432 Assignment 4 Answer Sketch

432 Staff

Due 2018-02-23 at 1 PM. Version: 2018-02-22

Contents

0.1 Setup

```
knitr::opts_chunk$set(comment=NA)

library(rms)
library(skimr)
library(broom)
library(tidyverse)

skim_with(numeric = list(hist = NULL),
          integer = list(hist = NULL))

bird <- read.csv("bird.csv") %>% tbl_df
```

1 Question 1 (60 points)

You will fit a logistic regression model to address the key research question here, which is: "After age, sex, socio-economic status and smoking have been controlled for, is there an additional risk associated with keeping a bird as a pet?"

You will need to:

1. specify an appropriate model for the data, then
2. evaluate the quality of that model, appropriately,
3. and then provide an estimate (odds ratio with associated 95% confidence interval and careful interpretation) that addresses the research question above directly, then state your conclusion about whether this additional risk exists, and if so, how large is it?

Some specific suggestions:

- Use complete English sentences, punctuated by (well-edited) critical statistical output. Include the code used to produce that output in your HTML file.
- Focus your presentation on the things that are *most important* for your reader to see.

- Feel free to fit the simplest possible model that meets the requirements of the question.
- Your model will need to include all of the effects that are supposed to be accounted for, but you need not fit complex interaction or other non-linear terms on the right-hand side of the model unless you choose to do so.
- You will have multiple decisions to make about how best to fit and analyze your model. Describe those choices well in your response.

1.1 Tidying data

Before doing any analysis, let's make sure the data is amenable to being analyzed:

```
skim(bird)
```

```
Skim summary statistics
 n obs: 147
 n variables: 8

Variable type: factor
 variable missing complete   n n_unique               top_counts ordered
      ses       0      147 147        2 Low: 102, Hig: 45, NA: 0   FALSE
      sex       0      147 147        2     M: 111, F: 36, NA: 0   FALSE

Variable type: integer
 variable missing complete   n    mean    sd   p0    p25 median    p75
      age       0      147 147   56.97  7.35   37   52       59   63
     cigs       0      147 147   15.75  9.7     0   10       15   20
    lungc       0      147 147    0.33  0.47    0    0        0    1
  petbird       0      147 147    0.46  0.5     0    0        0    1
  smokeyr       0      147 147   27.85 13.98    0   20       30   39
  subject       0      147 147 1074    42.58 1001 1037.5   1074 1110.5
 p100
   67
   45
    1
    1
   50
 1147
```

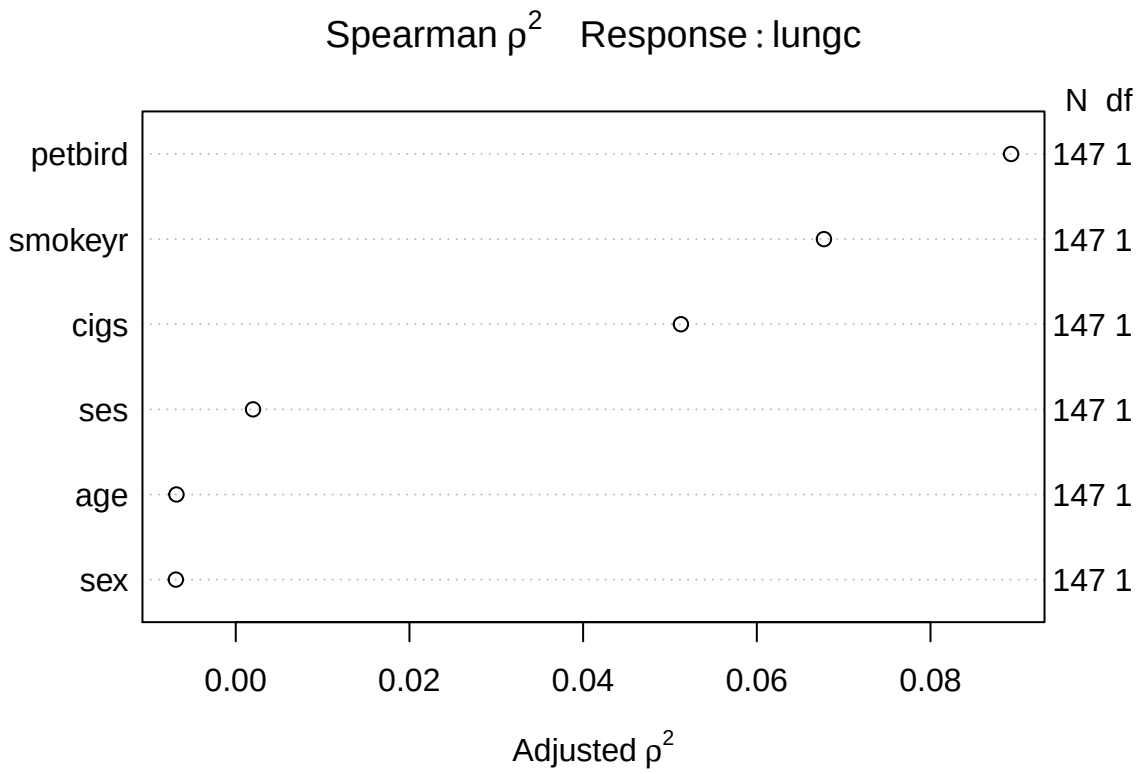
There is no missingness. None of the values seem out of line.

1.2 Specifying an appropriate model

The simplest possible model that meets the requirements of the problem would include only the main effects of our four controlling variables (age, sex, socio-economic status and smoking) as adjustors for the key predictor (`petbird`)'s effect on lung cancer. It's your decision how best to incorporate the two variables related to smoking.

Before specifying our model, we might look at the Spearman ρ^2 plot...

```
plot(spearman2(lungc ~ petbird + sex + ses +
                   cigs + smokeyr + age, data = bird))
```



Looking at the Spearman ρ^2 plot, it may be worth using `smokeyr` rather than `cigs`, so we'll try that first.

1.3 Building Model 1 (using glm)

```
model.1 <- glm(lungc ~ petbird + sex + ses + age + smokeyr,
               data = bird, family = binomial(logit))
summary(model.1)
```

```
Call:
glm(formula = lungc ~ petbird + sex + ses + age + smokeyr, family = binomial(logit),
    data = bird)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5193  -0.8722  -0.4522   0.9938   2.2207

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.75275    1.74785  -0.431 0.666706
petbird      1.33487    0.40913   3.263 0.001104 **
sexM        -0.52134    0.52960  -0.984 0.324914
sesLow      -0.13209    0.46417  -0.285 0.775976
age         -0.04634    0.03494  -1.326 0.184702
smokeyr      0.08288    0.02486   3.334 0.000857 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 155.24  on 141  degrees of freedom
AIC: 167.24

Number of Fisher Scoring iterations: 5
```

And now, here are the estimated odds ratios and confidence intervals.

```
exp(coef(model.1))
```

```
(Intercept)     petbird        sexM      sesLow         age     smokeyr
  0.4710694   3.7995037   0.5937247   0.8762651   0.9547130   1.0864141
```

```
exp(confint(model.1))
```

```
Waiting for profiling to be done...
                 2.5 %    97.5 %
(Intercept) 0.01404805 14.012382
petbird     1.73170213  8.676281
sexM        0.20749227  1.680832
sesLow      0.35178629  2.201361
age         0.88832369  1.020447
smokeyr     1.04001440  1.148195
```

1.3.1 Tidying the Model 1 Results

A better way to store these may be:

```
m1_coeffs <- tidy(model.1, exponentiate = TRUE, conf.int = TRUE)

m1_coeffs
```

```
         term  estimate  std.error  statistic      p.value   conf.low
1 (Intercept) 0.4710694 1.74784635 -0.4306728 0.6667062761 0.01404805
2     petbird 3.7995037 0.40913318  3.2626795 0.0011036424 1.73170213
3        sexM 0.5937247 0.52959566 -0.9844105 0.3249137213 0.20749227
4      sesLow 0.8762651 0.46416679 -0.2845670 0.7759758708 0.35178629
5         age 0.9547130 0.03493961 -1.3264160 0.1847019567 0.88832369
6     smokeyr 1.0864141 0.02486266  3.3336129 0.0008572586 1.04001440
  conf.high
1 14.012382
2  8.676281
3  1.680832
4  2.201361
5  1.020447
6  1.148195
```

1.3.2 Describing the Odds Ratio for `petbird`

The odds ratio for lung cancer associated with `petbird` after adjustment for `age`, `sex`, `ses` (socio-economic status) and `smokeyr` (years of smoking) is 3.8 with 95% CI for that odds ratio of (1.73, 8.68).

This means that the odds of lung cancer for someone who keeps a pet bird were 3.8 times higher than the odds of lung cancer for someone who did not keep a bird, assuming that these two people had the same values of `sex`, `ses`, `age` and `smokeyr`.

1.3.3 Fitting Model 1 with `lrm` to estimate C, R^2

We could also have fit model.1 with the `lrm` function from the `rms` package.

```
d <- datadist(bird)
options(datadist="d")

mod1_lrm <- lrm(lungc ~ petbird + sex + ses + age + smokeyr,
             data=bird, x = TRUE, y = TRUE)
mod1_lrm
```

```
Logistic Regression Model

 lrm(formula = lungc ~ petbird + sex + ses + age + smokeyr, data = bird,
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs            147    LR chi2      31.89    R2       0.271    C       0.771
  0              98    d.f.             5    g        1.435    Dxy     0.542
  1              49    Pr(> chi2) <0.0001    gr       4.200    gamma   0.543
 max |deriv| 1e-06                          gp       0.242    tau-a   0.243
                                            Brier    0.177

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -0.7527 1.7478 -0.43  0.6667
 petbird    1.3349 0.4091  3.26  0.0011
 sex=M     -0.5213 0.5296 -0.98  0.3249
 ses=Low   -0.1321 0.4642 -0.28  0.7760
 age       -0.0463 0.0349 -1.33  0.1847
 smokeyr    0.0829 0.0249  3.33  0.0009
```

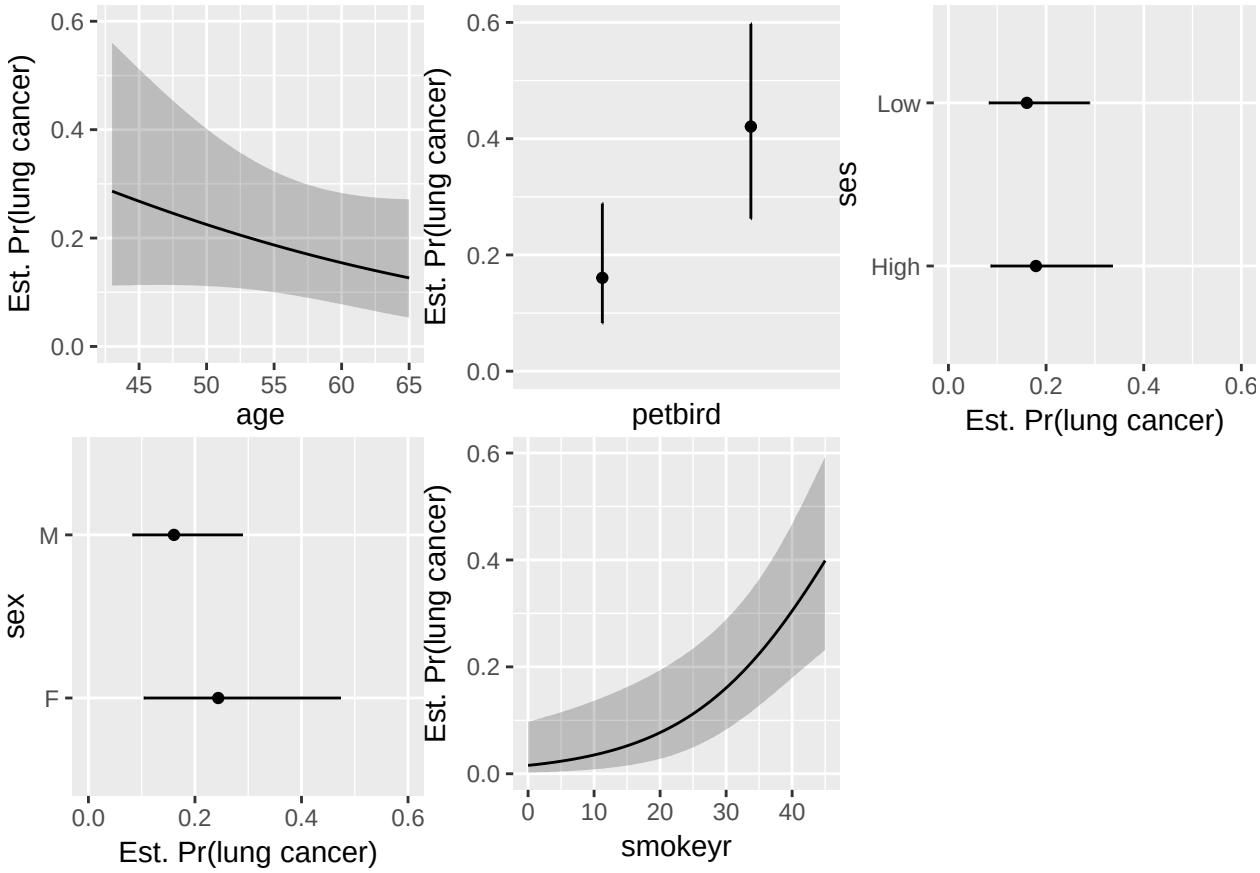
1.3.4 Evaluating Model 1

The C statistic and R^2 values are almost good.

1.3.5 Visualizing Model 1's Predictions

Note that, by specifying `plogis`, we are telling R to hold several variables at their median values and give us the *probabilities* of our outcome rather than the coefficients or odds ratios.

```
ggplot(Predict(mod1_lrm, fun = plogis),
       ylab = "Est. Pr(lung cancer)")
```



1.4 An Alternative Model (Model 2)

An alternative model using `cigs` rather than `smokeyr` would be even more aggressive, yielding the following:

```
model.2 <- glm(lungc ~ petbird + sex + ses + age + cigs,
               data = bird, family = "binomial")

m2_coeffs <- tidy(model.2, exponentiate = T, conf.int = T)

m2_coeffs
```

```
         term   estimate  std.error  statistic      p.value   conf.low
1 (Intercept) 0.03965134 1.66147354 -1.9426314 0.0520607160 0.00129277
2     petbird 4.26406531 0.40045953  3.6213971 0.0002930163 1.98103984
3        sexM 0.89762064 0.48587267 -0.2222964 0.8240831427 0.34801203
4      sesLow 1.19940606 0.44747050  0.4063430 0.6844906274 0.50303456
5         age 1.01357598 0.02751408  0.4901003 0.6240629473 0.96120187
6        cigs 1.06070725 0.02174522  2.7102921 0.0067223967 1.01775334
  conf.high
1 0.9138759
2 9.5920233
3 2.3672757
4 2.9479734
5 1.0716149
6 1.1091053
```

1.4.1 Looking at the C statistic and R^2 for Model 2

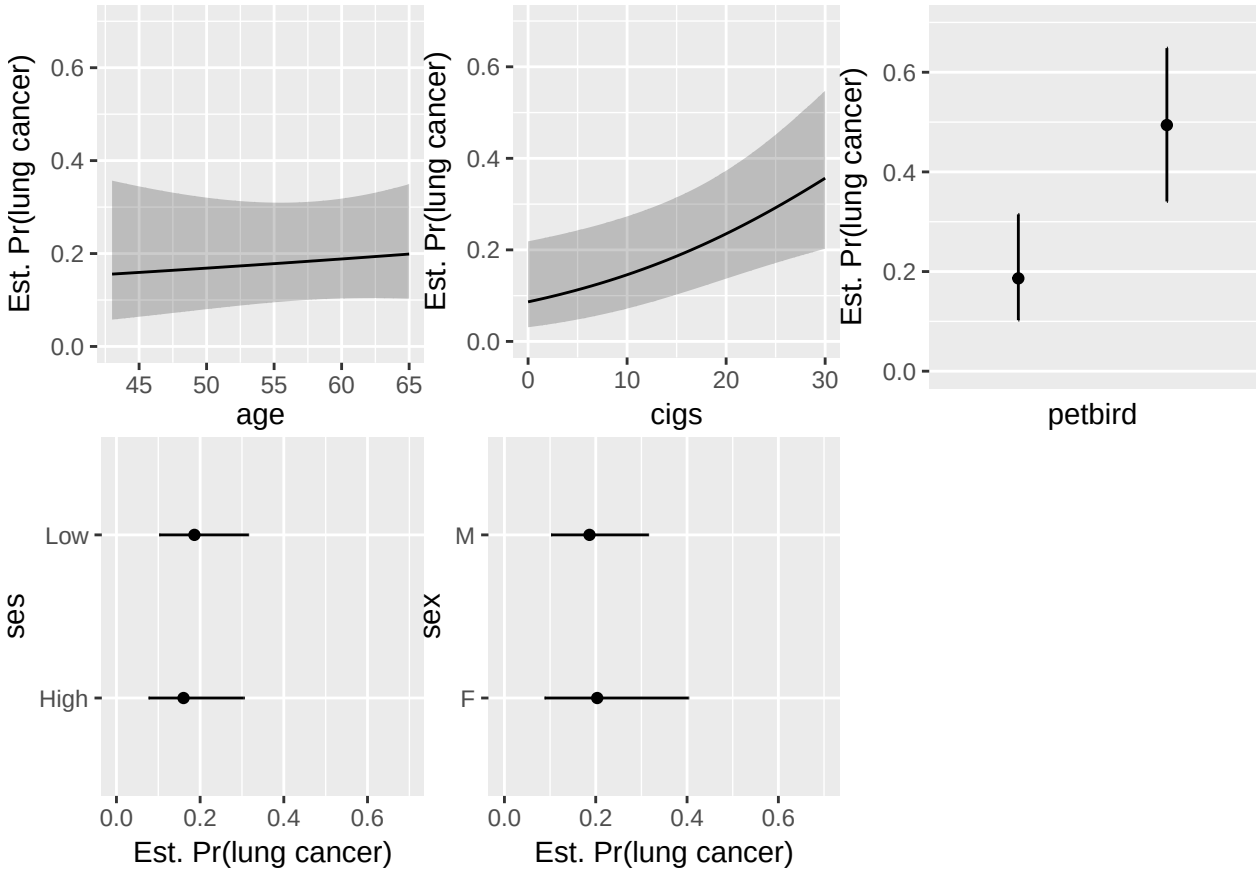
```
mod2_lrm <- lrm(lungc ~ petbird + sex + ses + age + cigs,
                data = bird, x = TRUE, y = TRUE)
mod2_lrm
```

```
Logistic Regression Model

 lrm(formula = lungc ~ petbird + sex + ses + age + cigs, data = bird,
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs            147    LR chi2     23.21    R2       0.203    C       0.741
  0              98    d.f.            5    g        1.074    Dxy     0.482
  1              49    Pr(> chi2) 0.0003    gr       2.927    gamma   0.483
 max |deriv| 2e-06                         gp       0.210    tau-a   0.216
                                           Brier    0.189

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -3.2276 1.6615 -1.94  0.0521
 petbird    1.4502 0.4005  3.62  0.0003
 sex=M     -0.1080 0.4859 -0.22  0.8241
 ses=Low    0.1818 0.4475  0.41  0.6845
 age        0.0135 0.0275  0.49  0.6241
 cigs       0.0589 0.0217  2.71  0.0067
```

1.4.2 Visualizing Model 2's Predictions

Here's the set of plots for Model 2.

```
ggplot(Predict(mod2_lrm, fun = plogis),
       ylab = "Est. Pr(lung cancer)")
```



1.4.3 Describing the Odds Ratio for `petbird` in Model 2

The odds ratio for lung cancer associated with `petbird` after adjustment for `age`, `sex`, `ses` (socio-economic status) and `cigs` (cigarettes smoked per day) is 4.26 with 95% CI for that odds ratio of (1.98, 9.59).

This means that the odds of lung cancer for someone who keeps a pet bird were 4.26 times higher than the odds of lung cancer for someone who did not keep a bird, assuming that these two people had the same values of `sex`, `ses`, `age` and `cigs`.

1.4.4 Evaluating Model 2

The C statistic and R^2 statistic adjusting for `cigs` turn out to be a bit worse than the model with `smokeyr`.

1.5 Yet another approach (Model 3)

An alternative model using `cigs` and`smokeyr` to calculate *pack years* `packyr` by multiplying # packs of cigarettes smoked per day (20 cigarettes = 1 pack) by the number of years the person has smoked to account for smoking history is worth a look, too.

```
bird <- bird %>%
    mutate(packs = cigs/20,
           packyr = packs*smokeyr)
```

```
model.3 <- glm(lungc ~ petbird + sex + ses + age + packyr,
               data = bird, family = "binomial")

m3_coeffs <- tidy(model.3, exponentiate = T, conf.int = T)

m3_coeffs
```

```
         term  estimate  std.error  statistic      p.value    conf.low
1 (Intercept) 0.1602469 1.66060458 -1.1026341 0.2701860965 0.005429896
2     petbird 4.1155399 0.40316578  3.5091521 0.0004495377 1.901275491
3        sexM 0.8636468 0.48792882 -0.3004361 0.7638445376 0.333141395
4      sesLow 1.1461406 0.45417523  0.3003252 0.7639291088 0.473084707
5         age 0.9906535 0.02876348 -0.3264723 0.7440670390 0.936450390
6      packyr 1.0355485 0.01177294  2.9670790 0.0030064369 1.012912783
  conf.high
1  3.811592
2  9.307369
3  2.285138
4  2.849116
5  1.049136
6  1.061255
```

1.5.1 Looking at the C statistic and R^2 for Model 2

```
mod3_lrm <- lrm(lungc ~ petbird + sex + ses + age + packyr,
                data = bird, x = TRUE, y = TRUE)
mod3_lrm
```

```
Logistic Regression Model

 lrm(formula = lungc ~ petbird + sex + ses + age + packyr, data = bird,
     x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           147    LR chi2     25.14    R2       0.218    C       0.749
  0             98    d.f.            5    g        1.128    Dxy     0.499
  1             49    Pr(> chi2) 0.0001    gr       3.089    gamma   0.499
 max |deriv| 4e-06                          gp       0.218    tau-a   0.223
                                            Brier    0.187

          Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -1.8310 1.6606 -1.10  0.2702
 petbird    1.4148 0.4032  3.51  0.0004
 sex=M     -0.1466 0.4879 -0.30  0.7638
 ses=Low    0.1364 0.4542  0.30  0.7639
 age       -0.0094 0.0288 -0.33  0.7441
 packyr     0.0349 0.0118  2.97  0.0030
```

1.5.2 Describing the Odds Ratio for `petbird` in Model 3

The odds ratio for lung cancer associated with `petbird` after adjustment for `age`, `sex`, `ses` (socio-economic status) and `packyr` (smoking pack-years) is 4.12 with 95% CI for that odds ratio of (1.9, 9.31).

This means that the odds of lung cancer for someone who keeps a pet bird were 4.12 times higher than the odds of lung cancer for someone who did not keep a bird, assuming that these two people had the same values of `sex`, `ses`, `age` and `packyr`.

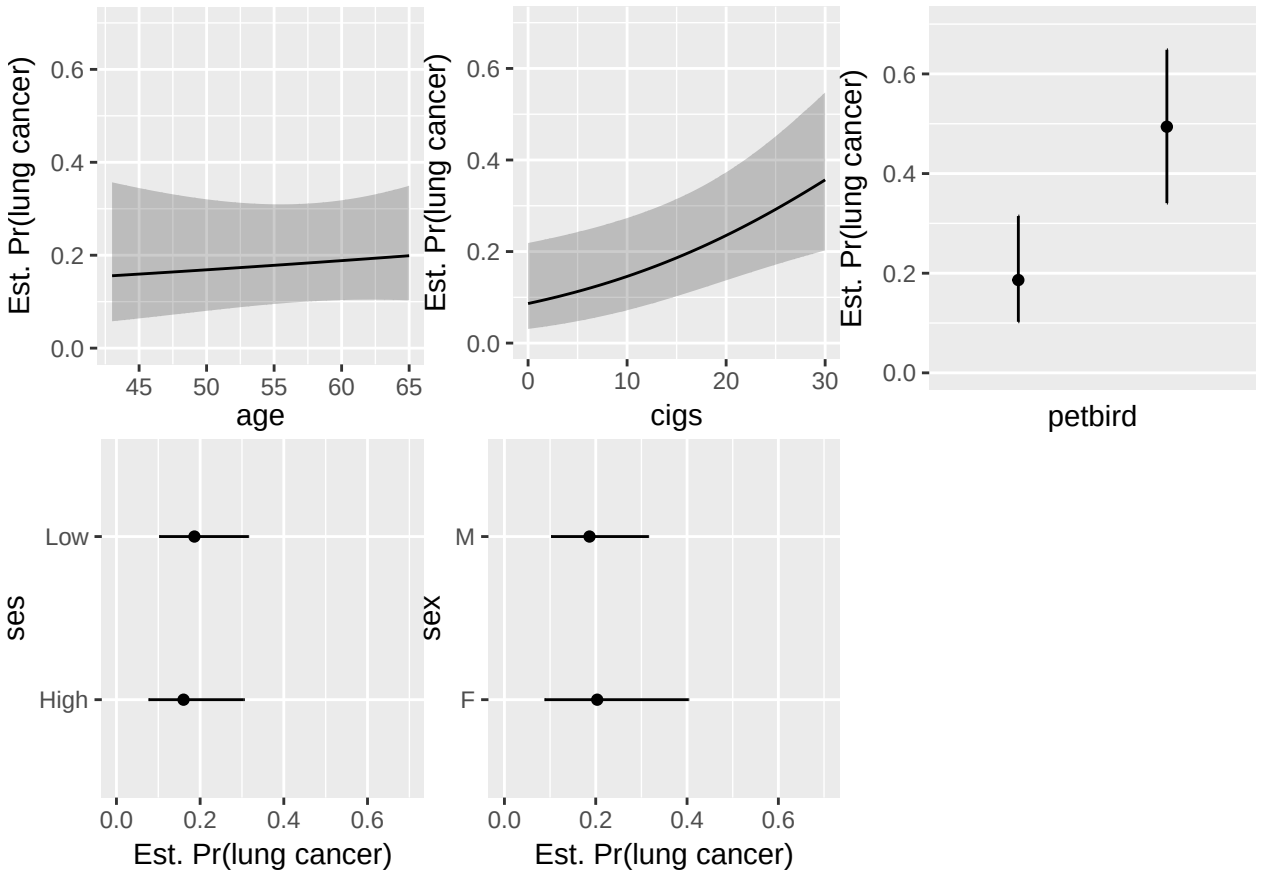
1.5.3 Evaluating Model 3

The C statistic and R^2 statistic adjusting for `packyr` turn out to be in between our previous two models.

1.5.4 Visualizing Model 3's Predictions

Here's the set of plots for Model 3.

```
ggplot(Predict(mod2_lrm, fun = plogis),
       ylab = "Est. Pr(lung cancer)")
```



2 Question 2 (40 points)

1. First, in 2-4 complete English sentences, I want you to specify, using your own words and complete English sentences, the most useful and relevant piece of advice you took away from

reading Jeff Leek's *How To Be A Modern Scientist*. Please provide a reference to the section of the book that provides this good advice. (For those of you who can more easily find things to gripe about in the book, don't worry - you will get that chance down the line.)

2. Then, in an essay of 4-8 additional sentences, describe why this particular piece of advice was meaningful or useful for you, personally, and how it will affect the way you move forward. You are encouraged to provide a specific example of a past or current scientific experience of yours that would have been (or is being) helped by this new approach or idea. Why is this idea important and worth sharing?

We don't write answer sketches for essay questions.