Genetic Evaluation - Solution 2

Peter von Rohr

2021-04-26

Problem 1: Model Selection

We assume that we have a dataset for the response variable `carcass weight` (CW) and for some predictor variables

- sex (`sex`)
- slaughterhouse (`slh`)
- herd (`hrd`)
- age at slaughter (`age`)
- day of month when animal was slaughtered (`day`) and
- humidity (`hum`)

Use a fixed linear effects model and determine which of the predictor variables are important for the response.

The data is available from https://charlotte-ngs.github.io/gelasmss2021/data/gel_model_sel_ex02.csv.

Hint

- Use the function `lm` in R to fit the fixed linear effects model
- Use either Mallow C_p statistic or the adjusted coefficient of determination R^2_{adj} or AIC as model selection criteria
- Use the backward model selection approach

Solution

As preparatory step we have to first read the data from the file

```
s_data_file <- "https://charlotte-ngs.github.io/gelasmss2021/data/gel_model_sel_ex02.csv"
tbl_modsel <- readr::read_csv2(s_data_file)
```

Before we can do any model fits, we have to convert all fixed effects into `factors`. Fixed effects will be

- `sex`
- `slh`
- `hrd`
- `day`

These must be converted into factors. All other predictors are fit as covariables and can stay as numeric types.

```
tbl_modsel$sex <- as.factor(tbl_modsel$sex)
tbl_modsel$slh <- as.factor(tbl_modsel$slh)
tbl_modsel$hrd <- as.factor(tbl_modsel$hrd)
tbl_modsel$day <- as.factor(tbl_modsel$day)
```

The backward model selection approach starts with the full model.

```r
lm_full <- lm(cw ~ sex + slh + hrd + age + day + hum, data = tbl_modsel)
summary(lm_full)
```

```
##
## Call:
## lm(formula = cw ~ sex + slh + hrd + age + day + hum, data = tbl_modsel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.9503  -5.0785  -0.0034   4.9371  25.3859
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.848384   7.424203   1.731   0.0836 .
## sex2        -74.326113   1.270106 -58.520   <2e-16 ***
## slh2         22.260154   0.251693  88.442   <2e-16 ***
## slh3          3.633450   0.253731  14.320   <2e-16 ***
## hrd2         88.051103   0.324615 271.248   <2e-16 ***
## hrd3          8.715901   0.325158  26.805   <2e-16 ***
## hrd4         58.733786   0.322198 182.291   <2e-16 ***
## hrd5         19.830919   0.321711  61.642   <2e-16 ***
## age           0.646483   0.018124  35.669   <2e-16 ***
## day2         -0.823091   0.799581  -1.029   0.3033
## day3         -0.502529   0.780698  -0.644   0.5198
## day4         -1.144556   0.780938  -1.466   0.1428
## day5         -1.061056   0.808272  -1.313   0.1893
## day6         -1.380825   0.777552  -1.776   0.0758 .
## day7         -1.037485   0.752821  -1.378   0.1682
## day8         -1.773093   0.793269  -2.235   0.0254 *
## day9         -1.572124   0.782887  -2.008   0.0447 *
## day10        -0.548560   0.794306  -0.691   0.4898
## day11        -0.920831   0.760181  -1.211   0.2258
## day12        -1.212207   0.768703  -1.577   0.1149
## day13        -0.578945   0.813871  -0.711   0.4769
## day14        -0.230919   0.783872  -0.295   0.7683
## day15        -0.674826   0.795888  -0.848   0.3965
## day16        -1.081408   0.794644  -1.361   0.1736
## day17        -0.721491   0.794795  -0.908   0.3640
## day18        -0.100078   0.801605  -0.125   0.9006
## day19        -1.728759   0.783159  -2.207   0.0273 *
## day20        -1.031175   0.792600  -1.301   0.1933
## day21        -0.058945   0.804225  -0.073   0.9416
## day22        -0.184605   0.826888  -0.223   0.8233
## day23        -0.006881   0.797887  -0.009   0.9931
## day24        -1.872135   0.790999  -2.367   0.0180 *
## day25        -1.515168   0.776605  -1.951   0.0511 .
## day26        -1.403853   0.771310  -1.820   0.0688 .
## day27        -1.280929   0.796001  -1.609   0.1076
## day28        -1.278467   0.776949  -1.645   0.0999 .
## day29        -0.389556   0.820790  -0.475   0.6351
## day30        -1.127890   0.774005  -1.457   0.1451
## hum           0.127239   0.101636   1.252   0.2107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7.466 on 5286 degrees of freedom
## Multiple R-squared:  0.9571, Adjusted R-squared:  0.9568
## F-statistic:  3102 on 38 and 5286 DF,  p-value: < 2.2e-16
```

Using `stepAIC()` to do the backward selection results in

```
lm_back <- MASS::stepAIC(lm_full, direction = "backward")
```

```
## Start:  AIC=21448.59
## cw ~ sex + slh + hrd + age + day + hum
##
##        Df Sum of Sq     RSS   AIC
## - day  29      1554  296169 21419
## - hum   1        87  294703 21448
## <none>               294615 21449
## - age   1     70911  365526 22595
## - sex   1    190867  485482 24106
## - slh   2    508924  803540 26787
## - hrd   4   5795837 6090452 37569
##
## Step:  AIC=21418.61
## cw ~ sex + slh + hrd + age + hum
##
##        Df Sum of Sq     RSS   AIC
## - hum   1        86  296256 21418
## <none>               296169 21419
## - age   1     71363  367532 22566
## - sex   1    191473  487643 24072
## - slh   2    511678  807847 26758
## - hrd   4   5835440 6131609 37547
##
## Step:  AIC=21418.16
## cw ~ sex + slh + hrd + age
##
##        Df Sum of Sq     RSS   AIC
## <none>               296256 21418
## - age   1     71332  367588 22565
## - sex   1    191461  487716 24071
## - slh   2    511719  807974 26757
## - hrd   4   5835356 6131612 37545
```

```
summary(lm_back)
```

```
##
## Call:
## lm(formula = cw ~ sex + slh + hrd + age, data = tbl_modsel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.1701  -5.1196  -0.0517   4.9396  26.2927
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.69871    7.37800   1.586    0.113
## sex2        -74.26071    1.26695 -58.614   <2e-16 ***
```

```
## slh2         22.25705    0.25093  88.697   <2e-16 ***
## slh3          3.63425    0.25300  14.365   <2e-16 ***
## hrd2         88.00687    0.32358 271.978   <2e-16 ***
## hrd3          8.70555    0.32368  26.895   <2e-16 ***
## hrd4         58.70436    0.32126 182.732   <2e-16 ***
## hrd5         19.80659    0.32085  61.731   <2e-16 ***
## age           0.64693    0.01808  35.777   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.465 on 5316 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9568
## F-statistic: 1.473e+04 on 8 and 5316 DF,  p-value: < 2.2e-16
```

Comparing the above result from `MASS::stepAIC()` to the real model that was used in the simulation shows that they agree.

```
lm_relevant <- lm(cw ~ sex + slh + hrd + age, data = tbl_modsel)
summary(lm_relevant)
```

```
##
## Call:
## lm(formula = cw ~ sex + slh + hrd + age, data = tbl_modsel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.1701  -5.1196  -0.0517   4.9396  26.2927
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.69871    7.37800   1.586    0.113
## sex2        -74.26071    1.26695 -58.614   <2e-16 ***
## slh2         22.25705    0.25093  88.697   <2e-16 ***
## slh3          3.63425    0.25300  14.365   <2e-16 ***
## hrd2         88.00687    0.32358 271.978   <2e-16 ***
## hrd3          8.70555    0.32368  26.895   <2e-16 ***
## hrd4         58.70436    0.32126 182.732   <2e-16 ***
## hrd5         19.80659    0.32085  61.731   <2e-16 ***
## age           0.64693    0.01808  35.777   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.465 on 5316 degrees of freedom
## Multiple R-squared:  0.9568, Adjusted R-squared:  0.9568
## F-statistic: 1.473e+04 on 8 and 5316 DF,  p-value: < 2.2e-16
```