Conor Desmond

Credit: Worked with Kyle Mulkins

**Problem Set 6**

1. I loaded in the sparrows data set, which contains various characteristics from 1898 of many house sparrows on the ground after a winter storm. Some survived the storm, and others died from it. This data set contains 11 variables, two of them are Boolean variables and many of them are highly correlated. The FL variable, (measures length of femur in inches), and HL variable, (measures length of humerus in inches), are the two variables with the highest correlation together, (not counting pairing a variable with itself), therefore I graphed those two variables together.

> sparrows <- read.csv(file.choose(), stringsAsFactors = T)

> head(sparrows)

Status AG TL AE WT BH HL FL TT SK KL

1 Survived Adult 154 241 24.5 31.2 0.69 0.67 1.02 0.59 0.83

2 Survived Adult 160 252 26.9 30.8 0.74 0.71 1.18 0.60 0.84

3 Survived Adult 155 243 26.9 30.6 0.73 0.70 1.15 0.60 0.85

4 Survived Adult 154 245 24.3 31.7 0.74 0.69 1.15 0.58 0.84

5 Survived Adult 156 247 24.1 31.5 0.71 0.71 1.13 0.57 0.82

6 Survived Adult 161 253 26.5 31.8 0.78 0.74 1.14 0.61 0.89

> summary(sparrows)

Status AG TL AE WT

Perished:36 Adult :59 Min. :153.0 Min. :236.0 Min. :23.2

Survived:51 Juvenile:28 1st Qu.:158.0 1st Qu.:245.0 1st Qu.:24.7

Median :160.0 Median :247.0 Median :25.8

Mean :160.4 Mean :247.5 Mean :25.8

3rd Qu.:162.5 3rd Qu.:251.0 3rd Qu.:26.7

Max. :167.0 Max. :256.0 Max. :31.0

BH HL FL TT

Min. :29.80 Min. :0.6600 Min. :0.6500 Min. :1.010

1st Qu.:31.40 1st Qu.:0.7250 1st Qu.:0.7000 1st Qu.:1.110

Median :31.70 Median :0.7400 Median :0.7100 Median :1.130

Mean :31.64 Mean :0.7353 Mean :0.7134 Mean :1.131

3rd Qu.:32.10 3rd Qu.:0.7500 3rd Qu.:0.7300 3rd Qu.:1.160

Max. :33.00 Max. :0.7800 Max. :0.7600 Max. :1.230

SK KL

Min. :0.5600 Min. :0.7700

1st Qu.:0.5900 1st Qu.:0.8300

Median :0.6000 Median :0.8500

Mean :0.6032 Mean :0.8511

3rd Qu.:0.6100 3rd Qu.:0.8800

Max. :0.6400 Max. :0.9300

> library(ggplot2)

> round(cor(sparrows[,-c(1,2)]),2)

TL AE WT BH HL FL TT SK KL

TL 1.00 0.62 0.53 0.30 0.39 0.42 0.37 0.41 0.32

AE 0.62 1.00 0.50 0.37 0.73 0.70 0.63 0.39 0.46

WT 0.53 0.50 1.00 0.42 0.47 0.44 0.47 0.35 0.41

BH 0.30 0.37 0.42 1.00 0.52 0.53 0.52 0.40 0.46

HL 0.39 0.73 0.47 0.52 1.00 0.84 0.74 0.44 0.49

FL 0.42 0.70 0.44 0.53 0.84 1.00 0.79 0.43 0.46

TT 0.37 0.63 0.47 0.52 0.74 0.79 1.00 0.35 0.42

SK 0.41 0.39 0.35 0.40 0.44 0.43 0.35 1.00 0.25

KL 0.32 0.46 0.41 0.46 0.49 0.46 0.42 0.25 1.00

1. cont. I graphed FL on the x-axis with HL on the y-axis and saw that the two variables appeared linearly correlated, therefore I did not make any transformations on the graph. I then made another plot with FL on the x-axis and HL on the y-axis, but this time I had Status, (variable for whether the sparrow survived or perished), determine the color of the points. This plot contained two graphs with the one graph being for adult sparrows and the other one being for juvenile sparrows. Both graphs seemed linearly correlated as well, therefore I did not make any transformations on these graphs.

> ggplot(sparrows, aes(FL, HL)) + geom\_point(size = 4)

Chart, scatter chart

Description automatically generated

> ggplot(sparrows, aes(FL, HL, color = Status)) + geom\_point(size = 4)+ facet\_wrap(~AG)

Chart, scatter chart

Description automatically generated

1. I ran a summary of the glm of the full model for this data set. For this model I regressed AG, TL, AE, WT, BH, HL, FL, TT, SK, and KL onto Status. I removed AG from the model because AGjuvenile was not significant and it had the highest p-value of 0.876225. I ran a summary of the glm of the model without AG and then removed FL from the model because it had the highest p-value of 0.82724. I ran a summary of the glm of the model without AG and FL and then removed TT from the model because it was not significant and it had the highest p-value of 0.77200. I ran a summary of the glm of the model without AG and then removed FL from the model because it was not significant and it had the highest p-value of 0.82724. I ran a summary of the glm of the model without AG, FL, and TT and then removed AE from the model because it was not significant and it had the highest p-value of 0.48111. I ran a summary of the glm of the model without AG, FL, TT, and AE and then removed SK from the model because it was not significant and it had the highest p-value of 0.42906. I ran a summary of the glm of the model without AG, FL, TT, AE, and SK and then removed BH from the model because it was not significant and it had the highest p-value of 0.23716. I ran a summary of the glm of the model without AG, FL, TT, AE, SK, and BH and then did not remove any more variables because all of the variables were significant. Therefore, I believe a good logistic model of Status on a subset of variables is TL, WT, HL, and KL regressed onto Status.

> summary(glm(Status ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ AG + TL + AE + WT + BH + HL + FL + TT +

SK + KL, family = "binomial", data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2252 -0.5232 0.1397 0.5131 2.0134

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 27.56606 26.42412 1.043 0.296848

AGJuvenile 0.10631 0.68253 0.156 0.876225

TL -0.73634 0.18965 -3.883 0.000103 \*\*\*

AE 0.08275 0.12622 0.656 0.512060

WT -0.88860 0.34182 -2.600 0.009333 \*\*

BH 0.58293 0.59735 0.976 0.329131

HL 56.03494 31.05541 1.804 0.071176 .

FL -6.64680 31.73442 -0.209 0.834096

TT 5.05213 14.05263 0.360 0.719210

SK 21.53121 27.28482 0.789 0.430037

KL 23.56111 12.03826 1.957 0.050326 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.01 on 86 degrees of freedom

Residual deviance: 65.92 on 76 degrees of freedom

AIC: 87.92

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + AE + WT + BH + HL + FL + TT + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + AE + WT + BH + HL + FL + TT + SK +

KL, family = "binomial", data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1979 -0.5187 0.1380 0.5195 1.9932

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 27.6154 26.5200 1.041 0.29773

TL -0.7315 0.1865 -3.922 8.79e-05 \*\*\*

AE 0.0795 0.1246 0.638 0.52332

WT -0.8930 0.3417 -2.613 0.00897 \*\*

BH 0.5793 0.5973 0.970 0.33216

HL 56.3324 30.9729 1.819 0.06895 .

FL -6.9118 31.6706 -0.218 0.82724

TT 5.1116 14.1011 0.362 0.71698

SK 21.7633 27.2708 0.798 0.42485

KL 23.5683 12.0734 1.952 0.05093 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 65.945 on 77 degrees of freedom

AIC: 85.945

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + AE + WT + BH + HL + TT + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + AE + WT + BH + HL + TT + SK + KL,

family = "binomial", data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2211 -0.5397 0.1404 0.5014 1.9806

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 28.03611 26.36607 1.063 0.28763

TL -0.73512 0.18643 -3.943 8.04e-05 \*\*\*

AE 0.07978 0.12459 0.640 0.52196

WT -0.88694 0.33935 -2.614 0.00896 \*\*

BH 0.55984 0.59108 0.947 0.34356

HL 52.95693 26.71797 1.982 0.04747 \*

TT 3.39882 11.72971 0.290 0.77200

SK 22.04688 27.18725 0.811 0.41741

KL 23.41190 12.05676 1.942 0.05216 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 65.992 on 78 degrees of freedom

AIC: 83.992

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + AE + WT + BH + HL + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + AE + WT + BH + HL + SK + KL, family = "binomial",

data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2095 -0.5401 0.1474 0.5000 2.0130

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 26.0639 25.4931 1.022 0.30660

TL -0.7323 0.1854 -3.951 7.79e-05 \*\*\*

AE 0.0864 0.1226 0.705 0.48111

WT -0.8791 0.3373 -2.607 0.00915 \*\*

BH 0.5958 0.5761 1.034 0.30099

HL 56.0686 24.7957 2.261 0.02375 \*

SK 22.2755 27.1914 0.819 0.41267

KL 23.3670 12.0203 1.944 0.05190 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 66.075 on 79 degrees of freedom

AIC: 82.075

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + WT + BH + HL + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + WT + BH + HL + SK + KL, family = "binomial",

data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1559 -0.5221 0.1523 0.5308 1.9600

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 35.1445 21.8668 1.607 0.10801

TL -0.6916 0.1744 -3.966 7.31e-05 \*\*\*

WT -0.8473 0.3283 -2.581 0.00985 \*\*

BH 0.5345 0.5576 0.959 0.33780

HL 65.1081 21.3385 3.051 0.00228 \*\*

SK 20.9032 26.4328 0.791 0.42906

KL 24.6188 11.8527 2.077 0.03780 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 66.581 on 80 degrees of freedom

AIC: 80.581

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + WT + BH + HL + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + WT + BH + HL + KL, family = "binomial",

data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9635 -0.5645 0.1492 0.6004 2.1646

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 41.5162 20.1743 2.058 0.03960 \*

TL -0.6888 0.1758 -3.919 8.91e-05 \*\*\*

WT -0.8604 0.3240 -2.655 0.00792 \*\*

BH 0.6371 0.5390 1.182 0.23716

HL 68.0605 20.9756 3.245 0.00118 \*\*

KL 25.4462 11.9957 2.121 0.03390 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 67.214 on 81 degrees of freedom

AIC: 79.214

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + WT + HL + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + WT + HL + KL, family = "binomial",

data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2234 -0.5648 0.1540 0.6094 2.2701

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 49.9861 18.4879 2.704 0.006857 \*\*

TL -0.6573 0.1683 -3.907 9.35e-05 \*\*\*

WT -0.7896 0.3097 -2.549 0.010800 \*

HL 72.3327 20.7640 3.484 0.000495 \*\*\*

KL 27.3775 11.7780 2.324 0.020101 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 68.612 on 82 degrees of freedom

AIC: 78.612

Number of Fisher Scoring iterations: 6

1. I created a model matrix with the 10 variables not called Status regressed onto Status and stored it into a variable called x. I stored the Status values into a variable called y. I made a grid and plotted cv.glmnet of x and y with alpha equal to 1 and lambda equal to the grid. The plot shows me that the Mean-Squared Error is between about .95 and about 1.39 for each value of the log of lambda shown and that it tends to be about 1.39. The best lambda is found by finding the minimum lambda of cv.glmnet of x and y with alpha equal to 1 and lambda equal to the grid. I predicted using glmnet of x and y with alpha equal to 1 and lambda equal to the grid, type equal to the “coefficients”, and s equal to the best lambda. This gave me the lasso coefficients of some of the variables in the model. Using the lasso, a good model is TL, AE, WT, BH, HL, TT, SK, and KL regressed onto Status because these variables had lasso coefficients.

> x <- model.matrix(Status ~ ., sparrows)[,-1]

> y <- sparrows$Status

> grid <- 10^seq(10, -2, length = 100)

> cv.out <- cv.glmnet(x, y, alpha = 1, lambda = grid, family = "binomial")

> plot(cv.out)

Chart

Description automatically generated

> bestlam <- cv.out$lambda.min

> out <- glmnet(x, y, alpha = 1, lambda = grid, family = "binomial")

> lasso.coef <- predict(out, type = "coefficients", s = bestlam)

> lasso.coef

11 x 1 sparse Matrix of class "dgCMatrix"

s1

(Intercept) 30.627830188

AGJuvenile .

TL -0.498731224

AE 0.006239689

WT -0.522488726

BH 0.266010434

HL 43.499999502

FL .

TT 1.173040783

SK 12.213400799

KL 14.981704386

4 and 5. The model from the lasso is bigger, (has more variables), and has a higher AIC, 83.992, than the model from the elimination method using p-values, that one has an AIC of 78.612. Therefore, based on the AIC values of the two models, the model from the lasso explains has more information lost than the model from the elimination method using p-values. The model from the lasso has the variables AE, BH, TT, and SK in it, while the model from the elimination method using p-values does not. The p-values of AE (0.52196), BH (0.34356), TT (.77200), and SK (.41741) are non-significant.

The factor variable AG appears to be not important vis-à-vis survival of the sparrows because AGJuvenile has a high p-value of 0.876225 and it is not in either of the two good logistic models I found. The variables TL, WT, HL, and KL appear to be important vis-à-vis survival of the sparrows because they have significant p-values (all are less than 0.03) and had lasso coefficients. They seem to have an effect on whether a sparrow survives. Also, the variables AE, BH, TT, and SK appear to be important vis-à-vis survival of the sparrows because they had lasso coefficients. They also seem to have an effect on whether a sparrow survives.

> summary(glm(Status ~ TL + WT + HL + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + WT + HL + KL, family = "binomial",

data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2234 -0.5648 0.1540 0.6094 2.2701

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 49.9861 18.4879 2.704 0.006857 \*\*

TL -0.6573 0.1683 -3.907 9.35e-05 \*\*\*

WT -0.7896 0.3097 -2.549 0.010800 \*

HL 72.3327 20.7640 3.484 0.000495 \*\*\*

KL 27.3775 11.7780 2.324 0.020101 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 68.612 on 82 degrees of freedom

AIC: 78.612

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ TL + AE + WT + BH + HL + TT + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ TL + AE + WT + BH + HL + TT + SK + KL,

family = "binomial", data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2211 -0.5397 0.1404 0.5014 1.9806

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 28.03611 26.36607 1.063 0.28763

TL -0.73512 0.18643 -3.943 8.04e-05 \*\*\*

AE 0.07978 0.12459 0.640 0.52196

WT -0.88694 0.33935 -2.614 0.00896 \*\*

BH 0.55984 0.59108 0.947 0.34356

HL 52.95693 26.71797 1.982 0.04747 \*

TT 3.39882 11.72971 0.290 0.77200

SK 22.04688 27.18725 0.811 0.41741

KL 23.41190 12.05676 1.942 0.05216 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008 on 86 degrees of freedom

Residual deviance: 65.992 on 78 degrees of freedom

AIC: 83.992

Number of Fisher Scoring iterations: 6

> summary(glm(Status ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL, sparrows, family = "binomial"))

Call:

glm(formula = Status ~ AG + TL + AE + WT + BH + HL + FL + TT +

SK + KL, family = "binomial", data = sparrows)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2252 -0.5232 0.1397 0.5131 2.0134

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 27.56606 26.42412 1.043 0.296848

AGJuvenile 0.10631 0.68253 0.156 0.876225

TL -0.73634 0.18965 -3.883 0.000103 \*\*\*

AE 0.08275 0.12622 0.656 0.512060

WT -0.88860 0.34182 -2.600 0.009333 \*\*

BH 0.58293 0.59735 0.976 0.329131

HL 56.03494 31.05541 1.804 0.071176 .

FL -6.64680 31.73442 -0.209 0.834096

TT 5.05213 14.05263 0.360 0.719210

SK 21.53121 27.28482 0.789 0.430037

KL 23.56111 12.03826 1.957 0.050326 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.01 on 86 degrees of freedom

Residual deviance: 65.92 on 76 degrees of freedom

AIC: 87.92

Number of Fisher Scoring iterations: 6