Extra Problems Solutions

Stephanie DeGraaf

December 6, 2018

Problem 1

This is not complicated to prove, but the algebra is rather extensive. A nice writeup can be found here.

Problem 2

We can write the RSS as

$$RSS = \sum_{i \in G_1} (y_i - \bar{y}_1) + \sum_{i \in G_2} (y_i - \bar{y}_2).$$

We can write the TSS as

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y}) = \sum_{i \in G_1} (y - \bar{y}) + \sum_{i \in G_2} (y_i - \bar{y}).$$

Then $\sum_{i \in G_1} (y_i - \bar{y}_1) \leq \sum_{i \in G_1} (y - \bar{y})$ and $\sum_{i \in G_2} (y_i - \bar{y}_2) \leq \sum_{i \in G_2} (y_i - \bar{y})$. Hence $RSS \leq TSS$.

Problem 3

a) The design matrix will be $X = \mathbf{1}_n$. Thus, $\tilde{\beta} = (n + \lambda)^{-1} \sum_{i=1}^{n} y_i = \frac{n}{n+\lambda} \bar{Y}$.

b) Each y_i has expectation β, so

$$E(\tilde{\beta}) = E((n + \lambda)^{-1} \sum_{i=1}^{n} y_i) = \frac{1}{n + \lambda} \sum_{i=1}^{n} E(y_i) = \frac{n}{n + \lambda} \beta.$$

Each y_i has variance σ^2, so

$$Var(\tilde{\beta}) = Var((n + \lambda)^{-1} \sum_{i=1}^{n} y_i) = \frac{1}{(n + \lambda)^2} \sum_{i=1}^{n} Var(y_i) = \frac{n}{(n + \lambda)^2} \sigma^2.$$

c)

$$\begin{aligned}
MSE(\tilde{\beta}) &= E(\tilde{\beta} - \beta)^2 \\
&= E(\tilde{\beta}^2 - 2\tilde{\beta}\beta + \beta^2) \\
&= Var(\tilde{\beta}) + E(\tilde{\beta})^2 - 2\beta E(\tilde{\beta}) + \beta^2 \\
&= \frac{n}{(n + \lambda)^2} \sigma^2 + (\frac{n}{n + \lambda} \beta)^2 - 2\beta \frac{n}{n + \lambda} \beta + \beta^2 \\
&= \frac{n}{(n + \lambda)^2} \sigma^2 + \beta^2 (\alpha^2 - 2\alpha + 1) \\
&= \frac{\alpha^2 \sigma^2}{n} + (1 - \alpha)^2 \beta^2.
\end{aligned}$$

The MSE of the OLS estimator is σ^2/n, so $MSE(\tilde{\beta}) < MSE(\bar{Y})$ if and only if

$$\begin{aligned}
MSE(\tilde{\beta}) &< MSE(\bar{Y}) \\
\frac{\alpha^2\sigma^2}{n} + (1-\alpha)^2\beta^2 &< \sigma^2/n \\
(1-\alpha)^2\beta^2 &< \frac{\sigma^2}{n}(1-\alpha^2) \\
\frac{\beta^2}{\sigma^2} &< \frac{(1-\alpha)(1+\alpha)}{n(1-\alpha)^2} \\
\frac{\beta^2}{\sigma^2} &< \frac{1+\alpha}{n(1-\alpha)}.
\end{aligned}$$

Problem 4

a) We can interpret the rpart output as follows:

The root node contains all 891 observations, the predicted class for this node is 0 (not survived), there are 342 passengers that survived in this node, and the predicted probabilities are 0.616 (not survived) and 0.3838 (survived).

The tree then splits according to whether the passenger had Fare $<$ 10.48. This creates 2 nodes: Node 2 if Fare $<$ 10.48 and Node 3 if Fare $>=$ 10.48. Node 2 has 339 observations, the predicted class for this node is 0, there are 67 passengers that survived in this node, and the predicted probabilities are 0.802 (not survived) and 0.1876 (survived). Node 2 is a terminal node. Node 3 has 552 passengers, the predicted class for this node is 0, there are 275 passengers that survived, and the predicted probabilities are 0.502 (not survived) and 0.498 (survived).

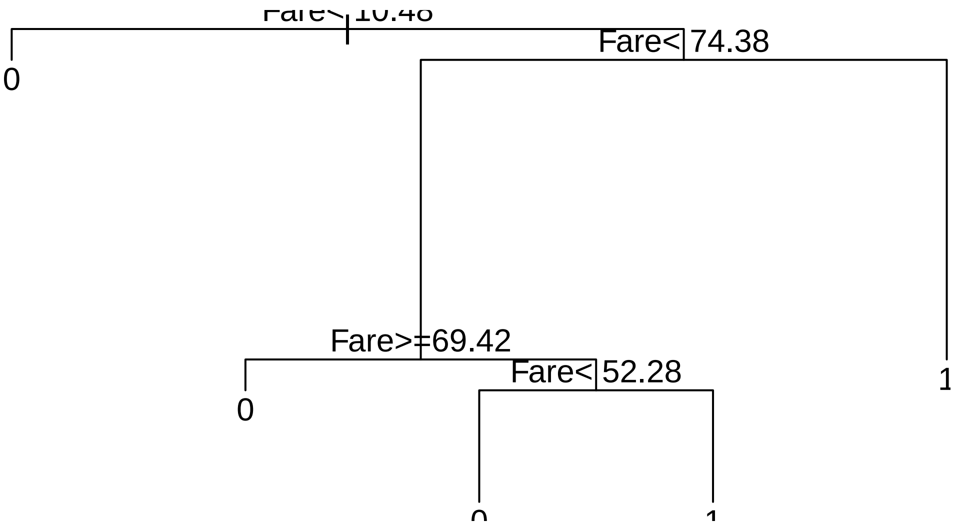
The tree then splits according to Fare $<$ 74.375. This creates 2 nodes: Node 6 if Fare $<$ 74.375, and Node 7 if Fare $>=$ 74.375. Node 6 has 455 passengers. We are given that the probability of survival is 0.4417582, so we can compute that $0.4417582 * 455 = 201$ passengers survived. The predicted class for this node is 0, since this class has the higher probability of 0.558.

For Node 7, the predicted class is 1 (survived), with a probability of 0.76288. There are 23 passengers that did not survive, and we know that this makes up 0.23711 of all the passengers in this node. So the total number of passengers in this node is $23/0.2371134 = 97$.

The rest of the nodes can be interpreted similarly. For Node number 26, we see that the predicted class is 0, so we know that 171 passengers survived out of the total 403. We can then compute the probabilities as $((403 - 171)/403, 171/403)$.

The completed output is shown here:

```
## n= 891
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 891 342 0 (0.6161616 0.3838384)
##    2) Fare< 10.48125 339  67 0 (0.8023599 0.1976401) *
##    3) Fare>=10.48125 552 275 0 (0.5018116 0.4981884)
##      6) Fare< 74.375 455 201 0 (0.5582418 0.4417582)
##       12) Fare>=69.425 15   2 0 (0.8666667 0.1333333) *
##       13) Fare< 69.425 440 199 0 (0.5477273 0.4522727)
##         26) Fare< 52.2771 403 171 0 (0.5756824 0.4243176) *
##         27) Fare>=52.2771 37   9 1 (0.2432432 0.7567568) *
##      7) Fare>=74.375 97  23 1 (0.2371134 0.7628866) *
```



c) A passenger with a ticket fare of 50 will end up in Node 26 (Fare <52.2771). The predicted probability of survival in this node is 0.4243176.

d) We can build our confusion matrix from the terminal nodes in the rpart output, which gives us the number of passengers correctly and incorrectly predicted at each terminal node. The confusion matrix will be:

```
##              True0    True1
## Pred0 272+13+232 67+2+171
## Pred1       23+9    74+28
```

```
##       True0 True1
## Pred0   517   240
## Pred1    32   102
```

From the confusion matrix, we calculate precision = 102/(240+102) = 0.2982456, and recall = 102/(32+103) = 0.761194.

e) This passenger belongs to node 107, which has probability of survival 0.8125.

f) This passenger belongs to node 7, which has probability of survival 0.94705882.

Problem 5

The likelihood for each Y_i is given by the Poisson distribution,

$$f(y_i|\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

The log likelihood of our data is

$$l(\beta_0) = \sum_{i=1}^{n} [\beta_0 Y_i - \exp(\beta_0) - \log Y_i!].$$

Taking the derivative with respect to β_0, we find

$$dl/d\beta_0 = \sum_{i=1}^{n} Y_i - n\exp(\beta_0) = 0,$$

so the MLE estimate for β_0 must be

$$\hat{\beta}_0 = \log \bar{Y}.$$

Problem 7

The formula for adjusted R squared is $1 - (1 - R^2)(n-1)/(n-p-1)$. Since R^2 is between 0 and 1, and $(n-1)/(n-p-1)$ is always positive, the second term will be nonnegative. 1 minus a nonnegative term will always be less than or equal to 1, so adjusted R squared is always less than or equal to 1.

Problem 8

The link function for a Poisson GLM is the log. The log likelihood of our data is

$$l(\beta) = \sum_{i=1}^{n} [Y_i \log \mu_i - \mu_i - \log Y_i!].$$

We can find MLE estimates by using Newton's method to maximize $l(\beta)$.

Problem 9

The link function is the inverse cdf of a standard normal. The log likelihood of our data is

$$l(\beta) = \sum_{i=1}^{n} [Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i)].$$

We can find MLE estimates by using Newton's method to maximize $l(\beta)$.