STA 138 Final Project

STA 138 | Andrew Muench, Michelle Tsang, Connor Young | December 2023

Introduction

In this study we will be trying to understand whether the workplace dustiness had an effect for contracting byssinosis. This is important to investigate because if workplace conditions, in this scenario dustiness, had an effect on whether employees contracted byssinosis which would make the company liable. This would cost the company in compensating for the medical conditions of the employees with byssinosis. In order to investigate whether workplace conditions had an effect on contracting byssinosis, we will try our best to model the data and to observe the effect of the variables on the data.

Results

Below is the results of our analysis to find the model of best fit for the variables and to analyze the relationship between Workplace and byssonnossis. As we see in Figure 1, we have the summary of the logistic model of the additive terms of all variables. The variables that were statistically significant if the significance level was less than 0.0001 are SmokingYes, Employment >=20, and Workspace 1. The variables that were statistically significant if the significance level is 0.01 is employment 10-19, SmokingYes, Employment >=20, and Workspace 1 have positive coefficients which would suggest that as these variables increase so does the likelihood of having byssinosis. If we were to increase the significance level to 0.01 we could say the same of employment 10-19.Based on the confidence intervals of the coefficients of these variables, these variables all have positive coefficients. Which means that we are 95% confident for all three variables that they are positive relationships. In order to see which predictors we should include in our regression to fit the best model, we will use forward stepwise selection. It will start with the empty model and continue adding predictor variables with the AIC as the selection criterion. The final model after using forward stepwise selection includes the following predictor variables: the intercept (the base case is if Workspace is rated a 3, the worker doesn't smoke, and was employed for less than 10 years), Workspace, Smoking, and Employment.

In order to see which predictors we should include in our regression to fit the best model, we will use forward stepwise selection(Figure 2). It will start with the empty model and continue adding predictor variables with the AIC as the selection criterion. The final model after using forward stepwise selection includes the following predictor variables: the intercept (the base case is if Workspace is rated a 3, the worker doesn't smoke, and was employed for less than 10 years), Workspace, Smoking, and Employment.

We will use the Wald test to determine whether or not each variable from the stepwise model selection is significant with $H_0 : \beta_{Workspace2} = 0, \beta_{Workspace1} = 0, ..., \beta_{Employment10-19} = 0$. When conducting the Wald test on the model using a 0.05 significance level, we are able to reject the null hypothesis and find that every variable is significant *except* for Workspace2 since its p-value of 0.499 is greater than 0.05. There is not enough evidence to conclude that there is a difference in the chance of getting byssinosis if the Workspace was rated a 3 or 2.

We reach the same conclusion when we create a 95% confidence interval of the odds ratio for determining if there is an effect on that variable on the chances of getting byssinosis. Having an odds ratio of 1 means that variable is does not create a difference in the chances of getting byssinosis. In this case, only the Workspace2 variable contains 1 in its confidence interval of odds ratio (0.693, 2.119), so it is possible that there is no difference in the chance of getting byssinosis if the Workspace was rated a 3 or 2.

The odds ratio confidence intervals for the variables Workspace1, SmokingYes, Employment>=20, and Employment10-19 of (0.693, 2.119), (1.308, 2.783), (1.353, 2.762), and (1.017, 2.701) respectively, are all greater than 1, meaning having a Workspace rating of 1, being a smoker, being employed for more than 20 years, or being employed for 10-19 years increases the chance of having byssinosis. The intercept (base case of Workspace with rating of 3, no smoking, and being employed for less than 10 years) has an odds ratio confidence interval of

(0.003, 0.008) is less than 1, meaning if the worker is the base case, they are less likely to have byssinosis.

We will use the Wald test to test the independence between the categorical variable Workspace and the chance of getting byssinosis with $H_0 : \beta_{Workspace2} = 0$ and $H_0 : \beta_{Workspace1} = 0$ at 0.05 significance level(Figure 3). We find that both the intercept (when Workspace is rated a 3) and Workspace1 (Workspace rated a 3 vs 1) have a p-value approximately 0, so we have enough evidence to reject the null hypothesis and conclude that there is a relationship between the Workspace being rated a 3 or a 1 and the chance of having byssinosis.

However, the same can't be said for Workspace2 since its p-value of 0.588 is greater than 0.05. We fail to reject the null hypothesis, there is not enough evidence to say there is a relationship between Workspace being rated a 2 (vs rated a 3) and the chance of having byssinosis (Workspace2 is independent of having byssionsis).

We reach the same conclusion when constructing a 95% confidence interval of the odds ratio for each variable (Figure4). We are 95% confident that the odds ratio (OR) of the intercept (Workspace is rated a 3) and Workspace1 are between (0.008, 0.016) and (10.666, 22.436) respectively. Since OR = 1 is not in either of these interval, both these variables have a relationship on the chance on having byssinosis. Since the odds ratio interval of (0.008, 0.016) is less than 1, the chance of having byssinosis when Workspace is rated a 3 is less likely to occur. Conversely, the odds ratio interval of (10.666, 22.436) is greater than 1, so the chance of having byssinosis when Workspace is rated a 3 vs a rating of 1 is more likely to occur.

We are 95% confident that the odds ratio interval of the Workspace2 is between (0.667, 2.038). Since OR = 1 is contained in the interval, it is possible that Workspace being rated a 2 (vs rated a 3) is independent from the chance of having byssinosis. After combining Workplace2 and Workplace 3 to be less dusty and Workplace1 to be more dusty, we conduct another Wald test(Figure 4: Part 1) on all our predictors after doing forward stepwise model selection using 0.05 significance level. The new model contains the intercept, Workspace1 (more dusty), SmokingYes, Employment>=20, and Employment10-19.

The p-values for all the variables are less than 0.05, meaning we can reject the null hypothesis that and have sufficient evidence to conclude there is a difference between the variables listed above and the baseline case (workspace is less dusty, no smoking, and employment is less than 10 years) on the chance of having byssinosis.

The confidence intervals for the variables Workspace1, SmokingYes, Employment>=20, and Employment10-19 of (10.487, 20.456), (1.306, 2.778), (1.350, 2.755), and (1.012, 2.668) respectively, are have odds ratios greater than 1, meaning having a Workspace that is more dusty, being a smoker, being employed for more than 20 years, or being employed for 10-19 years increases the chance of having byssinosis.

After combining Employment>20 and Employment10-19 to be being employed for more than 10 years and Employment<10 to be employed for less than 10 years (Figure 4 Part 2), we conduct another Wald test on all our predictors after doing forward stepwise model selection using 0.05 significance level. The new model contains the intercept, Workspace1 (more dusty), SmokingYes, Employment>=10.

The p-values for all the variables are less than 0.05, meaning we can reject the null hypothesis that and have sufficient evidence to conclude there is a difference between the variables listed above and the baseline case (workspace is less dusty, no smoking, and employment is less than 10 years) on the chance of having byssinosis.

The confidence intervals for the variables Workspace1, SmokingYes, and Employment>=10 of (10.474, 20.426), (1.296, 2.753), (1.324, 2.582) respectively, are have odds ratios greater than 1, meaning having a Workspace that is more dusty, being a smoker, being employed for more than 20 years, or being employed for 10-19 years increases the chance of having byssinosis.

In order to see if any of the models were better fits compared to each other, we used the LRT test to compare all 3 combinations of the variables(Figure 5). Based on the p-values of the tests from these tests, which are all above 0.5, no complex model is a better fit than the simpler model. Which means we would employ the simplest model, which is the forward stepwise model with all interaction terms and the employee modifications.

To determine how well our model fits the data, we will look at its deviance residuals (Figure 6). A residual is a measure of the difference between an observed value and a predicted value. The most basic residual for our models is the raw difference between the observed value (Byssinosis status: $y_i \in \{0, 1\}$) and the predicted probability ($\hat{\pi_i} \in [0, 1]$). This residual will be denoted as e_i and has the form.

The deviance residual, d_i, is based on deviance which is derived from a likelihood ratio test comparing the fitted logistic model and a *saturated* model which fits each data point perfectly. The formula to calculate the deviance residuals is

$$\begin{aligned} d_i &= \text{sign}(e_i)\sqrt{\text{Deviance}} \\ &= \text{sign}(e_i)\sqrt{-2\left[y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)\right]}. \end{aligned}$$

The values of these deviance residuals measure how much the predicted probabilities from our model differ from the observed data and whether our model is under- or over-fitting for certain predictor variable values.

Large values for deviance residuals indicate poor fit and small values indicate good fit. A positive deviance residual indicates our model is under-predicting for a certain set of predictor variable values, i.e., the observed value is 1 and the predicted probability is less than 1. A negative deviance residual indicates our model is over-predicting for a certain set of predictor variable values, i.e., the observed value is 0 and the predicted probability is greater than 0.

A good model fit is indicated by deviance residuals which are evenly distributed and relatively close to 0. Thus, if a smooth line is drawn to "average" the deviance residuals, we would like this line to be relatively flat and close to the line at $d_i = 0$.

From the Figure 7, we see that the fitted smooth line is relatively straight (after log-odds of -4) and close to 0, with a peak at around $(-5, 0.8)$ and the rest of the line at $d_i \approx -0.35$. It appears that at low log-odds, this model is somewhat over-predicting, but at log-odds greater than around -4.75, this model is somewhat under-predicting. Overall, the deviance residuals appear to be relatively equally distributed above and below 0 (~37% and ~63% respectively), and ~95% of the residuals have an absolute value less than 1.5. Thus, we consider this model to be a relatively good fit.

Discussion

For the project we made use of these following models:

Model 1 : logistic regression model with all additive terms Model 2: forward stepwise logistic regression with all additive terms Model 3: the forward stepwise model with only additive terms of the statistically significant variables, Model 4: the forward stepwise model with all the interaction terms of the statistically significant variables and the previous model a modification to the employment term. This modification to the employment variable was to change employment to be from 10-19 years to be all employees with over 10 years or greater of employment.

Of all these models, the Model 4 had the smallest AIC() and BIC(). AIC and BIC are measurements of goodness of fit of models that account for model complexity. Having a lower AIC and BIC value would mean that the model would be a better fit comparatively to other models. This would suggest that the Model 4 had the best fit. We also conducted a likelihood ratio test between all pairings of the models with lowest AICs and BICS which were Models 2,3 and 4 From these tests we all had large p-values(respectively) between the three combinations of the LRT test which would mean that the simpler model is not significantly worse than more complex plots. However, this does not mean that the simplest model is the best fit. The model with all the interaction terms of the statistically significant variables with the employee modification had the best fit. Observing the diagnostic plot of model 4 shows that the deviance residuals are relatively equally distributed which would suggest that the model is a relatively good fit. With the AICs, BICs, LRT and the diagnostic plots we conclude that model 4 is the best fit

Conclusion

Beginning from the model that best fits the data from above, we can look at the summary of the model. This summary of the model would suggest that Workspace 1, Employment >=10 and SmokingYes all have positive and statistically significant effects on the chance of contracting byssinosis. Using the Wald tests and the confidence intervals of the odds ratios for the variables, we see that all of these variables have statistically significant effects on the chance of contracting byssinosis. Because Workspace 1, the most dusty workplace, is among these statistically significant variables with a positive effect on the chance of contracting byssinosis, we conclude that Workplace dustiness positively contributes to the chance of byssinosis.

Figure 1: Logistic Regression: All Additive Terms and No Interactions

```
## $Employment
## [1] "<10"   "10-19" ">=20"
##
## $Smoking
## [1] "Yes" "No"
##
## $Sex
## [1] "M" "F"
##
## $Race
## [1] "W" "O"
##
## $Workspace
## [1] 1 2 3
## Levels: 3 2 1
```

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Employment +
##     Smoking + Sex + Race + Workspace, family = binomial(), data = byss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5076  -0.8218  -0.2693   0.2879   1.5709
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -5.1028     0.2643 -19.308  < 2e-16 ***
## Employment>=20    0.7444     0.2169   3.432 0.000600 ***
## Employment10-19   0.5678     0.2619   2.168 0.030177 *
## SmokingYes        0.6709     0.1961   3.421 0.000625 ***
## SexM             -0.1503     0.2296  -0.655 0.512777
## RaceW            -0.1176     0.2077  -0.566 0.571270
## Workspace2        0.1713     0.2870   0.597 0.550546
## Workspace1        2.7617     0.2172  12.713  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.114  on 64  degrees of freedom
## Residual deviance:  42.679  on 57  degrees of freedom
## AIC: 164.98
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 182.3771
```

```
##                       2.5 %     97.5 %
## (Intercept)     -5.62081192 -4.5848382
## Employment>=20   0.31923027  1.1695809
## Employment10-19  0.05443102  1.0812555
## SmokingYes       0.28650604  1.0553657
## SexM            -0.60034204  0.2997582
## RaceW           -0.52467718  0.2894865
## Workspace2      -0.39121682  0.7338872
## Workspace1       2.33596023  3.1874919
```

Figure 2: Forward Step-Wise Regression: All interactions of statistically significant variables

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Workspace +
##     Smoking + Employment, family = binomial(), data = byss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6131  -0.8065  -0.2514   0.3084   1.6743
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -5.2077     0.2375 -21.931  < 2e-16 ***
## Workspace2        0.1925     0.2852   0.675 0.499666
## Workspace1        2.7393     0.1914  14.315  < 2e-16 ***
## SmokingYes        0.6464     0.1925   3.358 0.000786 ***
## Employment>=20    0.6596     0.1820   3.625 0.000289 ***
## Employment10-19   0.5055     0.2492   2.029 0.042508 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.114  on 64  degrees of freedom
## Residual deviance:  43.429  on 59  degrees of freedom
## AIC: 161.73
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 174.779
```

```
##                       2.5 %     97.5 %
## (Intercept)     -5.67315243 -4.7423365
## Workspace2      -0.36641854  0.7513960
## Workspace1       2.36420684  3.1143267
## SmokingYes       0.26909047  1.0236948
## Employment>=20   0.30292282  1.0161852
## Employment10-19  0.01708115  0.9938192
```

Figure 3: Testing for Independence between workspace and byssinosis.

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Workspace,
##     family = binomial, data = byss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3374  -0.9329  -0.2774   0.2242   3.4086
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.4200     0.1571 -28.133   <2e-16 ***
## Workspace2    0.1542     0.2846   0.542    0.588
## Workspace1    2.7389     0.1897  14.439   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.114  on 64  degrees of freedom
## Residual deviance:  70.192  on 62  degrees of freedom
## AIC: 182.5
##
## Number of Fisher Scoring iterations: 5
```

Figure 4: Variable modificationsPart 1

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Workspace +
##     Smoking + Employment, family = binomial(), data = byss2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6156  -0.7821  -0.2584   0.3073   1.8354
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -5.1495     0.2190 -23.516  < 2e-16 ***
## Workspace1        2.6842     0.1704  15.749  < 2e-16 ***
## SmokingYes        0.6445     0.1925   3.349 0.000812 ***
## Employment>=20    0.6571     0.1819   3.612 0.000304 ***
## Employment10-19   0.5007     0.2490   2.011 0.044352 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.114  on 64  degrees of freedom
## Residual deviance:  43.874  on 60  degrees of freedom
## AIC: 160.18
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 171.0489
```

```
##                       2.5 %     97.5 %
## (Intercept)     -5.57871322 -4.7203201
## Workspace1       2.35019087  3.0182868
## SmokingYes       0.26727387  1.0217511
## Employment>=20   0.30050913  1.0136781
## Employment10-19  0.01264627  0.9888278
```

Figure 4: Variable modificationsPart 2

```
##
## Call:
## glm(formula = cbind(Byssinosis, Non.Byssinosis) ~ Workspace +
##     Employment + Smoking, family = binomial(), data = byss2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7071  -0.8030  -0.2876   0.3195   1.8273
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.1427     0.2185 -23.538  < 2e-16 ***
## Workspace1       2.6829     0.1704  15.746  < 2e-16 ***
## Employment>=10   0.6148     0.1704   3.607  0.00031 ***
## SmokingYes       0.6364     0.1920   3.314  0.00092 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.114  on 64  degrees of freedom
## Residual deviance:  44.288  on 61  degrees of freedom
## AIC: 158.59
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 167.289
```

```
##                     2.5 %     97.5 %
## (Intercept)    -5.5708797 -4.7144360
## Workspace1      2.3489479  3.0168555
## Employment>=10  0.2807494  0.9488551
## SmokingYes      0.2600059  1.0128046
```

Figure 5: Comparing models with LRT

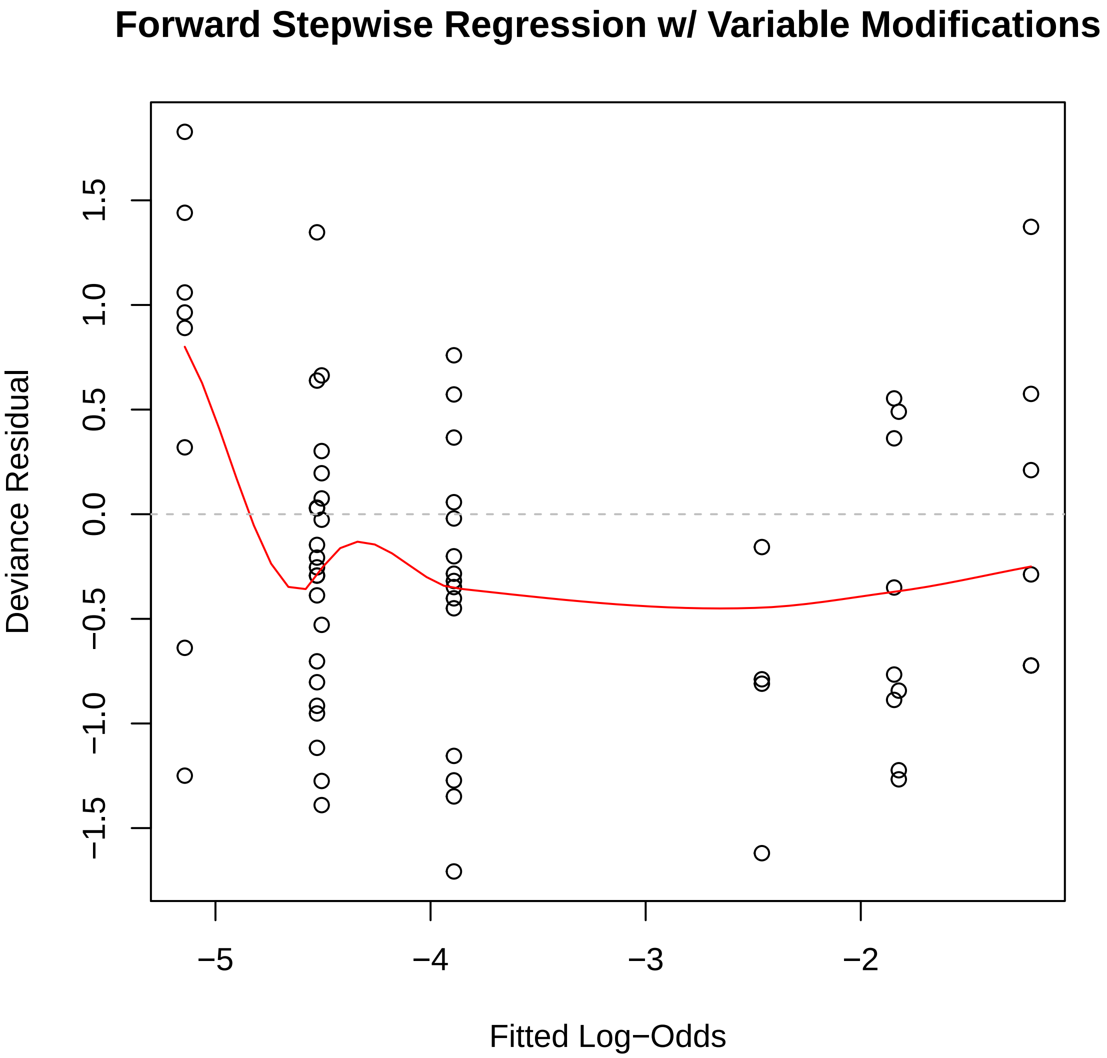
```
## [1] 0.6870723
```

```
## [1] 0.5050566
```

```
## [1] 0.650914
```

```
## [1] 0.5197143
```

Figure 6: Logistic Diagnostics



Appendix: R Script

```r
knitr::opts_chunk$set(echo = TRUE)
byss = read.csv("./Dataset/Byssinosis.csv")
# change Workspace to factor
byss$Workspace = factor(byss$Workspace, levels = c("3", "2", "1"))
# add Total column equal to Byssinosis + Non.Byssinosis
byss$Total = byss$Byssinosis + byss$Non.Byssinosis
# remove rows which have no total
byss = byss[byss$Total!=0,]
# unique values of each predictor variable
sapply(byss[1:5], function(x) unique(x))
glm1 = glm(cbind(Byssinosis, Non.Byssinosis)~Employment + Smoking + Sex + Race + Workspace,
           family=binomial(),
           data=byss)
summary(glm1)
BIC(glm1)
confint.default(glm1)
result1 = step(glm(cbind(Byssinosis, Non.Byssinosis)~1, binomial(), byss),
                  scope = ~Employment*Smoking*Sex*Race*Workspace,
                  k=log(72),
                  trace=0,
                  direction="forward")
summary(result1)
BIC(result1)
confint.default(result1)
glm2 = glm(cbind(Byssinosis, Non.Byssinosis)~Workspace, binomial, byss)
summary(glm2)
# changing Workplace var
# 2 & 3 -> 3 (less dusty)
# 1 -> 1 (more dusty)
byss2 = byss
byss2$Workspace[byss2$Workspace == 2] = 3
new1 = step(glm(cbind(Byssinosis, Non.Byssinosis)~1, binomial(), byss2),
                  scope = ~Employment*Smoking*Sex*Race*Workspace,
                  k=log(72),
                  trace=0,
                  direction="forward")
summary(new1)
BIC(new1)
confint.default(new1)
# changing Employment var
# 10-19 & >=20 -> >=10
byss2$Employment[byss2$Employment == "10-19" | byss2$Employment == ">=20"] = ">=10"
# new stepwise w/ Employment modification
new2 = step(glm(cbind(Byssinosis, Non.Byssinosis)~1, binomial(), byss2),
                  scope = ~Employment*Smoking*Sex*Race*Workspace,
                  k=log(72),
                  trace=0,
                  direction="forward")
summary(new2)
BIC(new2)
confint.default(new2)
# H_0: models are equivalent -> use simpler model
# H_1: models are not equivalent -> use more complex model
anv = anova(result1, glm1)
```

```
pvalue = 1-pchisq(anv[2, 4], anv[2,3])
pvalue
1-pchisq(anova(new1, result1)[2, 4], anova(new1, result1)[2,3]) # compare new1 & result1 w/ LRT
1-pchisq(anova(new2, result1)[2, 4], anova(new2, result1)[2,3]) # compare new2 & result1 w/ LRT
1-pchisq(anova(new2, new1)[2, 4], anova(new2, new1)[2,3]) # compare new2 & new1 w/ LRT
# new2
ry = residuals(new2, type="deviance")
rx = qlogis(fitted.values(new2))

plot(rx, ry,
     xlab = "Fitted Log-Odds",
     ylab = "Deviance Residual",
     main = "Forward Stepwise Regression w/ Variable Modifications",
     col = "black")
lines(loess.smooth(rx, ry), col="red")
abline(h=0, lty=2, col="grey")
```