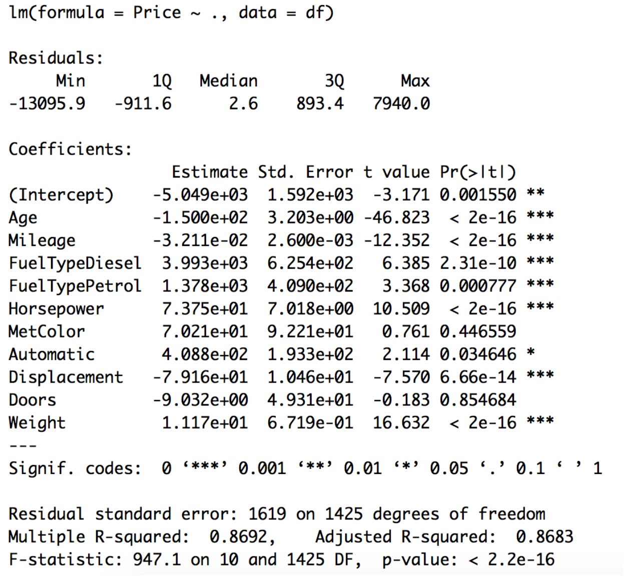
(b) (8 points) Fit a multiple linear regression with Price as the response and all other variables the predictors.

(i) Is there a relationship between the predictors and the response?

Solution: This question can be answered by testing the hypothesis

$$H_0 : \beta_{\text{Mileage}} = \beta_{\text{Age}} = \ldots \beta_{\text{Mileage}} = 0.$$

The F-statistic can be used to determine whether or not we should reject this null hypothesis. The p-value corresponding to the F-statistic in the summary output is very low, indicating clear evidence of a relationship between Price and the predictors.



```
lm(formula = Price ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-13095.9   -911.6      2.6    893.4   7940.0 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -5.049e+03  1.592e+03  -3.171 0.001550 ** 
Age            -1.500e+02  3.203e+00 -46.823  < 2e-16 ***
Mileage        -3.211e-02  2.600e-03 -12.352  < 2e-16 ***
FuelTypeDiesel  3.993e+03  6.254e+02   6.385 2.31e-10 ***
FuelTypePetrol  1.378e+03  4.090e+02   3.368 0.000777 ***
Horsepower      7.375e+01  7.018e+00  10.509  < 2e-16 ***
MetColor        7.021e+01  9.221e+01   0.761 0.446559    
Automatic       4.088e+02  1.933e+02   2.114 0.034646 *  
Displacement   -7.916e+01  1.046e+01  -7.570 6.66e-14 ***
Doors          -9.032e+00  4.931e+01  -0.183 0.854684    
Weight          1.117e+01  6.719e-01  16.632  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1619 on 1425 degrees of freedom
Multiple R-squared:  0.8692,	Adjusted R-squared:  0.8683 
F-statistic: 947.1 on 10 and 1425 DF,  p-value: < 2.2e-16
```

(ii) How strong is the relationship between the predictors and the response?

Solution: The predictors explain almost 86.9% of the variance in price, indicating a strong relationship between price and the predictors.

(iii) Which predictors appear to have a statistically significant relationship to the response?

Solution: The p-value corresponding to the T-statistic is very low for the following variables, indicating clear evidence of a relationship between Price and these predictors:

Age

Mileage

FuelType

Automatic

Displacement

Weight

Horsepower

assuming $\alpha = 0.05$. If we set $\alpha = 0.01$, then we can remove Automatic from this list.

(iv) What does the coefficient for the age variable suggest? How accurate can you estimate the effect of age on price?

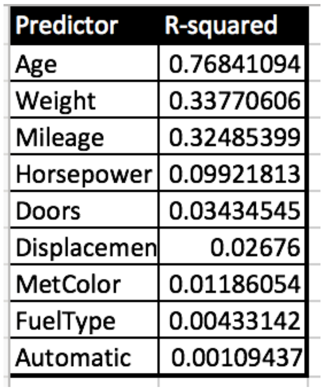
Solution: We have $\hat{\beta_{\text{Age}}} = -149.98 \simeq 150$ indicating the price is estimated to decrease by \$ 150 if Age is increased by 1 month assuming all other variables remain the same.
The standard error of $\hat{\beta_{\text{Age}}}$ can be used to construct confidence intervals for β_{Age}. Then, the 95% confidence interval for the effect of age on price is: (-156, -144). So the price is estimated to decrease between \$144 and 156 if Age is increased by 1 month assuming all other variables remain the same.

(v) What is the predicted price associated for a car with a mileage of 45000 miles, 30 months, diesel, automatic transmission, 4 doors, 2568 pounds, a displacement of 122 cu. inches, a horsepower of 90, and non-metallic color? What are the associated 90% confidence intervals? What are the associated 95% confidence intervals?

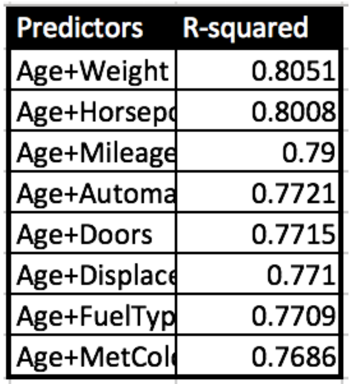
Solution: The estimate price is \$ 19,047.82 with 95% confidence interval (15874.61, 22221.04) and 90% confidence interval (16384.59, 21711.06)

(c) (3 points) Which predictors predictors matter most for predicting the price for a car? (Find the first and the second most important variables)

Solution: The first most important variable is AGE and the second most important variable is Weight.



Predictor	R-squared
Age	0.76841094
Weight	0.33770606
Mileage	0.32485399
Horsepower	0.09921813
Doors	0.03434545
Displacemen	0.02676
MetColor	0.01186054
FuelType	0.00433142
Automatic	0.00109437



Predictors	R-squared
Age+Weight	0.8051
Age+Horsepo	0.8008
Age+Mileage	0.79
Age+Automa	0.7721
Age+Doors	0.7715
Age+Displace	0.771
Age+FuelTyp	0.7709
Age+MetCol	0.7686

2. (Total: 10 points) This problem involves the Boston data set, which can be found in the file Boston.csv. This data set contains the following columns:

crim per capita crime rate by town.

zn proportion of residential land zoned for lots over 25,000 sq.ft.

indus proportion of non-retail business acres per town.

chas Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox nitrogen oxides concentration (parts per 10 million).

rm average number of rooms per dwelling.

age proportion of owner-occupied units built prior to 1940.

dis weighted mean of distances to five Boston employment centres.

rad index of accessibility to radial highways.

tax full-value property-tax rate per $10,000.

ptratio pupil-teacher ratio by town.

black $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

lstat lower status of the population (percent).

medv median value of owner-occupied homes in $1000s.

We want to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

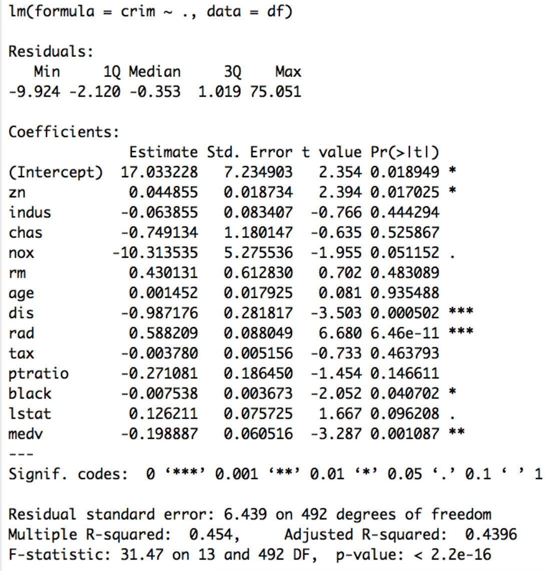
(a) (7 points) Fit a multiple regression model to predict the response using all of the predictors. Describe your results.

(i) Is there a relationship between the predictors and the response?

Solution: This question can be answered by testing the hypothesis

$$H_0 : \beta_{\text{zn}} = \beta_{\text{induc}} = \ldots \beta_{\text{medv}} = 0.$$

The F-statistic can be used to determine whether or not we should reject this null hypothesis. The p-value corresponding to the F-statistic is very low, indicating clear evidence of a relationship between the response and the predictors.



```
lm(formula = crim ~ ., data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-9.924 -2.120 -0.353  1.019 75.051 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.033228   7.234903   2.354 0.018949 *  
zn            0.044855   0.018734   2.394 0.017025 *  
indus        -0.063855   0.083407  -0.766 0.444294    
chas         -0.749134   1.180147  -0.635 0.525867    
nox         -10.313535   5.275536  -1.955 0.051152 .  
rm            0.430131   0.612830   0.702 0.483089    
age           0.001452   0.017925   0.081 0.935488    
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793    
ptratio      -0.271081   0.186450  -1.454 0.146611    
black        -0.007538   0.003673  -2.052 0.040702 *  
lstat         0.126211   0.075725   1.667 0.096208 .  
medv         -0.198887   0.060516  -3.287 0.001087 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,	Adjusted R-squared:  0.4396 
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

(ii) How strong is the relationship between the predictors and the response?

Solution: The predictors explain 45.4% of the variance in crim, indicating a weak relationship between the responce and the predictors.

(iii) What does the coefficient for the medv variable suggest? How accurate can you estimate the effect of medv on per capita crime rate?

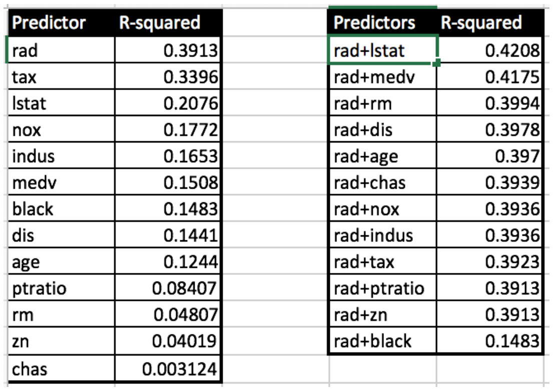
Solution: We have $\hat{\beta}_{\text{medv}} = -0.198887$ indicating per capita crime rate is estimated to decrease by 0.199 if the median value of owner-occupied homes is increased by \$1,000 assuming all other variables remain the same.
The standard error of $\hat{\beta}_{\text{medv}}$ can be used to construct confidence intervals for β_{medv}. Then, the 95% confidence interval for the effect of medv on per capita crime rate: (-0.317788478, -0.079985165). So the per capita crime rate is estimated to decrease between 0.07998 and 0.3178 if the median value of owner-occupied homes is increased by \$1,000 assuming all other variables remain the same.

(iv) For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

Solution: The p-value corresponding to the T-statistic for zn, dis, rad, ptratio, black, and medv is very low, indicating clear evidence of a relationship between the response and these predictors.

(b) (3 points) Which predictors predictors matter most for predicting per capita crime? (Find the first and the second most important variables)

Solution: The first most important variable is rad and the second most important variable is lstat.



Predictor	R-squared
rad	0.3913
tax	0.3396
lstat	0.2076
nox	0.1772
indus	0.1653
medv	0.1508
black	0.1483
dis	0.1441
age	0.1244
ptratio	0.08407
rm	0.04807
zn	0.04019
chas	0.003124

Predictors	R-squared
rad+lstat	0.4208
rad+medv	0.4175
rad+rm	0.3994
rad+dis	0.3978
rad+age	0.397
rad+chas	0.3939
rad+nox	0.3936
rad+indus	0.3936
rad+tax	0.3923
rad+ptratio	0.3913
rad+zn	0.3913
rad+black	0.1483

3. (Total: 8 points)

Suppose we collect data for a group of bank customers with variables

- default A variable with levels No and Yes indicating whether the customer defaulted on their debt
- student A variable with levels No and Yes indicating whether the customer is a student
- balance The average balance that the customer has remaining on their credit card after making their monthly payment
- income Income of customer

We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -13.75$, $\hat{\beta}_{\text{studentYes}} =$ -0.5149, $\hat{\beta}_{\text{balance}} = 0.003738$, $\hat{\beta}_{\text{income}} = -0.00000261$.

(a) (2 points) Estimate the probability that a student with a balance of $3,500 and income $75,000 does default on a loan.

Solution:

(b) (2 points) Estimate the probability that borrower a balance of $3,500 and income $75,000 who is not a student does default on a loan.

Solution:

(c) (2 points) How much income would the student in part (a) need to make to have a 85% chance of getting approved for a loan?

Solution:

(d) (2 points) How much income would the borrower in part (b) need to make to have a 85% chance of getting approved for a loan?

Solution:

4. (Total: 65 points) When a customer applies for a loan, banks and other credit providers use statistical models to determine whether or not to grant the loan based on the likelihood of the loan being repaid. The factors involved in determining this likelihood are complex, and extensive statistical analysis and modeling are required to predict the outcome for each individual case. You should analyze a loan dataset (in the file loan_dataset_training.csv) using logistic regression to predict loan repayment or default based on the data provided. The dataset consists of 82,176 loan records including the following fields:

- Loan ID: A unique Identifier for the loan information.
- Customer ID: A unique identifier for the customer. Customers may have more than one loan.
- Loan Status: A categorical variable indicating if the loan was paid back or defaulted. A Target variable
- Current Loan Amount: This is the loan amount that was either completely paid off, or the amount that was defaulted.
- Term: A categorical variable indicating if it is a short term or long term loan.
- Credit Score: A value between 0 and 800 indicating the riskiness of the borrowers credit history.
- Years in current job: A categorical variable indicating how many years the customer has been in their current job.
- Home Ownership: Categorical variable indicating home ownership. Values are“Rent”,“Home Mortgage”, and “Own”. If the value is OWN, then the customer is a home owner with no mortgage
- Annual Income: The customer’s annual income
- Purpose: A description of the purpose of the loan.
- Monthly Debt: The customer’s monthly payment for their existing loans
- Years of Credit History: The years since the first entry in the customerss credit history
- Months since last delinquent: Months since the last loan delinquent payment

- Number of Open Accounts: The total number of open credit cards
- Number of Credit Problems: The number of credit problems in the customer records.
- Current Credit Balance: The current total debt for the customer
- Maximum Open Credit: The maximum credit limit for all credit sources.
- Bankruptcies: The number of bankruptcies
- Tax Liens: The number of tax liens.

(a) (15 points) Analyze, process, clean the dataset and produce some numerical and graphical summaries of data. Answer the following questions:

  (i) Do there appear to be any patterns?

  (ii) What important fields and information does the dataset have?

  (iii) How do you clean the data and fill in the missing data?

(b) (5 points) Use 100 bootstrap samples to estimate the the standard errors of the coefficients from the logistic regression. Provide a 68% and 95% confidence interval for the model parameters of the logistic regression.

(c) (5 points) Calculate p-values for each feature. Do any of the features appear to be statistically significant? If so, which ones?

(d) (5 points) Use 5-fold cross-validation to estimate test AUC using all features. What test AUC do you obtain?

(e) (10 points) Fit a multiple regression model to calculate the probability that a loan will default using all of the features using the entire dataset. Use the test data set to answer the following questions:

  (i) Draw the ROC curve. What is the AUC?

  (ii) Compute the confusion matrix, accuracy, balanced accuracy, FP, and FP rates for Threshold=0.5.

  (iii) Determine the threshold for which (TP rate + (1-FP rate)) is maximal. Then, compute the confusion matrix, accuracy, balanced accuracy, FP, and FP rates for that threshold.

  (iv) Determine the threshold for which the ROC curve has the minimum distance to the upper left corner (where TP rate=1 and FP rate=0). Note that this distance is

$$\sqrt{(1 - \text{TP rate})^2 + (\text{FP rate})^2}.$$

  Then, compute the confusion matrix, accuracy, balanced accuracy, FP, and FP rates for that threshold.

(f) (15 points) Use *Forward Stepwise Selection* approach for feature selection. In each iteration, add the variable that results in the highest AUC. Then, use 5-fold cross-validation to choose the best model among $M_0, M_1, \ldots, M_m$. Report the features and AUC that appears to provide the best results. Use the test data set (data in file loan_dataset_test.csv) to answer the following questions:

  (i) Draw the ROC curve. What is the AUC?

  (ii) Compute the confusion matrix, accuracy, balanced accuracy, FP, and FP rates for Threshold=0.5.

  (iii) Determine the threshold for which (TP rate + (1-FP rate)) is maximal. Then, compute the confusion matrix, accuracy, balanced accuracy, FP, and FP rates for that threshold.

  (iv) Determine the threshold for which the ROC curve has the minimum distance to the upper left corner. Then, compute the confusion matrix, accuracy, balanced accuracy, FP, and FP rates for that threshold.

(g) (10 points) Use your best model and calculate the probability that a loan will default for the data given in the validation set (data in file loan_dataset_validation.csv). Submit a file including two columns: Loan_id and score (the probability of the probability that a loan will default).