Exploring the Impact of Workplace Factors on Byssinosis Prevalence: A Statistical Analysis of a 1973 Cotton Textile Company Study in North Carolina

STA 138 | Julie Lee and Michelle Wong | December 2023

INTRODUCTION

Byssinosis, a form of pneumoconiosis prevalent among workers in the textile industry, remains a critical occupational health concern. This study delves into the extensive dataset gathered from a prominent cotton textile company in North Carolina in 1973, encompassing 5,419 workers. The primary focus of this investigation lies in understanding the associations between byssinosis and potential predictor variables: workplace dustiness, years of employment, smoking status, gender, and race. The central question guiding this research is to determine whether workplace dustiness significantly contributes to the chance of byssinosis.

To answer this, the study employs a robust arsenal of statistical methods, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cross-Validation Information Criteria (AICc, CV-AIC), and Likelihood Ratio Tests. The exploration of these statistical tools aims to shed light on the role of workplace dustiness in the occurrence of byssinosis among a diverse workforce.

RESULTS

Data Checking for multicollinearity:

If the VIF values are high >5, it may indicate multicollinearity. However, we see that the numbers are low so there is not a lot of collinearity, which indicates independence of predictor variables and no need to fit Ridge or Lasso Regression Models. We can also proceed to use BIC and AIC in our analysis.

```
               GVIF Df GVIF^(1/(2*Df))
Employment 1.478891  2        1.102768
Smoking    1.058018  1        1.028600
Sex        1.259135  1        1.122112
Race       1.533393  1        1.238302
Workspace  1.264409  1        1.124459
```

BIC (Bayesian Information Criterion) Model Selection:

Method: The bidirectional stepwise selection is utilized in regression to construct the model that best predicts whether or not an individual is susceptible to Byssinosis by iteratively adding predictor variables to an empty model. The empty model is characterized by only the y-intercept (-3.466), representing a model for an individual that is not a smoker, is a female, is within the "other" race group, and does not take into account the type of workplace and employment years when predicting the onset of Byssinosis. This iterative approach utilizes the BIC Criterion (Bayesian Information Criterion as characterized by -2*log-likelihood+p*log(n)) which is based on likelihood and puts a heavier weight on the complexity of the model. In other words, BIC prefers a smaller model than AIC when adding additional predictors to the empty model as the penalty term(p*log*(n)) penalizes

the model with more parameters. To avoid overfitting, we randomly subset half of our observations (2709) as a testing set and perform a Bidirectional stepwise selection with BIC on the remaining training set.

Results: Bidirectional stepwise selection starting with the empty model (y-intercept B0 = -3.466) as using BIC results in the model: Byss ~ Workplace + Smoking: $\log(pi/1-pi) = B_0 + B_1x_{Workplace2} + B_2x_{Workplace3} + B_3x_{Smoking}$ where pi is the conditional probability of having Byssinosis. This indicates that the predictor variables Smoking status and type of workplace play the most influential role in predicting whether an individual is susceptible to Byssinosis. As each predictor variable was added to the empty model, R chooses to keep the model that has the lowest BIC, taking into account both the optimal fit of the model and complexity. The best fitted model has the following estimated parameters:

```
Call:  glm(formula = Byss ~ Workspace + Smoking, family = binomial,
    data = dataTrain)

Coefficients:
(Intercept)   Workspace2   Workspace3   SmokingYes
    -2.1997      -2.2303      -2.9558       0.7323

Degrees of Freedom: 2707 Total (i.e. Null);  2704 Residual
  (1 observation deleted due to missingness)
Null Deviance:      735
Residual Deviance: 591.4        AIC: 599.4
```

B_0 = -0.7341, $B_{1(Workspace\ 2)}$= -2.2303, $B_{2(Workspace3)}$= -2.9558, $B_{3(SmokingYes)}$= 0.7323

These estimated parameters suggest that the chance of having Byssinosis increases when an individual is a smoker and decreases when an individual is in both workspace 2 and 3, where workspace 3 has a heavier contribution in decreasing the chance of having Byssinosis (larger negative magnitude).

```
Start:  AIC=741.72
Byss ~ 1

             Df Deviance    AIC
+ Workspace   1   606.02 619.39
+ Sex         1   713.18 726.55
+ Smoking     1   722.48 735.85
<none>            735.04 741.72
+ Race        1   729.49 742.85
+ Employment  2   732.23 752.29
```

```
Step:  AIC=619.39
Byss ~ Workspace

             Df Deviance    AIC
+ Smoking     1   597.39 617.45
<none>            606.02 619.39
+ Sex         1   603.97 624.03
+ Race        1   605.96 626.02
+ Employment  2   600.74 627.48
- Workspace   1   735.04 741.72
```

```
Step:  AIC=617.45
Byss ~ Workspace + Smoking

                      Df Deviance    AIC
<none>                     597.39 617.45
- Smoking              1   606.02 619.39
+ Sex                  1   596.73 623.47
+ Race                 1   597.33 624.07
+ Smoking:Workspace    1   597.38 624.12
+ Employment           2   591.99 625.41
- Workspace            1   722.48 735.85
```

Wald Test on the Model Fit (obtained from BIC):

Method: Using the testing data, we perform Wald Tests to test for evidence of nonzero coefficients, using alpha = 0.05. The null hypothesis of the Wald Test states that the chance of an individual having Byssinosis is not associated with the type of workspace they work in, along with the smoking status. It is important to note that when performing multiple hypothesis tests, the overall chances for a type I error between different tests (in this case, we are testing the association between Byssionis and 2 predictor variables) increases. Therefore, we control for multiple testing by applying the Bonferroni Correction.

```
             Estimate Std. Error   z value     Pr(>|z|)
Workspace2 -2.9474305  0.4381760 -6.726591 1.736842e-11
Workspace3 -2.4846079  0.2535602 -9.798889 1.138306e-22
SmokingYes  0.5886574  0.2684215  2.193034 2.830495e-02
```

Results: By extracting the coefficient estimates from the testing set, we are able to see that they differ from those of the model fit we obtained before; such a result is expected as the testing and training sets utilize different observations. With an alpha value of 0.05, we apply the Bonferroni Correction and compare the 3 individual p-values to 0.05/3 = 0.0167. Because all the p-values are less than this value, we reject the null hypothesis. We are able to state that we have sufficient evidence to conclude that all 3 slopes are nonzero and therefore both smoking status and type of workplace has an association with the chance of predicting whether or not an individual has Byssinosis. However, it is important to note that the use of the Bonferroni Correction to test 3 Wald Tests can be overly conservative. Therefore, it is sometimes more ideal to utilize the likelihood ratio test (LRT) which is a statistical test that compares the goodness of fit between the null and alternative models which we will later be conducting.

If we re-utilized the training set for testing, we obtain the following coefficient estimates and smaller p-values for all 3 Wald Tests which is the expected observation as we re-used the data:

```
             Estimate Std. Error    z value     Pr(>|z|)
Workspace2 -2.2302529  0.3280188  -6.799161 1.052300e-11
Workspace3 -2.9557985  0.2923842 -10.109295 5.024433e-24
SmokingYes  0.7323049  0.2732616   2.679868 7.365125e-03
```

To further confirm our results, we conduct Wald Tests for the full model that additionally includes the 3 predictor variables, SexM, RaceW, and employment type. With an alpha value of 0.05, we apply the Bonferroni Correction and compare the 3 individual p-values to 0.05/7 = 0.00714. We are able to see that the p-values associated with the predictor variables, SexM, RaceW, and type of Employment are all greater than 0.00714 indicating that we have sufficient evidence to conclude that all 4 slopes are zero and therefore does NOT have an association with the chance of predicting whether or not an individual has Byssinosis.

```
                  Estimate Std. Error    z value     Pr(>|z|)
(Intercept)     -2.3406152  0.3800969 -6.1579435 7.369564e-10
SmokingYes       0.7573332  0.2814621  2.6907111 7.129992e-03
Workspace2      -2.2370455  0.3810170 -5.8712484 4.325255e-09
Workspace3      -2.9781570  0.3275942 -9.0909943 9.813741e-20
SexM            -0.1127755  0.3341558 -0.3374939 7.357446e-01
RaceW           -0.2526419  0.2939121 -0.8595833 3.900188e-01
Employment>=20   0.7337476  0.3066022  2.3931583 1.670403e-02
Employment10-19  0.4650135  0.3796974  1.2246951 2.206901e-01
```

Using Wald Test to obtain Confidence Intervals

Method: Confidence Intervals are an alternative way to hypothesis tests. We construct a Confidence interval of the odds ratio for the fit model to confirm whether or not the chance of an individual getting Byssinosis is due to whether or not they 1. Smoke or not AND 2. Type of workplace they work in:

Results: We observe that all 3 confidence intervals that correspond to each predictor variable do not include the number 1. We are 90 percent confident that the true odds ratios is not equal to 1 under this model for all predictor variables - there is sufficient evidence to conclude that a difference in whether or not someone smokes and is in different workspace is associated with differences in the chances of having Byssinosis or not because a odds ratio of 1 indicates independence between the response and predictor variables.

```
                   5 %      95 %
(Intercept) 0.07251266 0.1694122
Workspace2  0.06267480 0.1843886
Workspace3  0.03216978 0.0841740
SmokingYes  1.32687904 3.2601728
```

To further confirm our previous results, we obtain the 90 percent confidence intervals for the other 3 remaining predictor variables (Sexm, RaceW, and Employment.) We are able to see that the confidence intervals associated with these 3 predictor variables include 1 (with the exception of Employment >=20), indicating that a difference in the type of race, sex, and employment type of an individual is NOT associated with differences in the chances of having Byssinosis or not. We can further confirm that the best model obtained is the one that includes Smoking Status and type of workspace as the predictor variables.

```
                      5 %      95 %
(Intercept)     0.05151826 0.1798897
SmokingYes      1.34227942 3.3881943
Workspace2      0.05705368 0.1998220
Workspace3      0.02968832 0.0872208
SexM            0.51560573 1.5478422
RaceW           0.47898583 1.2596078
Employment>=20  1.25788516 3.4489279
Employment10-19 0.85254176 2.9729662
```

AIC Akaike Information Criterion) Model Selection:

$$AIC = -2 \times \text{log-likelihood} + 2 \times \text{number of parameters in the model}$$

To reiterate, AIC is another suitable criterion for our investigation as it provides a systematic way to compare and select models that effectively capture the associations between byssinosis and predictor variables, helping to answer the central question regarding the significance of workplace dustiness in the occurrence of byssinosis. AIC tends to select more complex models than BIC, especially when the sample size is large, due to BIC having a larger penalty term for model complexity.

The stepwise logistic regression below is aimed to optimize the model predicting byssinosis (Byss) using Employment, Smoking, Sex, Race, and Workspace. The iterative process, guided by the Akaike Information Criterion (AIC), revealed that removing Race and Sex led to a more efficient model, minimizing AIC while maintaining statistical significance. The final model retained Employment, Smoking, and Workspace, indicating their crucial roles in explaining the variability in Byssinosis.

```
Start:  AIC=737.04
Byss ~ 1

             Df Deviance    AIC
+ Workspace   2   599.26 605.26
+ Sex         1   713.18 717.18
+ Smoking     1   722.48 726.48
+ Race        1   729.49 733.49
<none>            735.04 737.04
+ Employment  2   732.23 738.23
```

```
Step:  AIC=605.26
Byss ~ Workspace

             Df Deviance    AIC
+ Smoking     1   591.36 599.36
+ Employment  2   594.06 604.06
<none>            599.26 605.26
+ Sex         1   598.91 606.91
+ Race        1   599.05 607.05
- Workspace   2   735.04 737.04
```

```
Step:  AIC=599.36
Byss ~ Workspace + Smoking

                    Df Deviance    AIC
+ Employment         2   586.12 598.12
<none>                   591.36 599.36
+ Smoking:Workspace  2   588.07 600.07
+ Race               1   591.14 601.14
+ Sex                1   591.36 601.36
- Smoking            1   599.26 605.26
- Workspace          2   722.48 726.48
```

```
Step:  AIC=598.12
Byss ~ Workspace + Smoking + Employment

                       Df Deviance    AIC
<none>                      586.12 598.12
+ Employment:Workspace  4   578.45 598.45
+ Smoking:Workspace     2   583.07 599.07
- Employment            2   591.36 599.36
+ Race                  1   585.37 599.37
+ Sex                   1   586.00 600.00
+ Employment:Smoking    2   584.55 600.55
- Smoking               1   594.06 604.06
- Workspace             2   719.60 727.60
```

The coefficients seen below in the final model with Employment, Smoking and Workplace, provide insights into the direction and strength of the associations between each predictor variable and the likelihood of byssinosis in the given model. The intercept serves as a baseline, representing the estimated log-odds of byssinosis when all predictor variables are zero. For each unit increase in the "Employment" category 20 years or more and 10-19 , the estimated log-odds of having byssinosis increase by 0.5818 and 0.3551 respectively. This indicates that longer employment durations are associated with an increased likelihood of byssinosis. If a person smokes, the estimated log-odds of having byssinosis increase by 0.7395. If the workplace is in category 2 ("Workspace2"), the estimated log-odds of having byssinosis decrease by 2.2440, and if the workplace is in category 3 ("Workspace3"), the estimated log-odds of having byssinosis decrease by 3.0001. This negative coefficient indicates that being in category 2 or 3 is associated with a further decreased likelihood of byssinosis compared to the baseline, with Workplace 3 having decreased risk for byssinosis.

```
Call:  glm(formula = Byss ~ Workspace + Smoking + Employment, family = binomial,
    data = dataTrain)

Coefficients:
    (Intercept)       Workspace2       Workspace3       SmokingYes   Employment>=20  Employment10-19
        -2.4656          -2.2440          -3.0001           0.7395           0.5818           0.3551

Degrees of Freedom: 2707 Total (i.e. Null);  2702 Residual
  (1 observation deleted due to missingness)
Null Deviance:      735
Residual Deviance: 586.1        AIC: 598.1
```

Comparing to Full Model and Other subset Models:

We can calculate the Akaike Information Criterion (AIC) values for different combinations of the predictor variables, allowing model comparison. Typically the best model is the one with the lowest AICc value, typically the model that best balances goodness of fit and complexity. In case the sample size is small, we are using Akaike Information Criterion corrected for small sample sizes (AICc) to correct/mitigate the risk of overfitting the model. If the sample size is large, the correction term won't have an impact.

Results: The above code performs model comparison based on the AICc. The table below shows subsets of predictor variables, along with their corresponding AICc, delta_AICc (difference from the minimum AICc), AICc weights (probability of being the best model), cumulative weights, and log-likelihood. The models are sorted in ascending order of AICc values, with "Subset Model - Employment, Smoking, Workspace" identified as having the lowest AICc, agreeing with our stepwise logistic regression analysis, and confirming it is indeed the most favorable model when comparing AIC values. In addition, the deviance is another way to measure the model fit from our logistic regression model and prefers a model with more variables, not taking into account whether or not an association was due to randomness and noise or whether it is real. The table below shows that the best fit model according to the AIC corresponds to the smallest Deviance value of 586.1224, further confirming that the best model fit according to AIC is the one that includes type of employment, smoking, and workspace.

Subset Models

	Model	Deviance	AIC
18	Subset Model - Employment, Smoking, Workspace	586.1224	598.1224
12	Subset Model - Smoking, Workspace	591.3641	599.3641
28	Subset Model - Employment, Smoking, Race, Workspace	585.3717	599.3717
27	Subset Model - Employment, Smoking, Sex, Workspace	585.9993	599.9993
24	Subset Model - Smoking, Race, Workspace	591.1425	601.1425
31	Subset Model - Employment, Smoking, Sex, Race, Workspace	585.2584	601.2584
23	Subset Model - Smoking, Sex, Workspace	591.3641	601.3641
30	Subset Model - Smoking, Sex, Race, Workspace	591.1404	603.1404
9	Subset Model - Employment, Workspace	594.0606	604.0606
21	Subset Model - Employment, Race, Workspace	593.2599	605.2599
5	Subset Model - Workspace	599.2648	605.2648
20	Subset Model - Employment, Sex, Workspace	593.9788	605.9788
14	Subset Model - Sex, Workspace	598.9082	606.9082
15	Subset Model - Race, Workspace	599.0460	607.0460
29	Subset Model - Employment, Sex, Race, Workspace	593.1705	607.1705
25	Subset Model - Sex, Race, Workspace	598.7377	608.7377
26	Subset Model - Employment, Smoking, Sex, Race	692.0299	704.0299
19	Subset Model - Employment, Sex, Race	697.7886	707.7886
22	Subset Model - Smoking, Sex, Race	701.1189	709.1189
13	Subset Model - Sex, Race	707.0284	713.0284
10	Subset Model - Smoking, Sex	707.2644	713.2644
17	Subset Model - Employment, Smoking, Race	703.6456	713.6456
16	Subset Model - Employment, Smoking, Sex	705.7514	715.7514
3	Subset Model - Sex	713.1783	717.1783
7	Subset Model - Employment, Sex	711.6361	719.6361
11	Subset Model - Smoking, Race	716.9619	722.9619
8	Subset Model - Employment, Race	715.6353	723.6353
2	Subset Model - Smoking	722.4817	726.4817
6	Subset Model - Employment, Smoking	719.5996	727.5996
4	Subset Model - Race	729.4853	733.4853
1	Subset Model - Employment	732.2315	738.2315

Wald Test on the Model Fit (AIC model):

Method: Using the testing data, we perform Wald Tests to test for evidence of nonzero coefficients, using alpha = 0.05. The null hypothesis of the Wald Test states that the chance of an individual having Byssinosis is not associated with the employment years, type of workspace they work in, and smoking status. It is important to note that when performing multiple hypothesis tests, the overall chances for a type I error between different tests (in this case, we are testing the association between Byssionis and 3 predictor variables) increases. Therefore, we control for multiple testing by applying the Bonferroni Correction.

```
                 Estimate Std. Error   z value     Pr(>|z|)
(Intercept)     -2.4835947  0.2939343 -8.449489 2.925799e-17
Workspace2      -2.9685886  0.4392647 -6.758087 1.398260e-11
Workspace3      -2.5303593  0.2554374 -9.905987 3.920803e-23
SmokingYes       0.5584513  0.2700401  2.068031 3.863712e-02
Employment>=20   0.7526646  0.2600270  2.894564 3.796861e-03
Employment10-19  0.6481212  0.3469683  1.867955 6.176831e-02
```

Results: By extracting the coefficient estimates from the testing set, we are able to see that they differ from those of the model fit we obtained before; such a result is expected as the testing and training sets utilize different observations. With an alpha value of 0.05, we apply the Bonferroni Correction and compare the 5 individual

p-values to 0.05/5= 0.01. Because all the p-values are less than this value (with the exception of employment 10-19), we reject the null hypothesis. We are able to state that we have sufficient evidence to conclude that all 4 out of 5 slopes are nonzero and therefore type of workspace, smoking status, and employment years has an association with the chance of predicting whether or not an individual has Byssinosis. However, it is important to note that the use of the Bonferroni Correction to test 3 Wald Tests can be overly conservative. Therefore, it is sometimes more ideal to utilize the likelihood ratio test (LRT) which is a statistical test that compares the goodness of fit between the null and alternative models which we will later be conducting.

If we re-utilized the training set for testing, we obtain the following coefficient estimates and smaller p-values for all 3 Wald Tests which is the expected observation as we re-used the data:

```
                  Estimate Std. Error    z value     Pr(>|z|)
(Intercept)     -2.4656466  0.2901667  -8.497345 1.939770e-17
Workspace2      -2.2439671  0.3289368  -6.821880 8.985685e-12
Workspace3      -3.0001483  0.2938855 -10.208562 1.815355e-24
SmokingYes       0.7394917  0.2750544   2.688529 7.176757e-03
Employment>=20   0.5818224  0.2558792   2.273816 2.297702e-02
Employment10-19  0.3551095  0.3599216   0.986630 3.238241e-01
```

Using Wald Test to obtain Confidence Intervals (AIC model)

Method: We construct a Confidence interval of the odds ratio for the fit model (using the AIC) to confirm whether or not the chance of an individual getting Byssinosis is due to whether or not they 1. Smoke or not AND 2. Type of workplace they work in AND 3. Employment years.

Results: We observe that all 5 confidence intervals that correspond to each predictor variable do not include the number 1 with the exception of Employment 10-19. We are 90 percent confident that the true odds ratios is not equal to 1 under this model for all predictor variables - there is sufficient evidence to conclude that a difference in whether or not someone smokes, is in different workspace, and how many employment years one has is associated with differences in the chances of having Byssinosis or not because a odds ratio of 1 indicates independence between the response and predictor variables.

```
                       5 %       95 %
(Intercept)     0.05271114 0.13691913
Workspace2      0.06172786 0.18215191
Workspace3      0.03069833 0.08072155
SmokingYes      1.33251417 3.29338472
Employment>=20  1.17461276 2.72564825
Employment10-19 0.78906316 2.57829374
```

To further confirm our previous results, we obtain the 90 percent confidence intervals for the other 2 remaining predictor variables (Sexm and RaceW) We are able to see that the confidence intervals associated with these 2 predictor variables include 1, indicating that a difference in the type of race and sex is NOT associated with

differences in the chances of having Byssinosis or not. We can further confirm that the best model obtained is the one that includes Smoking Status, type of workplace, and employment years.

```
                      5 %      95 %
(Intercept)     0.05151826 0.1798897
SmokingYes      1.34227942 3.3881943
Workspace2      0.05705368 0.1998220
Workspace3      0.02968832 0.0872208
SexM            0.51560573 1.5478422
RaceW           0.47898583 1.2596078
Employment>=20  1.25788516 3.4489279
Employment10-19 0.85254176 2.9729662
```

Likelihood Ratio Tests

The Likelihood Ratio Test (LRT) is a statistical test used for comparing the goodness of fit between two nested models. In the case below, we are using this test to compare a full model with all predictor variables to one without Workspace. From this test, we see that the test statistic is 200.81, with a p-value (Pr(>Chisq)) very close to zero (< 2.2e-16). The small p-value suggests that Model 1 significantly improves the fit compared to Model 2. Therefore, we reject the null hypothesis and conclude that including the "Workspace" variable in Model 1 is justified based on the likelihood ratio test. Workspace dustiness is a large contributor to Byss.

```
Likelihood ratio test

Model 1: Byss ~ Employment + Smoking + Sex + Race + Workspace
Model 2: Byss ~ Employment + Smoking + Sex + Race
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   8 -594.35
2   6 -694.76 -2 200.81  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We use LRT again below to compare two logistic regression models: Model 1, recommended by AIC and featuring the predictors "Employment," "Smoking," and "Workspace," as well as Model 2, recommended by BIC and comprising "Smoking" and "Workspace." Model 1 exhibited a log-likelihood of -594.73, while Model 2 had a log-likelihood of -601.69. The chi-square statistic, calculated as the difference in log-likelihoods, amounted to 13.932, leading to a p-value of 0.0009436. The '***' significance code indicates high statistical significance, suggesting that Model 1 significantly outperforms Model 2 in explaining the variance in the response variable "Byss." This outcome supports the inclusion of "Employment" as a predictor in Model 1, as it contributes significantly to the model's explanatory power compared to the simpler Model 2, as evidenced by the LRT results.

```
Likelihood ratio test

Model 1: Byss ~ Employment + Smoking + Workspace
Model 2: Byss ~ Smoking + Workspace
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   6 -594.73
2   4 -601.69 -2 13.932  0.0009436 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Measuring Deviance

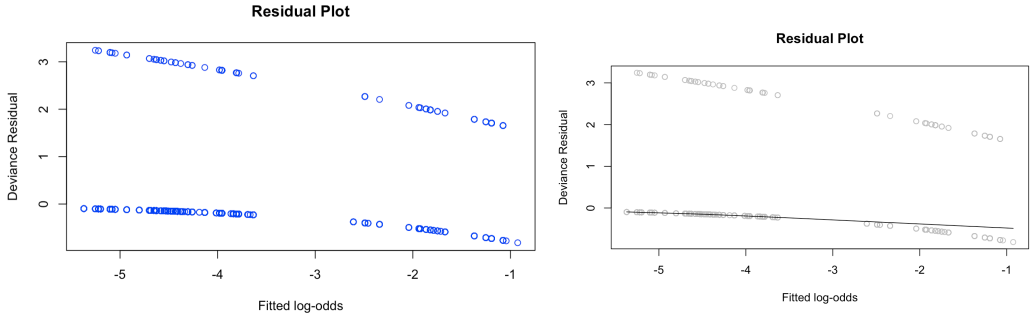
Deviance Calculation:

$$D = 2 \times (\text{LogLikelihood}_{\text{Saturated Model}} - \text{LogLikelihood}_{\text{Fitted Model}})$$

For the full model, including all predictor variables

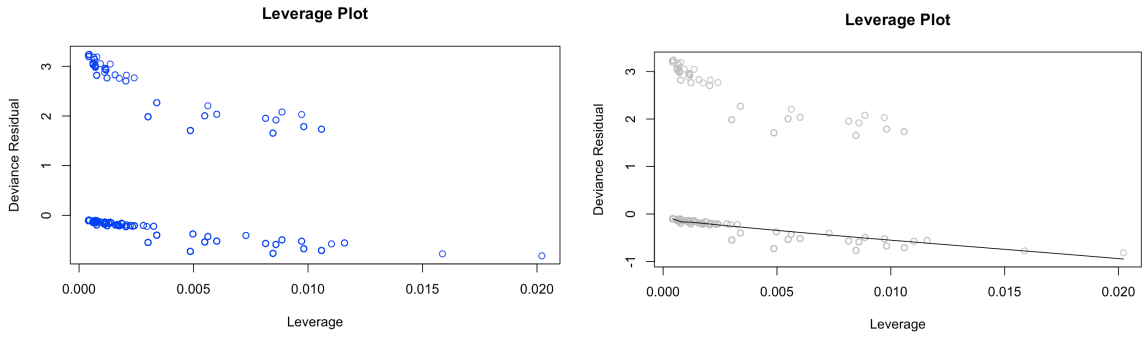
There are three plots we consider when measuring deviance: the residual plot, leverage plot, and influence plot. These are diagnostic plots that can help us interpret patterns and identify potential issues in the logistic regression model. For the residual plot, the deviance residuals are plotted against the fitted log-odds, and is a measure of the lack of fit of the model for each observation. In a well-fitted model, these residuals should be randomly scattered around zero.

Below, is the deviance for the linear regression model with all predictor values: Employment, Smoking, Workspace, Sex and Race. We can see here that there are two lines prominent in the figure, one with deviance residuals around 0, and the other with a downward slope, and deviance residuals around 2-3. There seems to be a pretty constant variability in the spread of residuals, and this Homoscedasticity (constant variance) is preferable, as heteroscedasticity (changing variance) may indicate issues with the model. There doesn't seem to be any real outliers in the spread, which indicates that the model is handling well.

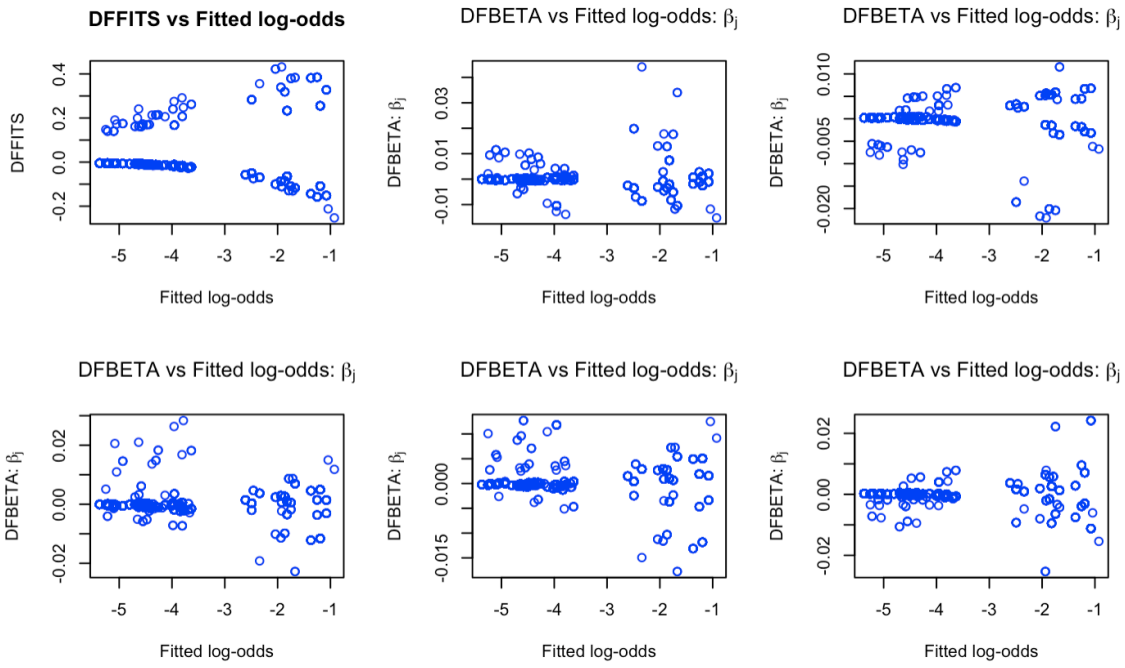


Leverage Plots can provide insight as to the points that have a large influence on the model. Observations with high leverage and high deviance residual are influential points that affect the model's performance, distorting the logistic regression coefficients and compromising the accuracy of predictions. Here, we see that there are few points that have high leverage (greater than 0.01), and we can see that the most of these points have low deviance residuals. This indicates that while these observations exert considerable influence on the logistic regression model due to their high leverage, their deviance residuals are low, suggesting that they align closely with the

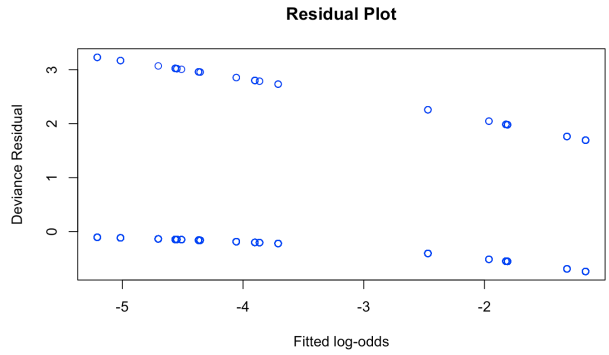
model's predicted values.

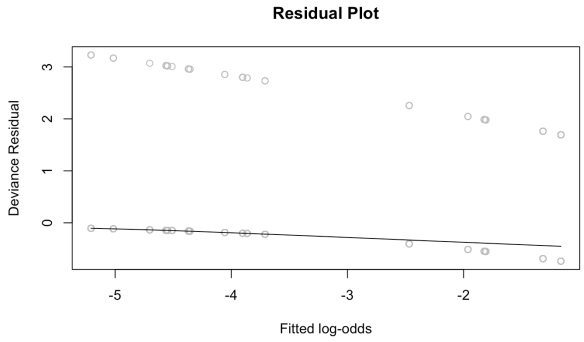


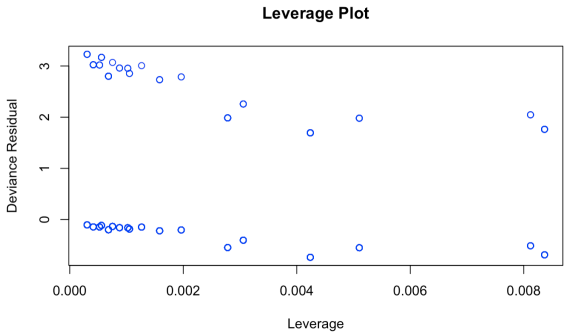
Influence Plots, featuring DFFITS and DFBETA, tells us the impact of observations on a logistic regression model. DFFITS identifies influential points by measuring changes in predicted values upon exclusion, while DFBETA assesses the influence on specific predictors by quantifying changes in regression coefficients. High DFFITS values indicate observations altering overall predictions, and elevated DFBETA values pinpoint influential variables. Here we see that for the DFFITS plot, there are a large number of points clustered around, and no one or two points that have a substantial impact on the predicted values, indicating no singular influential data points that significantly alter the model's fit when omitted. When looking at the DFBETA plot, we see that there are some betas that have more significance than the others, i.e. the top middle, bottom left, and bottom right figures have notably high DFBETA values for specific predictors, identifying variables where the influence is significant. My hypothesis is that these are the values that we are keeping in the model subset that are significant for the AIC or BIC model.

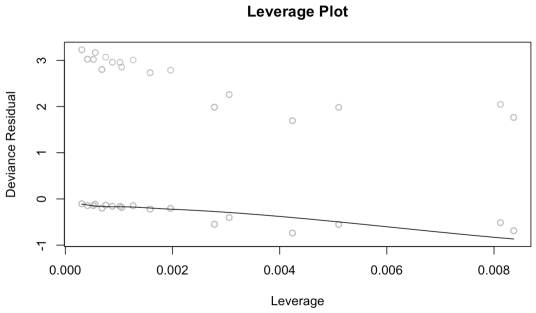


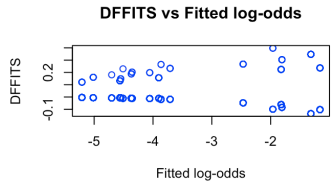
AIC Best Model

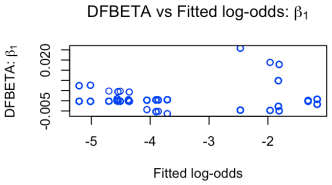


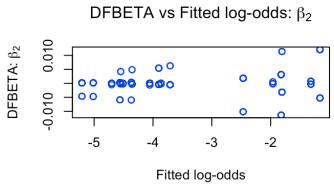


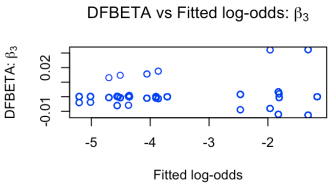


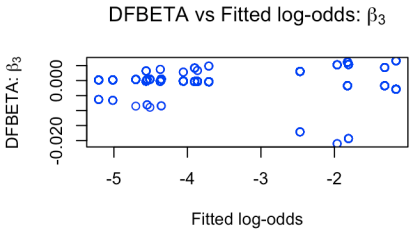






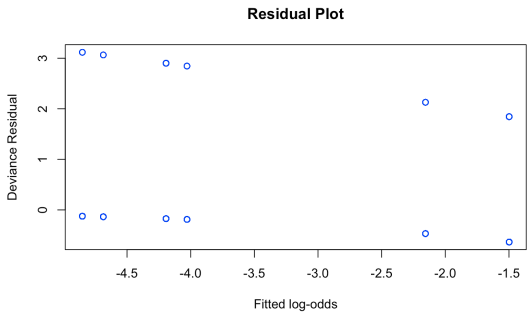


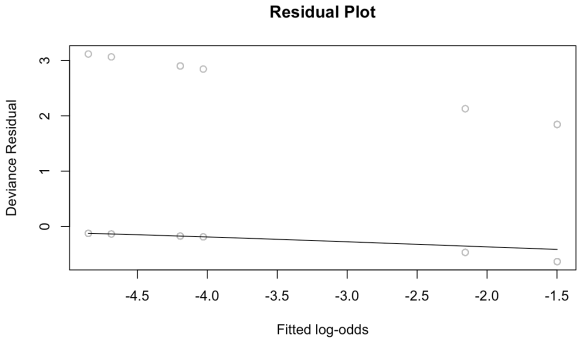


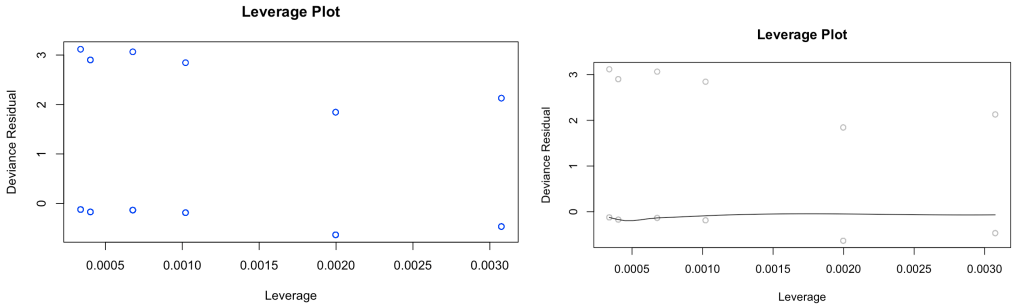


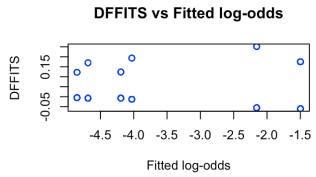
BIC Best Model

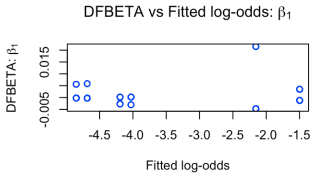
For this residual plot, we see scattering consistent with what we saw previously.

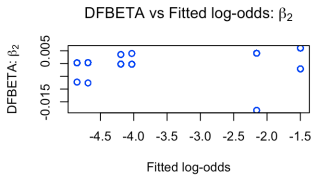






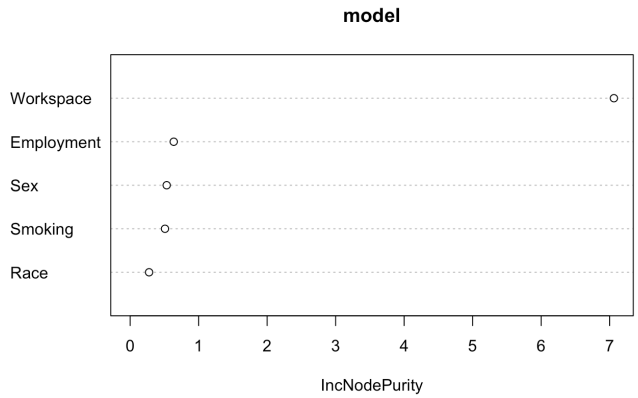






Random Forest: Can we exclude Workspace Dustiness from the Model?

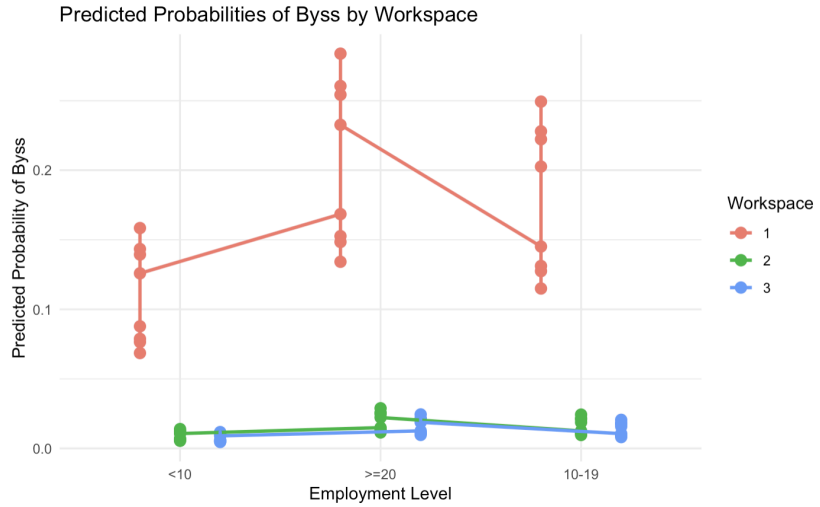
We can use Random Forest to capture any interaction, and to weigh the importance of each predictor variable. This fits a Random Forest model to predict the Byss variable based on the specified predictors. Below is the variable importance, providing insights into which predictors are more influential in predicting Byss. Workspace is the predictor variable with the highest IncNodePurity, indicating that it plays a crucial role in determining the Byss variable in the Random Forest model. This metric reflects the improvement in purity (homogeneity) achieved by considering the Workspace variable in the model, emphasizing its importance in understanding the dynamics of Byssinosis occurrence in the given dataset.



Graph: Predicted Probability of Byss Plotted against Employment Years and Workspace

The graph illustrates a distinct pattern concerning workplace dustiness levels and the likelihood of contracting

Byssinosis ("Byss"). Specifically, for workplaces characterized by 1, there is a substantially elevated probability of experiencing Byssinosis. In contrast, workplaces with 2 and 3 exhibit a comparatively diminished likelihood of encountering Byssinosis, suggesting a lower risk when someone is working in these settings. This aligns with the results we obtained from the model fit (AIC) where the slopes corresponding to both workplaces 2 and 3 were negative.



DISCUSSION

MODEL SELECTION:

We utilized 2 major model selection methods (BIC and AIC) to find the model that best predicts whether or not an individual has Byssinosis according to 5 predictor variables. When using the stepwise selection method for BIC values, the best model fit is characterized by the model that includes both the type of workspace and smoking status as the most influential predictor variables: $\log(pi/1\text{-}pi) = B_0 + B_1x_{Workplace2} + B_2x_{Workplace3} + B_3x_{Smoking}$. The corresponding estimated parameters suggest that the chance of having Byssinosis increases when an individual is a smoker but decreases when an individual is in both workspace 2 and 3 (less dusty and least dusty) To confirm these results, we ran a Wald Test and obtained 90 percent confidence intervals to find that there is high association between these 2 predictor variables with the response variable, additionally applying the Bonferroni Correction to take into account the increased Type I error associated with multiple hypothesis test.

We repeated the bidirectional stepwise logistic regression process guided by AIC, as well as comparing other subset models's AICc values, delta_AICc, AICc weights, cumulative weights, and log-likelihood. According to this model selection, the most influential predictor variables are Workspace, Smoking, and Employment years as characterized by: $\log(pi/1\text{-}pi) = B_0 + B_1x_{Workplace2} + B_2x_{Workplace3} + B_3x_{SmokingYes} + B_4x_{Employment >=20} + B_5x_{Employment\ 10\text{-}19}$. In other words, the model excluding both Race and Sex performed optimally, leading to a more efficient model while maintaining statistical significance.

Compared to the BIC method, the AIC wanted to keep an additional predictor variable (Employment status) in the final model. However, this observation is what is expected as the BIC criterion puts a heavier weight on the complexity of the model. In other words, BIC prefers a more simple model (smaller number of predictors) than AIC as the penalty term

(p*log*(n)) penalizes the model with more parameters. As a result, BIC tends to favor simpler models more strongly than AIC, especially when the sample size is relatively small. It is also important to note that the use of the Bonferroni Correction to test 3 Wald Tests can be overly conservative. Therefore, we utilized the likelihood ratio test (LRT) which is a statistical test that compares the goodness of fit between 1. Full Model with all predictors versus a Reduced model that excludes workspace and 2. Full Model (AIC) with 3 predictor variables and a Reduced Model (BIC) with 2 predictor variables.

When we conducted the likelihood ratio test to compare the BIC model and the AIC model, we found that Full Model (AIC) significantly outperforms Reduced Model (BIC) in explaining the variance in the response variable "Byss" with a chi-square statistic that had a p-value of 0.0009. This outcome supports the inclusion of "Employment Years" as a predictor in the Full Model (AIC), as it contributes significantly to the model's explanatory power compared to the simpler Reduced Model (BIC), as evidenced by the LRT results.

Deviance

We assess Deviance, a key measure in logistic regression through diagnostic plots to evaluate model fit and identify potential outliers or model fit issues. Ideally, the residual plot, contrasting deviance residuals against fitted log-odds, would be a well-fitted model with residuals scattered around zero. In the presented linear regression model encompassing predictors like Employment, Smoking, Workspace, Sex, and Race, the deviance plot displays two distinct lines—one with residuals around 0 and another with a downward slope and residuals around 2-3. This pattern suggests consistent variability, preferable for homoscedasticity, indicating that the model handles well without notable outliers. The Leverage plots identify influential points with high leverage and deviance residuals, influencing logistic regression coefficients, and despite a few high-leverage points, their low deviance residuals imply alignment with the model's predictions. Influence plots, featuring DFFITS and DFBETA, highlight the impact of observations. DFFITS shows no singular influential data points significantly altering predictions, while DFBETA identifies variables with notable influence, contributing to model selection criteria like AIC or BIC. Overall, these diagnostic plots collectively contribute to a comprehensive evaluation of the logistic regression model and its ability to capture associations within the data.

We also observe that the complexity in the models increases as more predictor variables are included. This complexity manifests in the form of more data points in diagnostic plots such as residual plots, leverage plots, and influence plots. The higher number of points may indicate a more nuanced interplay between predictors and the response variable, reflecting the intricate nature of the associations being modeled.

Why we should keep Workspace

In leveraging Random Forest modeling, we aimed to discern interaction effects and evaluate the significance of predictor variables in predicting Byss. The variable importance analysis highlighted Workspace as the most influential predictor, as reflected by its highest IncNodePurity. This underscores Workspace's pivotal role in shaping the outcome of Byssinosis in the dataset.

Confirming these findings, the Likelihood Ratio Test solidifies the significance of Workspace, as Model 1 (workplace dustiness, years of employment, smoking status, gender, and race) with Workspace outperforms Model 2 (years of employment, smoking status, gender, and race). The compelling evidence suggests that Workspace is a crucial contributor to Byssinosis, justifying its retention in the predictive model.

The ggplot graph, illustrating workplace dustiness levels and the likelihood of contracting Byssinosis ("Byss") gives us further insight into the Workspace variable, revealing a pronounced association between workplace dustiness levels and the likelihood of Byssinosis. The Workplace variable has three possibilities: 1,2,3, with level 1 being the most dusty. From this, we can see that Workplaces characterized by level 1 exhibit a substantially higher probability of Byssinosis, while levels 2 and 3 entail a comparatively lower risk. This makes sense intuitively as higher levels of workplace dustiness likely increases exposure to potentially harmful particles. In contrast, the diminished likelihood observed in workplaces with dustiness levels 2 and 3 aligns with the assumption that decreasing dustiness is associated with a lower risk of Byssinosis. The graph not only visually supports these insights but also quantifies the differential probabilities across the three workplace categories, providing a clear and interpretable depiction of the relationship between workplace dustiness and the occurrence of Byssinosis.

Code Appendix

```
n <- nrow(byss)#Importing dataset/converting from wide to long
countColumns <- c(which(names(byss) == "Byssinosis"),
which(names(byss) == "Non.Byssinosis"))
longByss <- rbind(
cbind(Byss=1,
byss[rep(1:n, byss[,"Byssinosis"]), -countColumns]),
cbind(Byss=0,
byss[rep(1:n, byss[,"Non.Byssinosis"]), -countColumns])
)
row.names(longByss) <- 1:nrow(longByss)

longByss$Employment <- as.factor(longByss$Employment)
longByss$Smoking <- as.factor(longByss$Smoking)
longByss$Sex <- as.factor(longByss$Sex)
longByss$Race <- as.factor(longByss$Race)
longByss$Workspace <- as.factor(longByss$Workspace)
```

```
library(car)
vif(fullModel)
```

```
#splitting the data between training and testing sets to avoid overfitting.
set.seed(19203842)
trainIndex <- sample(5419,2709)
dataTrain <- longByss[trainIndex,]
```

```
dataTest <- longByss[-trainIndex,]
```

```
#Bidirectional Step Wise Selection: BIC
result <- step(glm(Byss~1, binomial, dataTrain),
scope = ~Employment*Smoking*Sex*Race*Workspace,
direction = "both",
k = log(800), ## appropriate weight for BIC
trace = 1 ## set to 0 to omit trace
)
result
```

```
#Wald Hypothesis Testing on the BIC model
testFit <- glm (result$model, binomial, dataTest)
summary(testFit)$coefficients
```

```
testFit<-glm(result$model, binomial, dataTrain)
summary(testFit)$coefficients
```

```
#Utilizing Wald Tests to obtain Confidence intervals for BIC model
exp(confint.default(result, level=0.9))
```

```
myGLM1 <- glm(Byss~Smoking+Workspace+Sex+Race+Employment, family=binomial, dataTrain)
exp(confint.default(myGLM1, level=0.9))
```

```
#Stepwise Logistic regression for AIC
responseVar <- "Byss"
predictorVars <- c("Employment", "Smoking", "Sex", "Race", "Workspace")
initial_model <- glm(formula(paste(responseVar, "~", paste(predictorVars, collapse = "+"))),
data = dataTrain, family = binomial)

# Bidirectional stepwise selection with AIC
final_model <- step(initial_model, direction = "both", trace = 1)
summary(final_model)
```

```
#Defining themodel
responseVar <- "Byss"
predictorVars <- c("Employment", "Smoking", "Sex", "Race", "Workspace")
fullModel <- glm(as.formula(paste(responseVar, "~", paste(predictorVars, collapse = "+"))),
                 family = binomial,
                 data = dataTrain)
```

```
#Constructing the big AIC table with the Deviance
# Load the tidyverse package (if not already loaded)
library(tidyverse)

byss <- read.csv("~/Downloads/Byssinosis.csv")
n <- nrow(byss)
countColumns <- c(which(names(byss) == "Byssinosis"),
which(names(byss) == "Non.Byssinosis"))
longByss <- rbind(
cbind(Byss=1,
byss[rep(1:n, byss[,"Byssinosis"]), -countColumns]),
cbind(Byss=0,
byss[rep(1:n, byss[,"Non.Byssinosis"]), -countColumns])
)
row.names(longByss) <- 1:nrow(longByss)

# Assuming 'Employment,' 'Smoking,' 'Sex,' 'Race,' and 'Workspace' are column names in your
dataset
longByss$Employment <- as.factor(longByss$Employment)
longByss$Smoking <- as.factor(longByss$Smoking)
longByss$Sex <- as.factor(longByss$Sex)
longByss$Race <- as.factor(longByss$Race)
longByss$Workspace <- as.factor(longByss$Workspace)

# Assuming 'longByss' is your data frame
responseVar <- "Byss"
predictorVars <- c("Employment", "Smoking", "Sex", "Race", "Workspace")

# Fit the full model
fullModel <- glm(as.formula(paste(responseVar, "~", paste(predictorVars, collapse = "+"))),
                 family = binomial,
                 data = longByss)
allSubsets <- lapply(1:length(predictorVars), function(n) combn(predictorVars, n, simplify =
TRUE))

models <- list() # empty list to store models
mod.names <- character(0)
deviances <- numeric(0)
AIC_values <- numeric(0)

for (i in seq_along(allSubsets)) {
  for (j in seq_along(allSubsets[[i]][1, ])) {
    subsetFormula <- as.formula(paste(responseVar, "~", paste(allSubsets[[i]][, j], collapse
= "+")))
    subsetModel <- glm(subsetFormula,
                       family = binomial,
                       data = longByss)
    models[[paste("Subset Model -", paste(allSubsets[[i]][, j], collapse = ", "))]] <-
subsetModel
    mod.names <- c(mod.names, paste("Subset Model -", paste(allSubsets[[i]][, j], collapse =
", ")))
    deviances <- c(deviances, deviance(subsetModel))
```

```
    AIC_values <- c(AIC_values, AIC(subsetModel))
  }
}

# Create a data frame for the table
table_data <- data.frame(
  Model = mod.names,
  Deviance = deviances,
  AIC = AIC_values
)

table_data <- table_data[order(table_data$AIC), ]
library(knitr)
kable(table_data, format = "html", caption = "Subset Models")
```

```
#Likelihood ratio test comparing AIC model with BIC model
library(lmtest)
fullModel <- glm(Byss ~ Employment + Smoking+ Workspace,
                 family = binomial,
                 data = dataTrain)

reducedModel <- glm(Byss ~ Smoking+Workspace,
                    family = binomial,
                    data = dataTrain)

# likelihood ratio test
lrTest <- lrtest(fullModel, reducedModel)
print(lrTest)

#Likelihood ratio test comparing Full Model with model excluding Workspace

library(lmtest)
fullModel <- glm(Byss ~ Employment + Smoking+ Workspace + Sex+ Race,
                 family = binomial,
                 data = dataTrain)

reducedModel <- glm(Byss ~ Smoking+ Sex+ Race+ Employment,
                    family = binomial,
                    data = dataTrain)

# likelihood ratio test
lrTest <- lrtest(fullModel, reducedModel)
```

```
#Constructing the residual plot, leverage plot, and influence plot
# Assuming 'myGLM' is our logistic regression model

# Residual plots
```

```
ry <- residuals(myGLM, type = "deviance")
rx <- logit(fitted.values(myGLM))
plot(rx, ry,
     xlab = "Fitted log-odds",
     ylab = "Deviance Residual",
     main = "Residual Plot",
     col = 'blue')
scatter.smooth(rx, ry,
               xlab = "Fitted log-odds",
               ylab = "Deviance Residual",
               main = "Residual Plot",
               col = 'gray')
```

```
# Leverage plot
plot(hatvalues(myGLM), ry,
     xlab = "Leverage",
     ylab = "Deviance Residual",
     main = "Leverage Plot",
     col = 'blue')
scatter.smooth(hatvalues(myGLM), ry,
               xlab = "Leverage",
               ylab = "Deviance Residual",
               main = "Leverage Plot",
               col = 'gray')
```

```
# Influence plots
par(mfrow = c(2, 2))
plot(rx, dffits(myGLM),
     xlab = "Fitted log-odds",
     ylab = "DFFITS",
     main = "DFFITS vs Fitted log-odds",
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 1],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[1])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[1])),
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 2],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[2])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[2])),
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 3],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[3])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[3])),
     col = 'blue')


par(mfrow = c(1, 1))
```

```
#Deviance for Full Model
myGLM <- glm(Byss ~ Smoking + Workspace+Employment+Sex+Race,
                  family = binomial,
                  data = longByss)

# Residual plots
ry <- residuals(myGLM, type = "deviance")
rx <- logit(fitted.values(myGLM))
plot(rx, ry,
     xlab = "Fitted log-odds",
     ylab = "Deviance Residual",
     main = "Residual Plot",
     col = 'blue')
scatter.smooth(rx, ry,
               xlab = "Fitted log-odds",
               ylab = "Deviance Residual",
               main = "Residual Plot",
               col = 'gray')

# Leverage plot
plot(hatvalues(myGLM), ry,
     xlab = "Leverage",
     ylab = "Deviance Residual",
     main = "Leverage Plot",
     col = 'blue')
scatter.smooth(hatvalues(myGLM), ry,
               xlab = "Leverage",
               ylab = "Deviance Residual",
               main = "Leverage Plot",
               col = 'gray')

par(mfrow = c(2, 3))

plot(rx, dffits(myGLM),
     xlab = "Fitted log-odds",
     ylab = "DFFITS",
     main = "DFFITS vs Fitted log-odds",
     col = 'blue')

for (j in 1:6) {
  plot(rx, dfbeta(myGLM)[, j]
,
       xlab = "Fitted log-odds",
       ylab = expression(paste("DFBETA: ", beta[j])),
       main = expression(paste("DFBETA vs Fitted log-odds: ", beta[j])),
       col = 'blue')
}

par(mfrow = c(1, 1))
```

```
#Deviance for AIC preferred Model
myGLM <- glm(Byss ~ Employment + Smoking + Workspace,
                  family = binomial,
                  data = longByss)

# Residual plots
ry <- residuals(myGLM, type = "deviance")
rx <- logit(fitted.values(myGLM))
plot(rx, ry,
     xlab = "Fitted log-odds",
     ylab = "Deviance Residual",
     main = "Residual Plot",
     col = 'blue')
scatter.smooth(rx, ry,
               xlab = "Fitted log-odds",
               ylab = "Deviance Residual",
               main = "Residual Plot",
               col = 'gray')

# Leverage plot
plot(hatvalues(myGLM), ry,
     xlab = "Leverage",
     ylab = "Deviance Residual",
     main = "Leverage Plot",
     col = 'blue')
scatter.smooth(hatvalues(myGLM), ry,
               xlab = "Leverage",
               ylab = "Deviance Residual",
               main = "Leverage Plot",
               col = 'gray')

# Influence plots
par(mfrow = c(2, 2))
plot(rx, dffits(myGLM),
     xlab = "Fitted log-odds",
     ylab = "DFFITS",
     main = "DFFITS vs Fitted log-odds",
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 1],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[1])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[1])),
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 2],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[2])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[2])),
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 3],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[3])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[3])),
```

```
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 4],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[3])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[3])),
     col = 'blue')

par(mfrow = c(1, 1))
```

```
#Deviance for BIC Preferred Model
myGLM <- glm(Byss ~ Smoking + Workspace,
                  family = binomial,
                  data = longByss)

# Residual plots
ry <- residuals(myGLM, type = "deviance")
rx <- logit(fitted.values(myGLM))
plot(rx, ry,
     xlab = "Fitted log-odds",
     ylab = "Deviance Residual",
     main = "Residual Plot",
     col = 'blue')
scatter.smooth(rx, ry,
               xlab = "Fitted log-odds",
               ylab = "Deviance Residual",
               main = "Residual Plot",
               col = 'gray')

# Leverage plot
plot(hatvalues(myGLM), ry,
     xlab = "Leverage",
     ylab = "Deviance Residual",
     main = "Leverage Plot",
     col = 'blue')
scatter.smooth(hatvalues(myGLM), ry,
               xlab = "Leverage",
               ylab = "Deviance Residual",
               main = "Leverage Plot",
               col = 'gray')

# Influence plots
par(mfrow = c(2, 2))
plot(rx, dffits(myGLM),
     xlab = "Fitted log-odds",
     ylab = "DFFITS",
     main = "DFFITS vs Fitted log-odds",
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 1],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[1])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[1])),
```

```
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 2],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[2])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[2])),
     col = 'blue')
plot(rx, dfbeta(myGLM)[, 3],
     xlab = "Fitted log-odds",
     ylab = expression(paste("DFBETA: ", beta[3])),
     main = expression(paste("DFBETA vs Fitted log-odds: ", beta[3])),
     col = 'blue')


par(mfrow = c(1, 1))
```

```
#Random Forest
library(randomForest)
model <- randomForest(Byss ~ Employment + Smoking + Sex + Race + Workspace, data =
dataTrain)
importance <- importance(model)
varImpPlot(model)
```

```
#Likelihood Ratio Test: Full Model versus Model excluding Workspace
library(lmtest)
fullModel <- glm(Byss ~ Employment + Smoking + Sex + Race + Workspace,
                 family = binomial,
                 data = longByss)

reducedModel <- glm(Byss ~ Employment + Smoking + Sex + Race,
                    family = binomial,
                    data = longByss)

# likelihood ratio test
lrTest <- lrtest(fullModel, reducedModel)
print(lrTest)
```

```
#Graph for different levels of workspace dustiness and effects on Byss
library(ggplot2)

# Create a data frame for prediction
prediction_data <- expand.grid(Employment = levels(longByss$Employment),
                               Smoking = levels(longByss$Smoking),
                               Sex = levels(longByss$Sex),
                               Race = levels(longByss$Race),
                               Workspace = factor(1:3))

# Predict probabilities using the fitted model
prediction_data$Prob_Byss <- predict(fullModel, newdata = prediction_data, type =
```

```
"response")

# Plot predicted probabilities for different levels of Workspace
ggplot(prediction_data, aes(x = Employment, y = Prob_Byss, color = Workspace)) +
  geom_point(position = position_dodge(width = 0.6), size = 3) +
  geom_line(aes(group = Workspace), position = position_dodge(width = 0.6), size = 1) +
  labs(title = "Predicted Probabilities of Byss by Workspace",
       x = "Employment Level", y = "Predicted Probability of Byss",
       color = "Workspace") +
  theme_minimal()
```