STAT X12 HW 10

Odds ratios and Logistic regression

Read Ashley et al. (2016) available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159878

- Focus on the progression after 80 weeks binary response variable and results related to it that are reported in Table 2.

- Individually or in groups of up to 4, complete the following. A 3% bonus if you can find someone to work with you have not completed a HW with previously, noted in the last question.

1) I used their suggested cutoff for ICOS in row 5 of Table 2 to create a new binary variable (below/above 8031.61). What percent of the responses are below 8031.61? Then make a contingency table of above/ below the ICOS cutoff and the response of `Progression_Event`. Use that table to estimate and report the odds of progression at 80 weeks in the high and low ICOS groups.

```
d1 <- read_excel("journal.pone.0159878.s002.XLSX")
summary(d1$ICOS)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5197    7211    8746    9615   11041   18848

d1 <- d1 %>% mutate(ICOSF = factor(ifelse(ICOS<8031.61, "low", "high")))
summary(d1$ICOSF)

## high  low
##   36   24

24/ (36+24)

## [1] 0.4

tally(Progression_Event ~ ICOSF, data = d1)

##                   ICOSF
## Progression_Event high low
##                 0    9  16
##                 1   27   8
```

- Note that in `Progression_Event`, the 0s are the “non-progressors” and “progressors” are the 1’s.

- Percent of responses: 24 (low) / 36 + 24 (total) = 40% (percent under the cutoff)

- Odds of progression at 80 weeks: 1 = progressors

- When patients are above the ICOSF threshold of 8031.61 RFU, they are (27/9) 3 to 1 more likely to progress. When patients are below the threshold, they are (8/16) 0.5 to 1 more likely to progress.

2) Use the previous results to estimate and interpret (report in a sentence) the odds ratio of progression between the ICOS defined groups. Hint: this will match the reported results in the "Unadjusted OR" column. Show your work. No CI needed at this point.

- High threshold progress / low threshold progress = 3/0.5 = 6
- The estimated mean odds of success is 6 times higher in the high threshold group than in the low threshold group.

3) Use a logistic regression model to obtain the same result. So fit a logistic regression model with the binary ICOSF predictor and binary progression response. Then work with the correct coefficient to get the same estimated OR as in the previous question. Show your work. You can use `tbl_regression(intercept = T, exponentiate = T)` to get a start on the results you need or `summary` and `confint`.

```
m1 <- glm(Progression_Event ~ ICOSF, family = 'binomial', data = d1)
summary(m1)
```

```
##
## Call:
## glm(formula = Progression_Event ~ ICOSF, family = "binomial",
##     data = d1)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0986     0.3849   2.854  0.00431
## ICOSFlow     -1.7918     0.5794  -3.093  0.00198
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 81.503  on 59  degrees of freedom
## Residual deviance: 71.041  on 58  degrees of freedom
## AIC: 75.041
##
## Number of Fisher Scoring iterations: 4
```

```
tbl_regression(m1, intercept = T, exponentiate = T)
```

Characteristic	OR[1]	95% CI[1]	p-value
(Intercept)	3.00	1.47, 6.76	0.004
ICOSF			
high	—	—	
low	0.17	0.05, 0.50	0.002

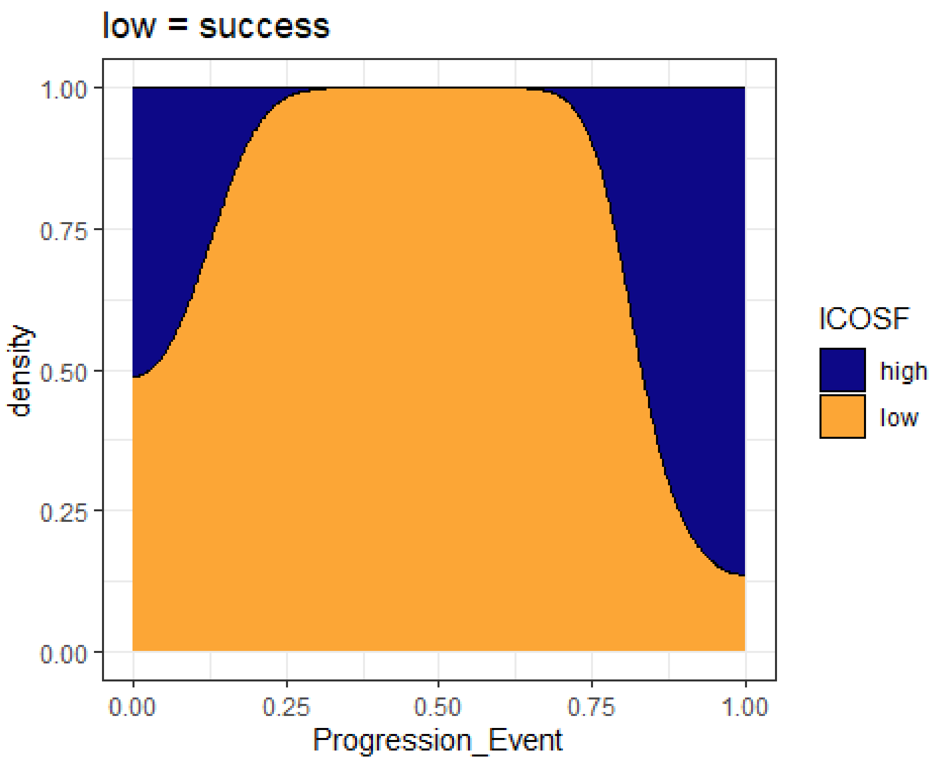
[1] OR = Odds Ratio, CI = Confidence Interval

```
exp(1.78)

## [1] 5.929856

exp(confint(m1))

##                  2.5 %    97.5 %
## (Intercept) 1.46606734 6.7591122
## ICOSFlow    0.05082888 0.5008416
```

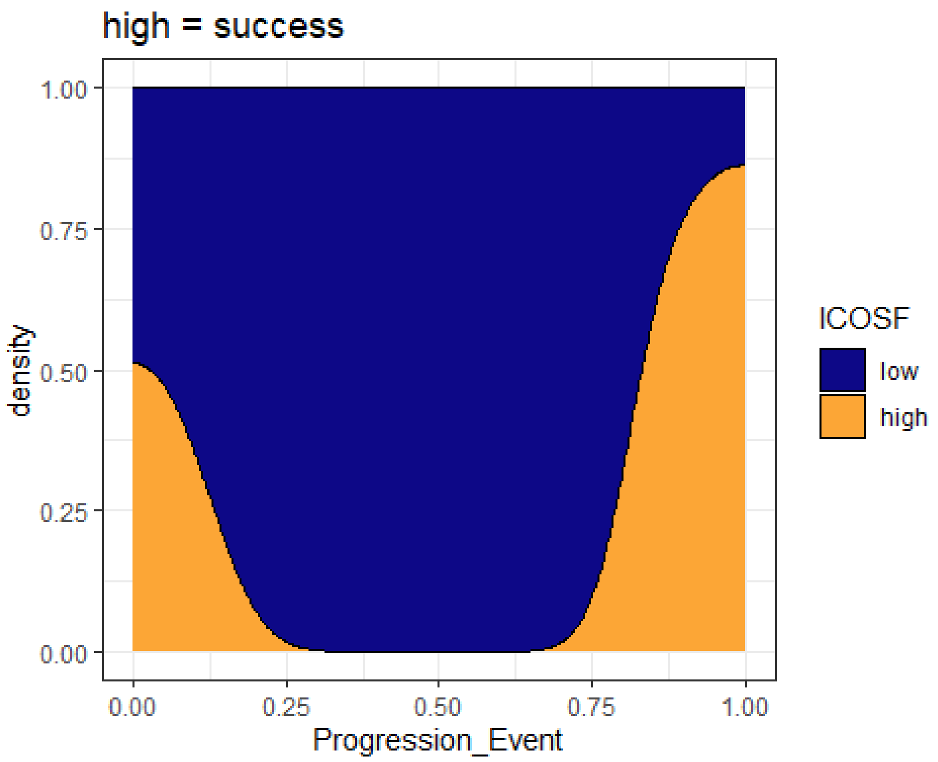
- The reference group for progression event is the no progress group. This means the intercept is the estimate for the progress group. By exponentiating this value, we report a OR of 5.93 ~ 6 with 95% CI (0.80143754, 4.2011535).

4) Now that you have matched the results in the “Unadjusted” situation with the logistic regression model, we can try to match their “Adjusted” inferences. Adjusted models contain `Age`, smoking status (code to rename to `Smoke` below), gender (`Male`), baseline FVC (`FVC_predicted_percent`), and DLCO (`DLCO_predicted_percent`) predicted percentages. Check for missing data on the combination of these variables, the ICOSF binary variable, and the binary response. “Clean” the data set if needed and then report the sample size total and as analyzed in the “adjusted” model that will contain these predictors.

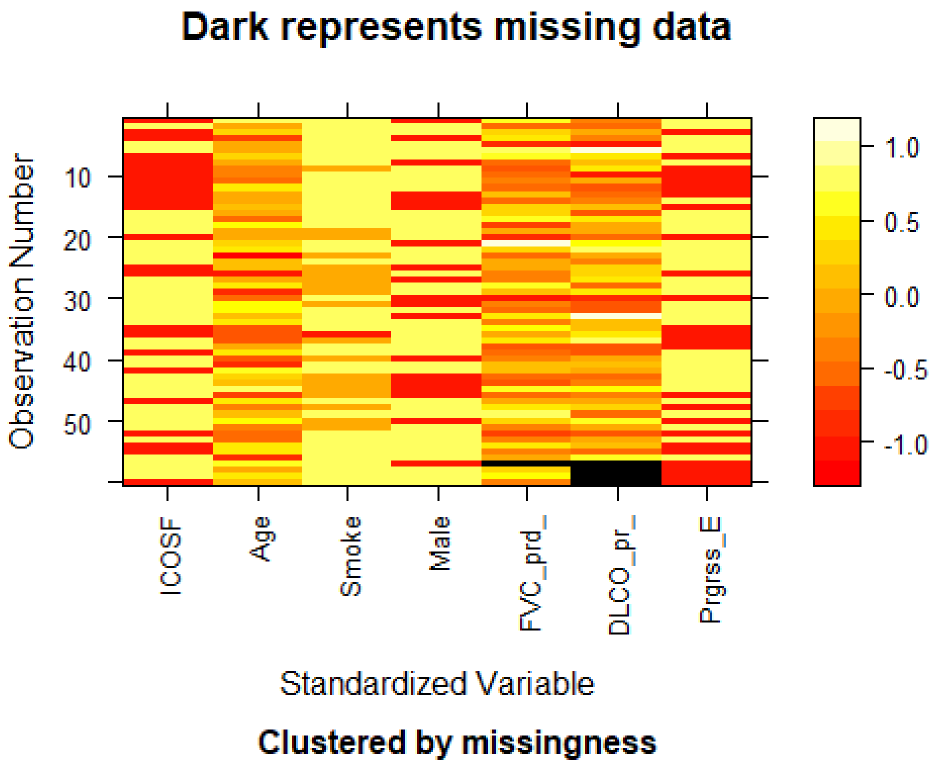
```
d1 <- d1 %>% mutate(Male = factor(Male)) %>%
             dplyr::rename(Smoke = 'Smoke Status') %>%
             mutate(Smoke = factor(Smoke))

ggplot(d1, aes(fill = ICOSF, x = Progression_Event))+
  geom_density(position = 'fill')+
  scale_fill_viridis_d(end = 0.8, option = "C")+
  labs(title = 'low = success')
```



```r
d1$ICOSF <- relevel(d1$ICOSF, ref = 'low')

ggplot(d1, aes(fill = ICOSF, x = Progression_Event))+
  geom_density(position = 'fill')+
  scale_fill_viridis_d(end = 0.8, option = "C")+
  labs(title = 'high = success')
```



```
d2 <- d1 %>%
  mutate(Age = as.numeric(Age),
         FVC_predicted_percent = as.numeric(FVC_predicted_percent),
         DLCO_predicted_percent = as.numeric(DLCO_predicted_percent)) %>%
  dplyr::select(ICOSF, Age, Smoke, Male, FVC_predicted_percent, DLCO_predicte
d_percent, Progression_Event)

library(mi)
tdf <- missing_data.frame(data.frame(d2))

## NOTE: In the following pairs of variables, the missingness pattern of the
first is a subset of the second.
##  Please verify whether they are in fact logically distinct variables.
##      [,1]                    [,2]
## [1,] "FVC_predicted_percent" "DLCO_predicted_percent"

image(tdf)
```



```
d2 <- d2 %>%
  na.omit()
```

- 56 is the final sample size

5) Fit a logistic regression model containing those components to create the “adjusted” model for the ICOS binary predictor. Make effects plots on both the link and response scales - do not include partial residuals in the plots. No discussion.

```
m2 <- glm(Progression_Event ~ ICOSF + Age + Smoke + Male + FVC_predicted_perc
ent + DLCO_predicted_percent, family = 'binomial', data = d2)
summary(m2)

##
## Call:
## glm(formula = Progression_Event ~ ICOSF + Age + Smoke + Male + 
##     FVC_predicted_percent + DLCO_predicted_percent, family = "binomial", 
##     data = d2)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.998e+01  2.400e+03  -0.008 0.993357
## ICOSFhigh                2.600e+00  7.865e-01   3.306 0.000948
## Age                      5.555e-02  4.726e-02   1.175 0.239837
## SmokeNever               1.458e+01  2.400e+03   0.006 0.995152
## SmokePast                1.556e+01  2.400e+03   0.006 0.994827
## Male1                   -8.020e-01  7.415e-01  -1.082 0.279450
## FVC_predicted_percent   5.671e-03  2.805e-02   0.202 0.839755
## DLCO_predicted_percent  1.130e-02  3.772e-02   0.300 0.764525
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.095  on 55  degrees of freedom
## Residual deviance: 55.148  on 48  degrees of freedom
## AIC: 71.148
##
## Number of Fisher Scoring iterations: 15
```

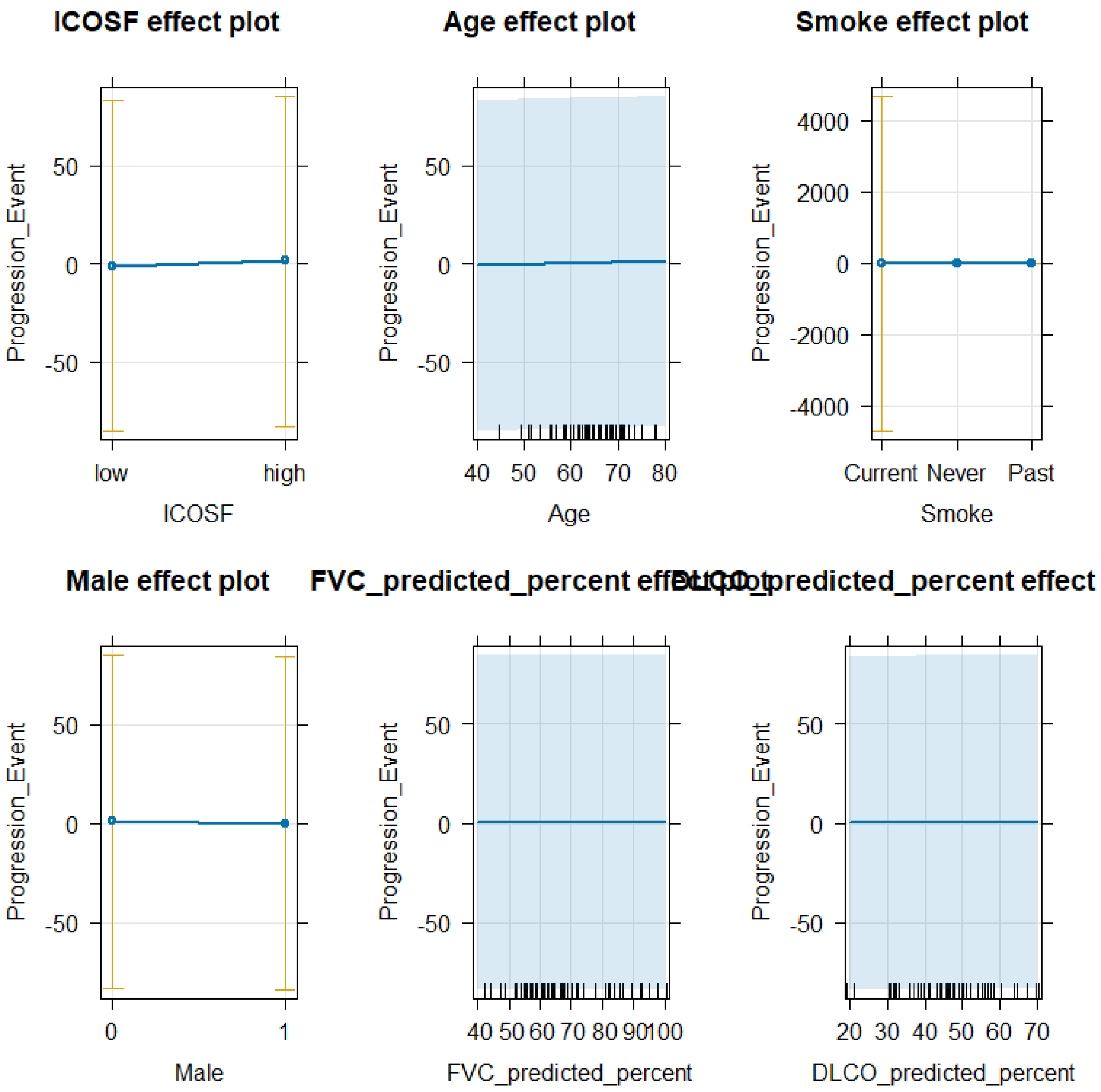
```
confint(m2)
```

```
##                                2.5 %       97.5 %
## (Intercept)                       NA 448.16239340
## ICOSFhigh                 1.18248907   4.33324293
## Age                      -0.03379142   0.15603486
## SmokeNever             -471.94649070           NA
## SmokePast              -472.01917316           NA
## Male1                    -2.33489565   0.61886452
## FVC_predicted_percent    -0.04899226   0.06358558
## DLCO_predicted_percent   -0.06143374   0.09028005
```

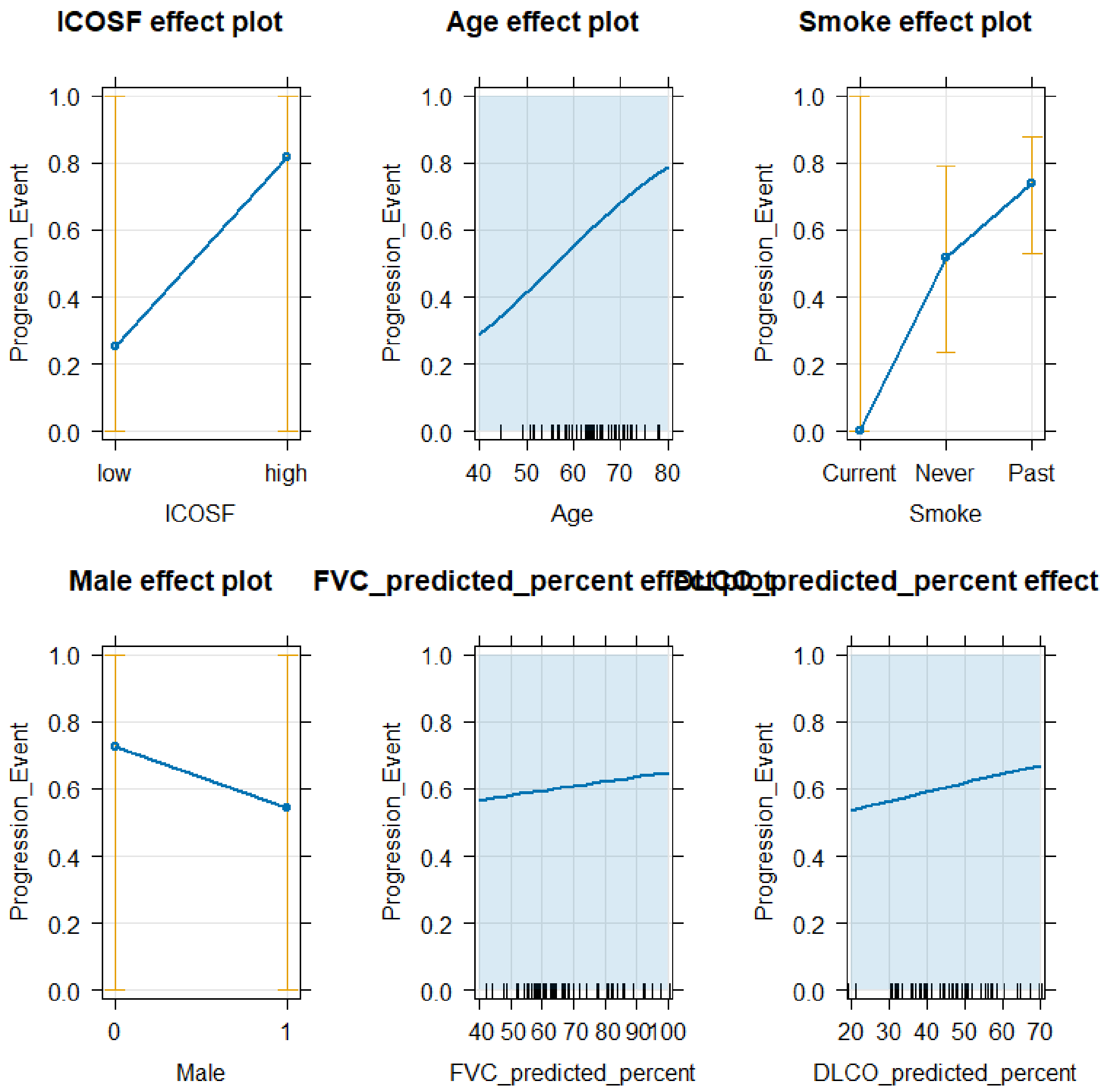
```
Anova(m2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Progression_Event
##                        LR Chisq Df Pr(>Chisq)
## ICOSF                   14.3095  1  0.0001551
## Age                      1.4602  1  0.2268922
## Smoke                    1.7694  2  0.4128362
## Male                     1.2091  1  0.2715008
## FVC_predicted_percent    0.0411  1  0.8393025
## DLCO_predicted_percent   0.0908  1  0.7631493
```

```
plot(allEffects(m2), type = 'link', grid = T)
```



```
plot(allEffects(m2), type = 'response', grid = T)
```



6) Report an evidence sentence for the ICOS predictor in the previous model.

- There is strong evidence against the null hypothesis of no difference in progress between ICOS threshold (z = 2.600e+00, p-value = 0.0009), controlled for Age, smoking status, gender, baseline FVC , and DLCO predicted percentages.

7) Then report a size interpretation on the odds scale for ICOSF from the adjusted model (include a 95% CI, noting that it might not match the one in the paper).

```
summary(m2)

##
## Call:
## glm(formula = Progression_Event ~ ICOSF + Age + Smoke + Male +
##     FVC_predicted_percent + DLCO_predicted_percent, family = "binomial",
##     data = d2)
##
```

```
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.998e+01  2.400e+03  -0.008 0.993357
## ICOSFhigh               2.600e+00  7.865e-01   3.306 0.000948
## Age                     5.555e-02  4.726e-02   1.175 0.239837
## SmokeNever              1.458e+01  2.400e+03   0.006 0.995152
## SmokePast               1.556e+01  2.400e+03   0.006 0.994827
## Male1                  -8.020e-01  7.415e-01  -1.082 0.279450
## FVC_predicted_percent   5.671e-03  2.805e-02   0.202 0.839755
## DLCO_predicted_percent  1.130e-02  3.772e-02   0.300 0.764525
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.095  on 55  degrees of freedom
## Residual deviance: 55.148  on 48  degrees of freedom
## AIC: 71.148
##
## Number of Fisher Scoring iterations: 15
```

```
# exp the inverse of the high estimate
exp(-2.600e+00)
```

```
## [1] 0.07427358
```

```
# exp the high estimate
exp(2.600e+00)
```

```
## [1] 13.46374
```

```
exp(confint(m2))
```

```
##                                 2.5 %        97.5 %
## (Intercept)                       NA 4.309774e+194
## ICOSFhigh               3.262485e+00  7.619097e+01
## Age                     9.667731e-01  1.168867e+00
## SmokeNever             1.087035e-205            NA
## SmokePast              1.010829e-205            NA
## Male1                   9.682059e-02  1.856818e+00
## FVC_predicted_percent   9.521885e-01  1.065651e+00
## DLCO_predicted_percent  9.404153e-01  1.094481e+00
```

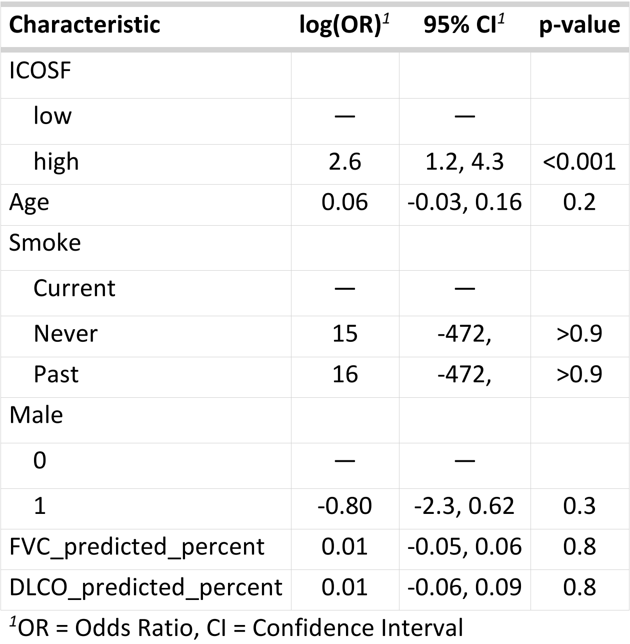
```
#low cis inverse and exp the high cis
exp(-3.262485e+00)
```

```
## [1] 0.03829312
```

```
exp(-7.619097e+01)
```

```
## [1] 8.141083e-34
```

```
tbl_regression(m2)
```



Characteristic	log(OR)[1]	95% CI[1]	p-value
ICOSF			
low	—	—	
high	2.6	1.2, 4.3	<0.001
Age	0.06	-0.03, 0.16	0.2
Smoke			
Current	—	—	
Never	15	-472,	>0.9
Past	16	-472,	>0.9
Male			
0	—	—	
1	-0.80	-2.3, 0.62	0.3
FVC_predicted_percent	0.01	-0.05, 0.06	0.8
DLCO_predicted_percent	0.01	-0.06, 0.09	0.8

[1] OR = Odds Ratio, CI = Confidence Interval

- High: The estimated mean odds of success is 13.46374 (95% CI (3.262485e+00, 7.619097e+01) times higher in the high threshold group than in the low threshold group.

- Low: The estimated mean odds of success is 0.07427358 (95% CI (0.03829312, 8.141083e-34) times lower in the low threshold group than in the high threshold group.

8) There is something undesirable happening in the previous model that is highlighted in the effects plots. What is it? Make a contingency table and plot of progressing/not vs the smoking status. What does this tell you?

- It appears that the smoking distribution cannot be used because there is 100% success and failure based on the progression event for the current smoking group. We know that if this is the case, the model will be unable to make an accurate log odds estimate because the denominator will either be 1 or 0.

```
tally(Progression_Event ~ Smoke, data = d2)

##                   Smoke
## Progression_Event Current Never Past
##                 0       1     7   13
##                 1       0    12   23
```

- This table shows that there is only one observation for current smoking. This means the estimate probability will be incorrect/ unreliable.

9) Write out the estimated "adjusted" model, defining all model components. Make sure you are clear about what is a "success" in the model.

- $Progression _0 + *1I*{ICOSF = high} + _2Age + *3I*{Smoke = Never} + *4I*{Smoke = Past} + *5I*{Male= 1} + _6FVCPredictedPercent + _7DLCOPredictedPercent $

Where beta 1 ICOSF high = 1, and 0 otherwise. Where beta 3 smoking never = 1, and 0 otherwise. Where beta 4 smoking past = 1, and 0 otherwise. Where beta 5 male = 1, and 0 otherwise.

- $logit(\hat{\pi}) = -1.998e + 01 + 2.600e + 00 + 5.555e - 02 + 1.458e + 01 + 1.556e + 01 - 8.020e - 01 + 5.671e - 03 + 1.130e - 02$

Where beta 1 ICOSF high = 1, and 0 otherwise. Where beta 3 smoking never = 1, and 0 otherwise. Where beta 4 smoking past = 1, and 0 otherwise. Where beta 5 male = 1, and 0 otherwise.

10) Use your estimated model to generate a predicted *probability* of progressing of a subject who is Age 60, male, past smoker, has both FVCpred and DLCOpred of 50%, and is in the high ICOS group. Do not use the `predict` function, although you can check your work with it. Show your work.

(-1.998e+01 (2.600e+00 * 0) + (5.555e-02*60) + (1.458e+01 * 0) + (1.556e+01* 1) (-8.020e-01*1) + (5.671e-03*.5) + (1.130e-02*.5))

```
-1.998e+01+ (5.555e-02*60) + (1.556e+01*1) -(8.020e-01*1)+ (5.671e-03*.5) + (
1.130e-02*.5)
```

```
## [1] -1.880514
```

11) Document any outside resources used. Also note if there is a new collaboration that qualifies you for the bonus. Report NONE if neither is the case.