Logistic Regression Applications Solutions

Lt Col Ken Horton   Lt Col Kris Pruitt   Professor Bradley Warner

18 November, 2020

Exercises

1. Possum classification

Let's investigate the `possum` data set again. This time we want to model a binary outcome variable. As a reminder, the common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum. We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia.

We use logistic regression to differentiate between possums in these two regions. The outcome variable, called `pop`, takes value `Vic` when a possum is from Victoria and `other` when it is from New South Wales or Queensland. We consider five predictors: `sex`, `head_l`, `skull_w`, `total_l`, and `tail_l`.

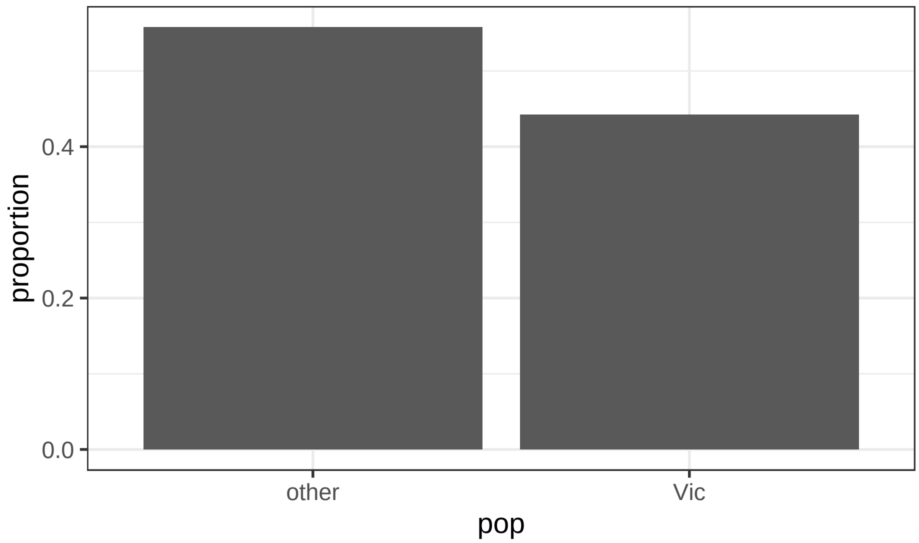
a. Explore the data by making histograms of the quantitative variables, and bar charts of the discrete variables. Are there any outliers that are likely to have a very large influence on the logistic regression model?

```
possum <- read_csv("data/possum.csv") %>%
  select(pop,sex,head_l,skull_w,total_l,tail_l) %>%
  mutate(pop=factor(pop),sex=factor(sex))
```

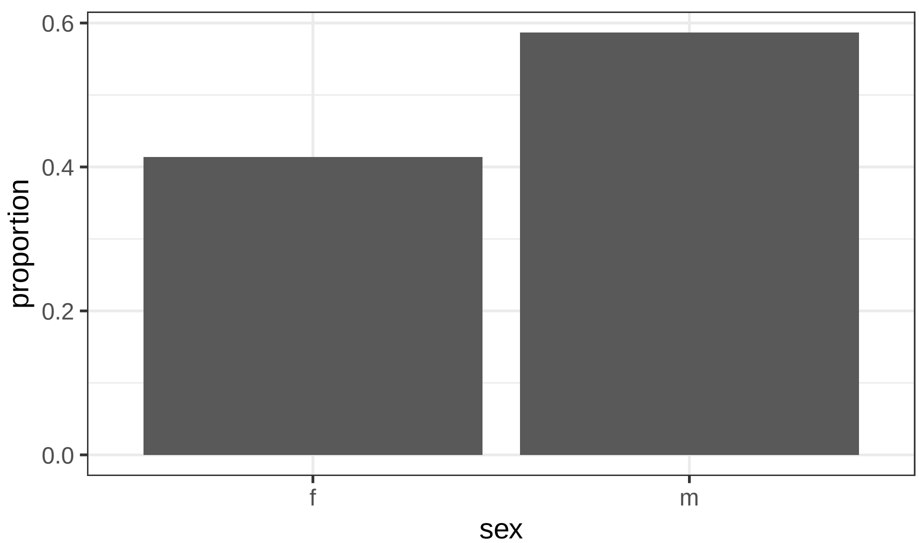
```
inspect(possum)
```

```
##
## categorical variables:
##   name  class levels   n missing                                  distribution
## 1  pop factor      2 104       0 other (55.8%), Vic (44.2%)
## 2  sex factor      2 104       0 m (58.7%), f (41.3%)
##
## quantitative variables:
##         name   class  min     Q1 median     Q3   max     mean       sd   n
## ...1  head_l numeric 82.5 90.675  92.80 94.725 103.1 92.60288 3.573349 104
## ...2 skull_w numeric 50.0 54.975  56.35 58.100  68.6 56.88365 3.113426 104
## ...3 total_l numeric 75.0 84.000  88.00 90.000  96.5 87.08846 4.310549 104
## ...4  tail_l numeric 32.0 35.875  37.00 38.000  43.0 37.00962 1.959518 104
##      missing
## ...1       0
## ...2       0
## ...3       0
## ...4       0
```

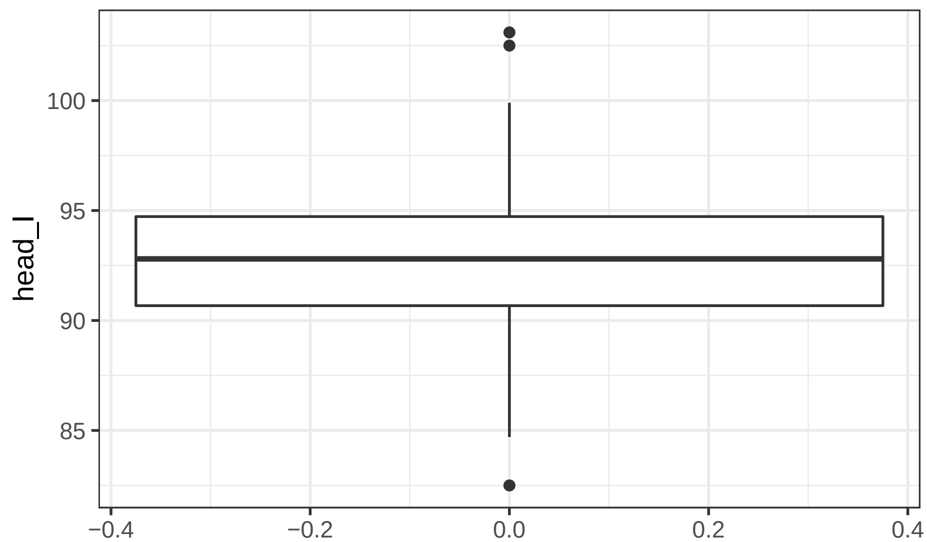
```r
possum %>%
  gf_props(~pop) %>%
  gf_theme(theme_bw())
```



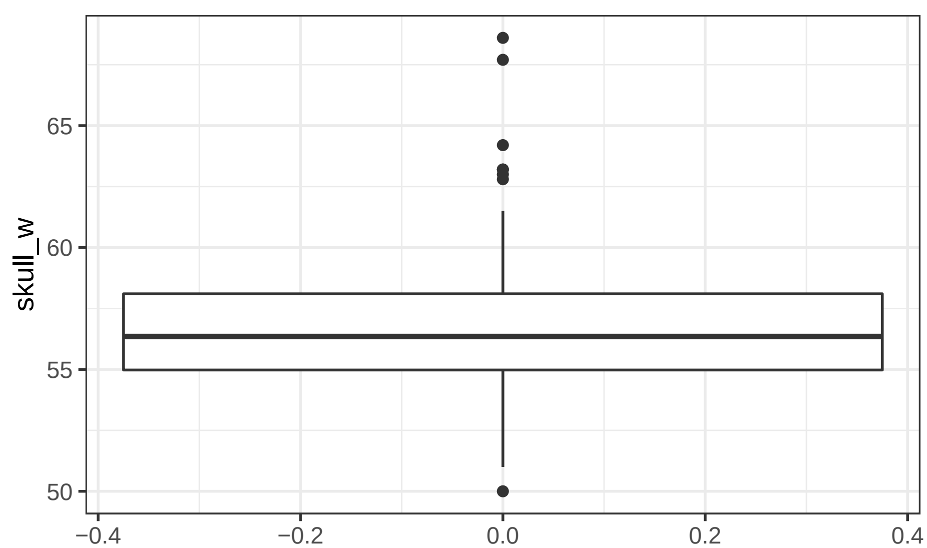

```r
possum %>%
  gf_props(~sex) %>%
  gf_theme(theme_bw())
```



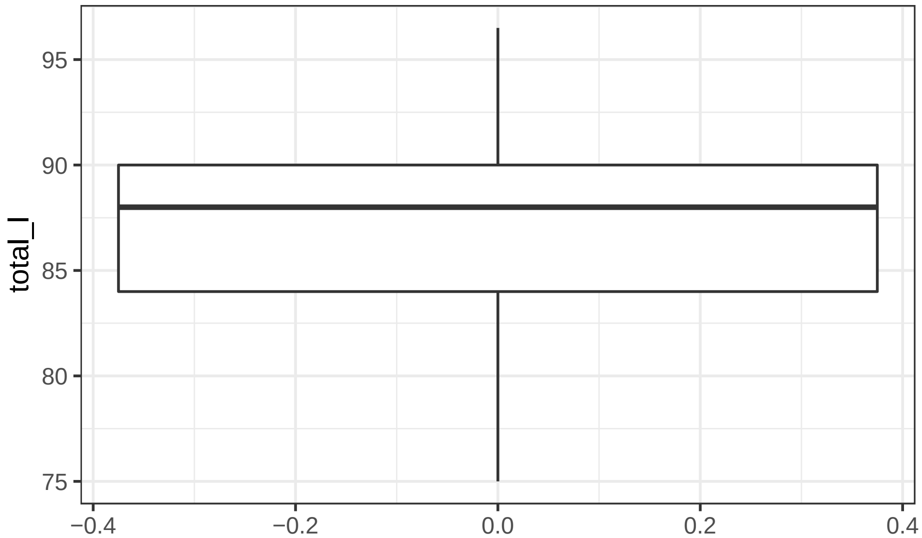

```r
possum %>%
  gf_boxplot(~head_l) %>%
  gf_theme(theme_bw())
```



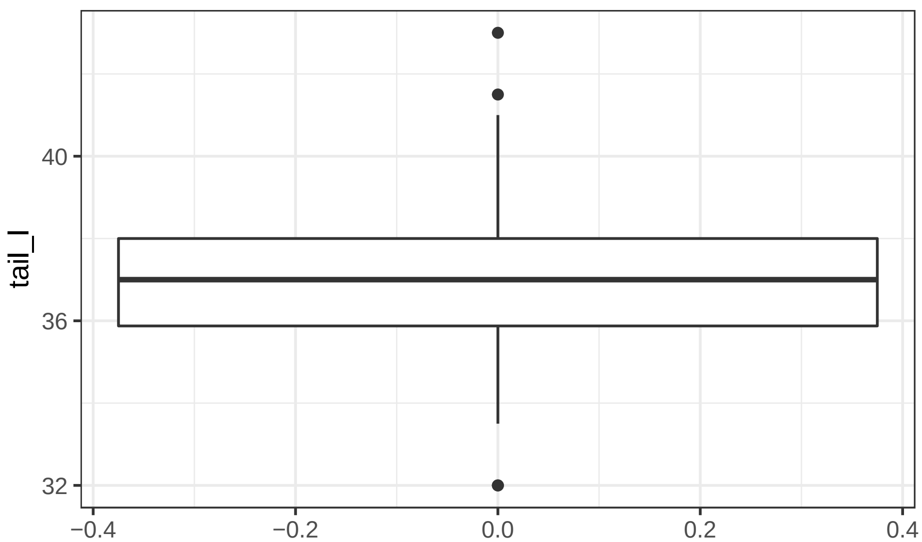

```
possum %>%
  gf_boxplot(~skull_w) %>%
  gf_theme(theme_bw())
```




```
possum %>%
  gf_boxplot(~total_l) %>%
  gf_theme(theme_bw())
```

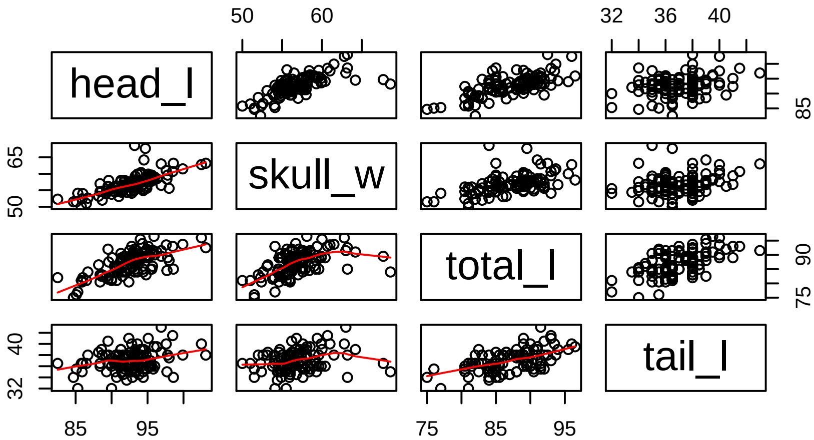



```
possum %>%
  gf_boxplot(~tail_l) %>%
  gf_theme(theme_bw())
```



There are some potential outliers for skull width but otherwise not much concern.

```
pairs(possum[,3:6],lower.panel = panel.smooth)
```



We can see that `head_l` is correlated with the other three variables. This will cause some multicollinearity problems.

b. Build a logistic regression model with all the variable. Report a summary of the model.

```r
possum_mod <- glm(pop=="Vic"~.,data=possum,family="binomial")
```

```r
summary(possum_mod)
```

```
##
## Call:
## glm(formula = pop == "Vic" ~ ., family = "binomial", data = possum)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6430  -0.5514  -0.1182   0.3760   2.8501
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  39.2349    11.5368   3.401 0.000672 ***
## sexm         -1.2376     0.6662  -1.858 0.063195 .
## head_l       -0.1601     0.1386  -1.155 0.248002
## skull_w      -0.2012     0.1327  -1.517 0.129380
## total_l       0.6488     0.1531   4.236 2.27e-05 ***
## tail_l       -1.8708     0.3741  -5.001 5.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142.787  on 103  degrees of freedom
## Residual deviance:  72.155  on  98  degrees of freedom
## AIC: 84.155
##
## Number of Fisher Scoring iterations: 6
```

```r
confint(possum_mod)
```

```
## Waiting for profiling to be done...

##                  2.5 %      97.5 %
## (Intercept) 18.8530781 64.66444839
## sexm        -2.6227018  0.02472167
## head_l      -0.4428559  0.10865739
## skull_w     -0.4933140  0.04479826
## total_l      0.3768179  0.98455786
## tail_l      -2.7170468 -1.23231969
```

c. Using the p-values decide if you want to remove a variable(S) and if so build that model.

Let's remove `head_l` first.

```r
possum_mod_red <- glm(pop=="Vic"~sex+skull_w+total_l+tail_l,data=possum,family="binomial")
```

```r
summary(possum_mod_red)
```

```
##
## Call:
## glm(formula = pop == "Vic" ~ sex + skull_w + total_l + tail_l,
##     family = "binomial", data = possum)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8102  -0.5683  -0.1222   0.4153   2.7599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  33.5095     9.9053   3.383 0.000717 ***
## sexm         -1.4207     0.6457  -2.200 0.027790 *
## skull_w      -0.2787     0.1226  -2.273 0.023053 *
## total_l       0.5687     0.1322   4.302 1.69e-05 ***
## tail_l       -1.8057     0.3599  -5.016 5.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142.787  on 103  degrees of freedom
## Residual deviance:  73.516  on  99  degrees of freedom
## AIC: 83.516
##
## Number of Fisher Scoring iterations: 6
```

Since `head_l` was correlated with the other variables, removing it has increased the precision, decreased the standard error, of the other predictors. There p-values are all now less than 0.05.

d. For any variable you decide to remove, build a 95% confidence interval for the parameter.

```r
confint(possum_mod)
```

```
## Waiting for profiling to be done...

##                  2.5 %      97.5 %
## (Intercept) 18.8530781 64.66444839
## sexm        -2.6227018  0.02472167
## head_l      -0.4428559  0.10865739
## skull_w     -0.4933140  0.04479826
## total_l      0.3768179  0.98455786
## tail_l      -2.7170468 -1.23231969
```

We are 95% confident that the true slope coefficient for `head_l` is between -0.44 and 0.108.

The bootstrap is not working for this problem. It may be that we have convergence issues when we resample the data. This is a reminder that we need to be careful and not just run methods without checking results. Here is the code:

```r
set.seed(952)
results<-do(1000)*glm(pop=="Vic"~.,data=resample(possum),family="binomial")
```

```r
head(results)
```

```
##     Intercept          sexm        head_l       skull_w       total_l
## 1 -1184.61875  2.122389e+01  3.861561e+00  2.749263e+00  7.274005e+00
## 2  6371.55550 -1.301514e+02  1.023732e+01 -2.738816e+01 -1.076913e+01
## 3 -9612.61941 -2.392900e+03 -1.875252e+02  5.782027e+02 -2.820691e+02
## 4   -25.18662 -1.852185e+01  2.097593e+01 -1.353619e+01  1.483815e+01
## 5   -26.56607 -1.398995e-14 -2.258909e-14  6.528545e-15  2.388756e-14
## 6 -1025.00035  6.159665e+01  2.526181e+01 -2.032143e+01  1.438639e+01
##          tail_l    orig.id102    orig.id103  orig.id104   orig.id12  orig.id15
## 1  3.651693e-01 -1.122791e+01  1.830254e+01   -30.06351    9.691673   38.52117
## 2 -1.268187e+02 -6.013872e+01 -1.755235e+02   315.41584          NA         NA
## 3  5.516366e+02            NA  3.566885e+02 -4457.23937 1654.519130 1509.96846
## 4 -6.444674e+01            NA -2.679856e+01   103.36587 -263.501068 -176.17714
## 5 -3.702267e-14  1.750610e-13  1.941549e-14          NA          NA         NA
## 6 -4.000390e+01            NA  2.039059e+01          NA          NA         NA
##    orig.id16    orig.id17  orig.id19    orig.id2 orig.id20 orig.id22  orig.id23
## 1   44.12043    -4.982253   23.26590    16.73273 29.062437 -19.09794   39.22422
## 2 -317.45001   132.433698 -380.82833 -100.12632  4.028447 390.84406 -231.54072
## 3         NA -6636.047683         NA -644.61784        NA        NA -206.66839
## 4         NA           NA -265.68649 -114.22065        NA  37.53452 -129.65294
## 5   53.13214    53.132137   53.13214    53.13214        NA  53.13214   53.13214
## 6         NA   207.594757 -109.17251    29.15153 40.603330  17.05419         NA
##    orig.id25  orig.id26  orig.id27  orig.id28  orig.id30   orig.id31
## 1  -19.34349  -10.76200   80.80735   26.15158   82.72554   -7.120802
## 2 -105.92746         NA -424.78233 -307.29246 -644.89802   10.658336
## 3 2398.36905 1447.85074         NA 2288.58868         NA 3975.534037
## 4 -217.18353 -160.13003 -114.45314 -195.91828 -231.20021 -139.402101
## 5   53.13214   53.13213   53.13213         NA   53.13213   53.132133
## 6 -137.52370  -90.83382   50.18438         NA  -37.88613 -112.726291
##     orig.id32 orig.id34  orig.id35  orig.id37  orig.id38 orig.id39   orig.id40
```

```
## 1   -6.841683  51.71343   34.99329  102.80146   80.67304  184.5574    80.77377
## 2  200.766711        NA -178.02745 -431.20393 -258.05139        NA          NA
## 3 -460.608750 798.15409  649.30597         NA  103.52953        NA -1121.24437
## 4          NA        NA  -99.49100  -48.12580  -16.48625   50.5334   -46.30721
## 5   53.132137  53.13214         NA   53.13213   53.13213        NA    53.13214
## 6          NA        NA  -20.83852  122.56094   55.81312        NA          NA
##    orig.id42  orig.id43 orig.id45  orig.id46  orig.id49    orig.id5 orig.id51
## 1  140.05083  110.18040  84.42337   63.65698  -26.12220    67.99495 -6.405154
## 2 -718.96127 -815.80810        NA -317.35251  -15.56427          NA        NA
## 3         NA -378.56648 811.04644 2104.18891 -795.41935 -1497.07611        NA
## 4  -66.90598         NA -86.26996 -101.32511  -59.33191   -54.03626        NA
## 5   53.13214   53.13214        NA         NA         NA    53.13213        NA
## 6  106.89994   20.67055        NA  -21.79366         NA    94.31185        NA
##   orig.id52 orig.id55 orig.id56     orig.id57   orig.id6     orig.id61
## 1 -102.4909 -121.7735 -113.4715 -6.394241e+01   29.75290 -3.436999e+01
## 2        NA  222.7481        NA  4.562610e+02         NA            NA
## 3        NA -100.4041  620.9428 -3.310210e+03 1356.43221 -8.930887e+02
## 4        NA        NA        NA            NA         NA -1.872334e+01
## 5        NA        NA        NA  3.569360e-15         NA  4.053929e-14
## 6        NA -193.9330 -164.9053  1.053552e+02  -68.52307  1.165930e+00
##       orig.id62    orig.id63     orig.id64     orig.id65     orig.id66
## 1 -2.650416e+01    -9.895249 -3.654841e+01 -1.843685e+01  1.666557e+01
## 2            NA   -23.084937  3.669481e+00 -1.105276e+01            NA
## 3 -2.815222e+03 -1771.331692 -9.040277e+02            NA -1.602168e+03
## 4  3.918779e+01           NA            NA            NA            NA
## 5 -1.189173e-13           NA  3.537457e-14  1.071007e-14  7.030941e-15
## 6  1.288583e+02    67.429596            NA  4.985831e+01  7.161574e+01
##    orig.id67     orig.id68     orig.id69    orig.id7  orig.id72     orig.id74
## 1  -13.43152 -3.812871e+01 -4.826951e+01   -2.222496   62.72098  5.495169e+01
## 2 -101.29629  7.461278e+01 -4.506194e+01  -67.154169 -406.64765 -1.690399e+02
## 3   11.72032  1.246652e+03  2.344065e+03 1638.110671         NA -1.397453e+03
## 4         NA            NA -1.748907e+02 -148.856764  -14.22850  5.934741e+01
## 5         NA -5.182242e-15  4.526154e-14   53.132137         NA  2.033970e-14
## 6  -12.27520            NA -1.748038e+02  -82.214506         NA  1.405044e+02
##       orig.id75     orig.id76     orig.id77    orig.id78     orig.id79
## 1 -4.638070e+01 -3.849598e+01 -3.284425e+00 6.006826e+01  6.552061e+01
## 2  1.405837e+02            NA            NA           NA -2.757429e+02
## 3            NA            NA            NA           NA            NA
## 4 -3.556608e+01  1.695434e+02            NA           NA            NA
## 5 -1.659973e-14  4.965611e-14  5.295878e-14 1.098947e-14  2.455674e-14
## 6 -1.228508e+01  1.166753e+02 -5.391930e+01 8.292921e+01            NA
##     orig.id8     orig.id82    orig.id83     orig.id84     orig.id85 orig.id88
## 1   11.30555  6.770113e+01 5.662433e+01  4.665353e+01 -4.728733e+01   2.24768
## 2  -65.64737            NA           NA            NA            NA        NA
## 3         NA -2.514791e+03           NA -1.660379e+02            NA        NA
## 4 -122.82285            NA           NA            NA  1.455141e+02  54.91457
## 5         NA -3.283904e-14 4.400705e-15  1.525840e-14 -1.296601e-13        NA
## 6         NA            NA           NA  9.528016e+01            NA 174.95861
##       orig.id89     orig.id90     orig.id91 orig.id92 orig.id94 orig.id96
## 1 -2.843704e+01 -3.210972e+00 -6.595084e+01        NA        NA        NA
## 2 -3.200426e+02            NA  1.407166e+02 -46.87452        NA        NA
## 3  1.060689e+03            NA            NA        NA        NA        NA
## 4 -2.398799e+02            NA -9.916551e+01        NA        NA        NA
## 5  5.515735e-14  1.140355e-14  3.439707e-14        NA        NA        NA
```

```
## 6 -1.599827e+02  2.445159e+01 -8.285017e+01  46.98960        NA        NA
##   orig.id97 orig.id99 .row orig.id10 orig.id101  orig.id13  orig.id14
## 1        NA        NA    1        NA         NA         NA         NA
## 2        NA        NA    1   23.2755   157.5174  -18.54683  -63.07253
## 3        NA        NA    1        NA -1277.1247         NA         NA
## 4        NA        NA    1        NA   188.3804 -121.65005 -188.75237
## 5        NA        NA    1        NA         NA         NA   53.13213
## 6        NA        NA    1  123.7400         NA  -42.61572 -125.70632
##    orig.id18  orig.id21 orig.id24  orig.id29   orig.id33  orig.id36  orig.id44
## 1         NA         NA        NA         NA          NA         NA         NA
## 2 -111.81282 313.062580 -461.1891 -374.93855   -4.391855 -159.52493 -355.02674
## 3 2887.10658         NA 2929.2997         NA 2295.620871         NA         NA
## 4         NA -59.621238 -221.9315 -186.90793  -43.915123         NA         NA
## 5         NA  53.132137        NA   53.13214   53.132138   53.13214   53.13214
## 6  -94.74002   1.501875 -127.4223         NA          NA  -27.76164         NA
##   orig.id47     orig.id48     orig.id50   orig.id53    orig.id54     orig.id59
## 1        NA            NA            NA          NA           NA            NA
## 2  24.67150 -3.192271e+02 -2.437457e+01    39.13670 9.040662e+02  5.092646e+02
## 3        NA -8.532396e+02 -1.987376e+03 -1842.82567           NA -9.766937e+01
## 4 -34.84794 -2.656503e+02            NA   -69.08107 2.508742e+02 -1.294101e+02
## 5        NA  1.434615e-13 -4.787834e-15          NA 4.342819e-14 -2.136850e-13
## 6        NA -1.323725e+02  8.472751e+01    38.22207           NA -1.572490e+02
##       orig.id70  orig.id71 orig.id73     orig.id86 orig.id87   orig.id9
## 1            NA         NA        NA            NA        NA         NA
## 2 -3.821490e+01  401.53342 -168.1978 -2.995413e+01 650.61940  -81.53516
## 3 -2.213188e+03 -602.42467  204.7150 -1.473635e+03        NA         NA
## 4 -7.568901e+00   92.69708  130.1489            NA  67.25214 -118.47267
## 5 -7.027829e-17         NA        NA -3.931598e-14        NA         NA
## 6            NA   68.43743  196.0079            NA        NA         NA
##   orig.id93 orig.id98 orig.id100    orig.id3  orig.id4   orig.id41
## 1        NA        NA         NA          NA        NA          NA
## 2        NA        NA         NA          NA        NA          NA
## 3        NA        NA  -762.9931 -2001.83141 -910.6692 -3464.13722
## 4        NA        NA    40.8204   -11.43343        NA    89.78318
## 5        NA        NA         NA          NA        NA    53.13214
## 6        NA        NA         NA          NA        NA          NA
##       orig.id60     orig.id81 orig.id11     orig.id58     orig.id80 orig.id95
## 1            NA            NA        NA            NA            NA        NA
## 2            NA            NA        NA            NA            NA        NA
## 3 -3.346934e+03 -2.558937e+02        NA            NA            NA        NA
## 4  7.222230e+01  5.506440e+01  54.37726            NA            NA        NA
## 5 -3.181852e-14  1.402462e-14  53.13213 -3.207989e-14 -1.600303e-14        NA
## 6  1.587278e+02  7.623006e+01 129.59604  1.052235e+02  1.272378e+02        NA
##   .index
## 1      1
## 2      2
## 3      3
## 4      4
## 5      5
## 6      6
```

```r
confint(results)
```

```
##        name       lower      upper level     method   estimate
```

```
## 1 Intercept -8030.43219 8566.11832  0.95 percentile 39.2349178
## 2      sexm  -201.28404  207.55196  0.95 percentile -1.2375895
## 3    head_l  -122.70294   63.61867  0.95 percentile -0.1600622
## 4   skull_w   -35.94883   92.78429  0.95 percentile -0.2012445
## 5   total_l   -32.84729   87.94284  0.95 percentile  0.6488131
## 6    tail_l  -138.14608  151.97362  0.95 percentile -1.8708001
```

These intervals are much too large.

e. Explain why the remaining parameter estimates change between the two models.

When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for sex male changed when we removed the head length variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.

f. Write out the form of the model. Also identify which of the following variables are positively associated (when controlling for other variables) with a possum being from Victoria: `head_l`, `skull_w`, `total_l`, and `tail_l`.

We dropped `head_l` from the model. Here is the equation:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = 33.5 - 1.42 \text{ sex} - 0.28 \text{ skull width} + 0.57 \text{ total length} - 1.81 \text{ tail length}$$

Only `total_l` is positively association with the probability of being from Victoria.

g. Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the reduced model's computed probability that this possum is from Victoria? How confident are you in the model's accuracy of this probability calculation?

Let's predict the outcome. We use `response` for the type to put the answer in the form of a probability. See the help menu on `predict.glm` for more information.

```r
predict(possum_mod_red,newdata = data.frame(sex="m",skull_w=63,tail_l=37,total_l=83),
        type="response",se.fit = TRUE)
```

```
## $fit
##           1
## 0.006205055
##
## $se.fit
##           1
## 0.008011468
##
## $residual.scale
## [1] 1
```

While the probability, 0.006, is very near zero, we have not run diagnostics on the model. We should also have a little skepticism that the model will hold for a possum found in a US zoo. However, it is encouraging that the possum was caught in the wild.

As a rough sense of the accuracy, we will use the standard error. The errors are really binomial but we are trying to use a normal approximation. If you remember back to our block on probability, with such a low probability, this assumption of normality is suspect. However, we will use it to give us an upper bound.

```
0.0062+c(-1,1)*1.96*.008
```

```
## [1] -0.00948  0.02188
```

So at most, the probability of the possum being from Victoria is 2%.

File Creation Information

- File creation date: 2020-11-18
- Windows version: Windows 10 x64 (build 18362)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.7.0
- `tidyverse` package version: 1.3.0
- `openintro` package version: 2.0.0