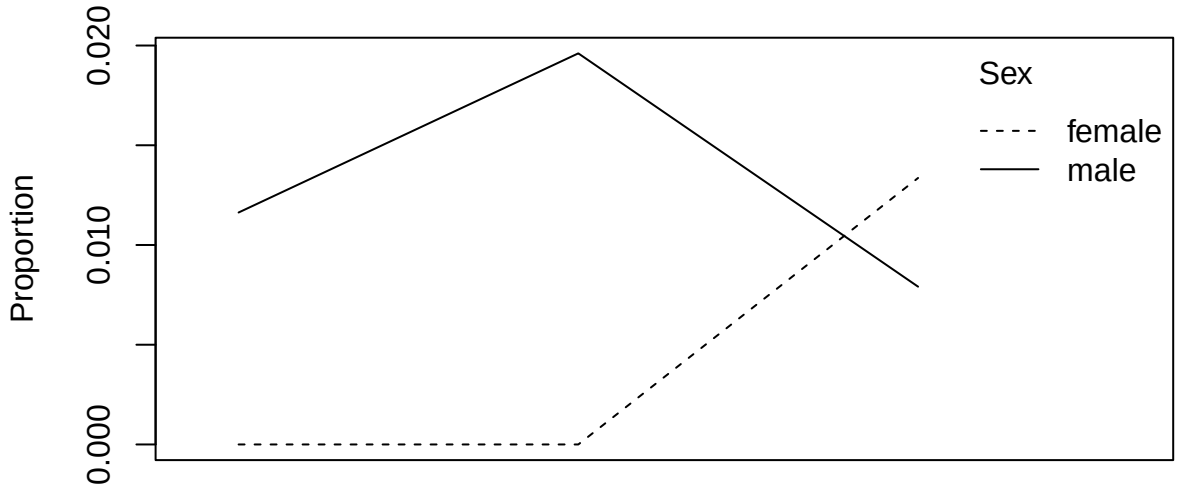
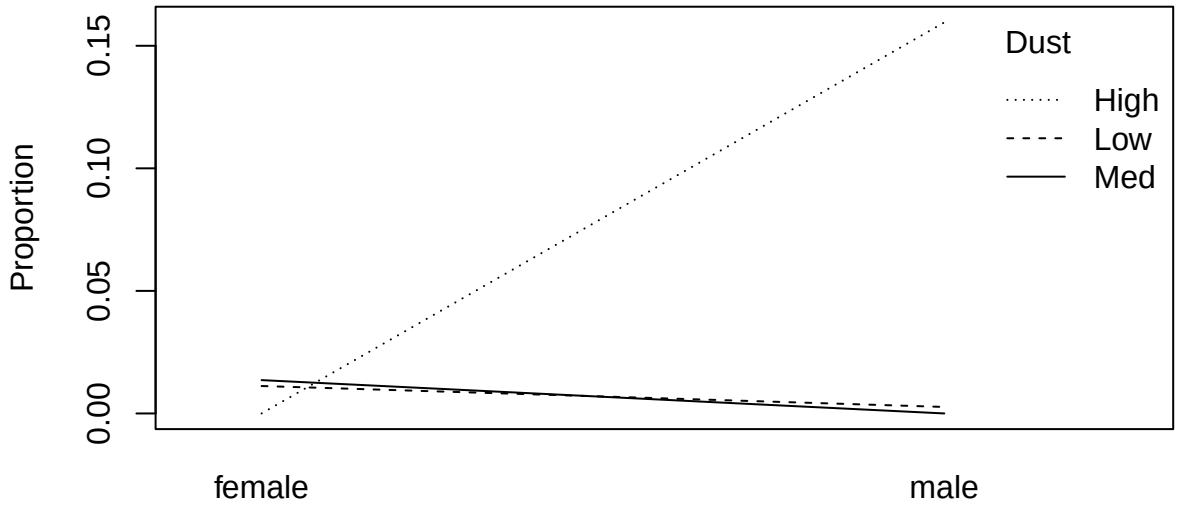
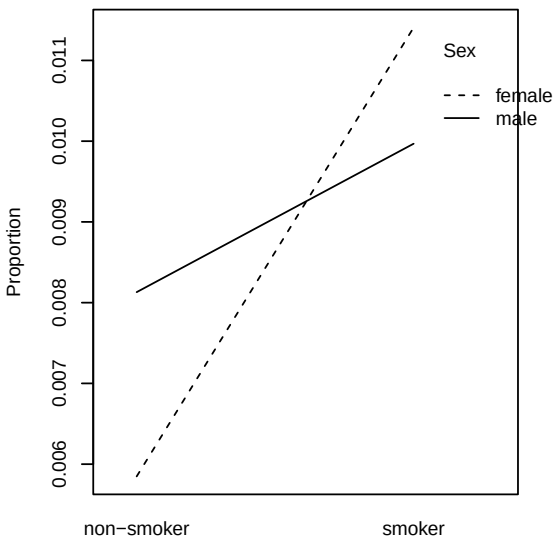
Assignment 2

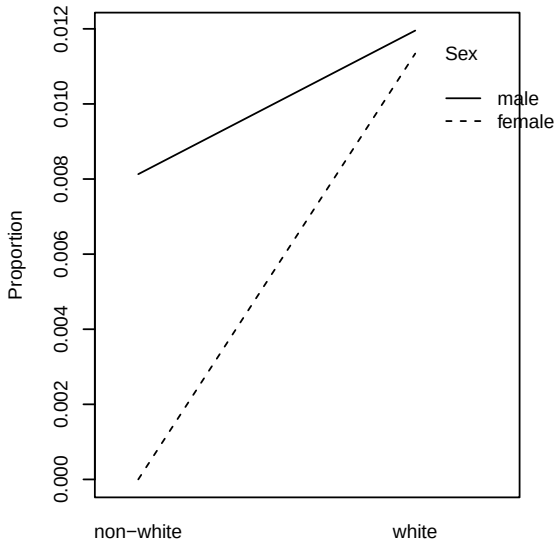
1.

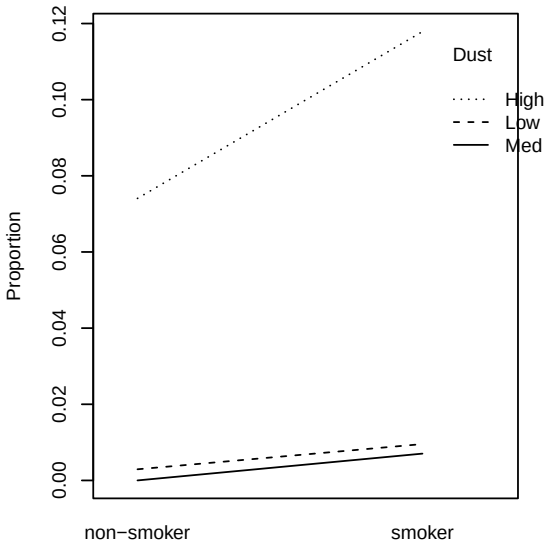
First, we look at some interaction plots to get an idea of the level of interactions in the data set.

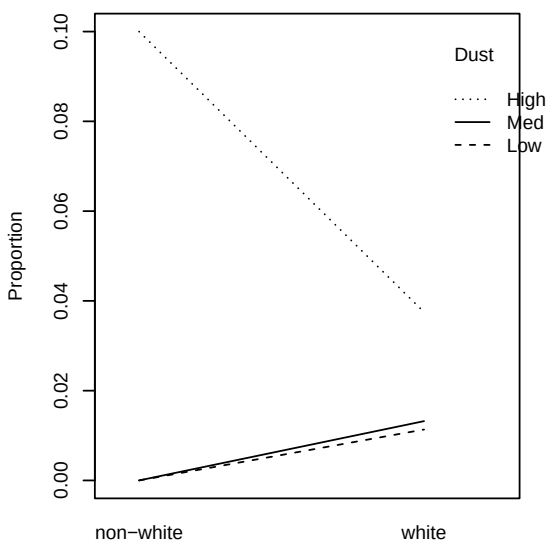


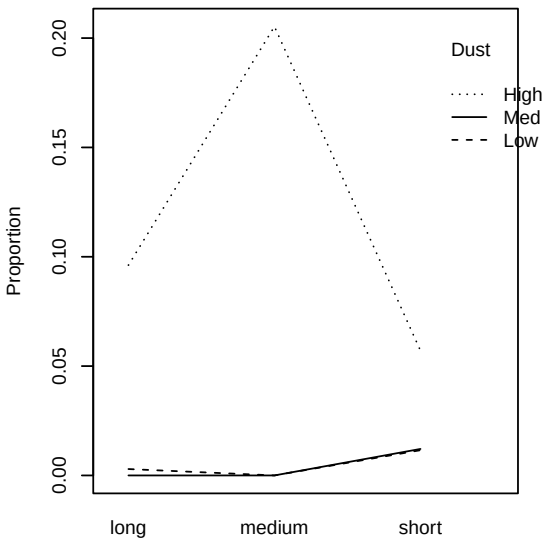


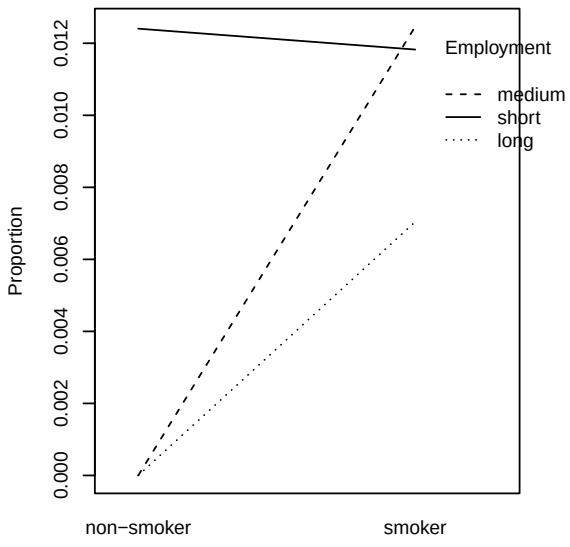

Sex
female
male
Proportion
0.006
0.007
0.008
0.009
0.010
0.011
non−smoker
smoker

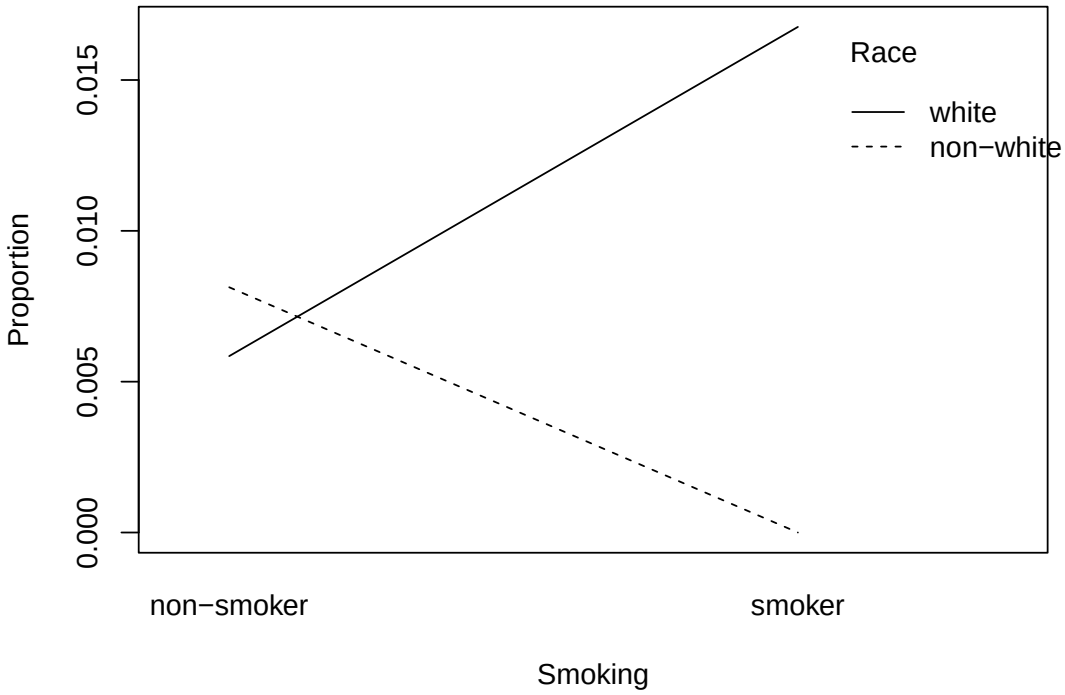

Sex
male
female
Proportion
0.000
0.002
0.004
0.006
0.008
0.010
0.012
non−white
white

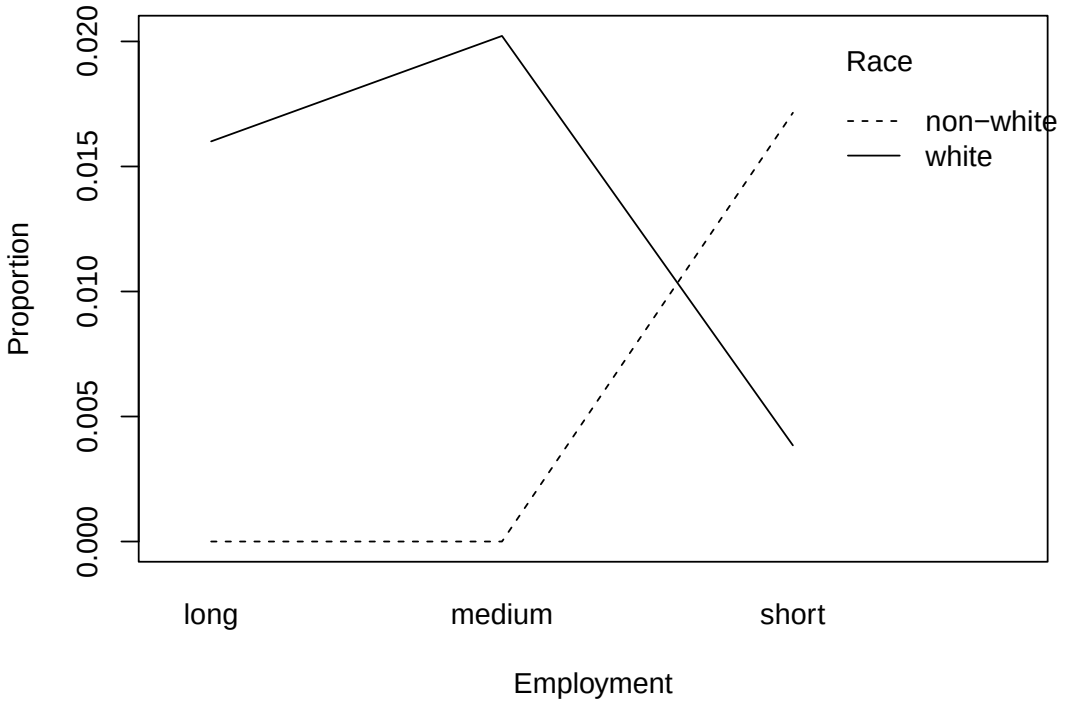

Dust
High
Low
Med
Proportion
0.00
0.02
0.04
0.06
0.08
0.10
0.12
non−smoker
smoker


Dust
High
Med
Low
Proportion
0.00
0.02
0.04
0.06
0.08
0.10
non−white
white


Dust
High
Med
Low
Proportion
0.00
0.05
0.10
0.15
0.20
long
medium
short


Employment
medium
short
long
Proportion
0.000
0.002
0.004
0.006
0.008
0.010
0.012
non−smoker
smoker


Race
white
non−white
Proportion
0.000
0.005
0.010
0.015
non−smoker
smoker
Smoking


Race
non−white
white
Proportion
0.000
0.005
0.010
0.015
0.020
long
medium
short
Employment

The interaction plots show the median of the proportions of people who develop byssinosis for that combination of predictors. Since many lines are not parallel, we can see that there may be some interaction effects. For example, smoking seemingly increases the probability of developing byssinosis for white people, but decreases the probability for non-white people. Another example is that smoking seems to be associated with an increase in the risk of developing byssinosis, but the increase is a lot higher for females than males. We should not infer too much from these interaction plots because covariate classes vary in size, they are just a visual aid to judge if there might be any interaction effects.

We start with the most basic model:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \sum_j \beta_j x_{ij},$$

where p_i denotes the probability of developing byssinosis for the i^{th} covariate class, x_{ij} is the j^{th} covariate of the i^{th} covariate class. In this context, the covariates include dust, race, sex, smoke, emp, **and** all their interaction terms. Treatment coding is used to code all variables.

```
##
## Call:
## glm(formula = cbind(yes, no) ~ .^2, family = binomial, data = byssinosis)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.11620    1.19834  -3.435 0.000593 ***
## dustLow                -1.95791    1.13856  -1.720 0.085498 .
## dustMed                -0.85278    1.43396  -0.595 0.552040
## sexmale                 2.26285    1.08732   2.081 0.037423 *
## smokesmoker             0.99892    0.79452   1.257 0.208660
## empmedium               0.52792    1.10052   0.480 0.631443
## empshort                0.68533    0.95520   0.717 0.473084
## racewhite               0.15085    0.99619   0.151 0.879639
## dustLow:sexmale        -1.23599    0.81086  -1.524 0.127433
## dustMed:sexmale        -1.99044    0.96614  -2.060 0.039381 *
## dustLow:smokesmoker    -0.77088    0.56463  -1.365 0.172163
## dustMed:smokesmoker    -1.39346    0.72649  -1.918 0.055101 .
## dustLow:empmedium      -0.98898    0.78700  -1.257 0.208882
## dustMed:empmedium      -0.10910    1.03360  -0.106 0.915937
## dustLow:empshort        0.94657    0.79859   1.185 0.235901
## dustMed:empshort        0.46424    1.14065   0.407 0.684009
## dustLow:racewhite       1.07680    0.81491   1.321 0.186375
## dustMed:racewhite       0.68166    1.14303   0.596 0.550932
## sexmale:smokesmoker    -0.26956    0.57189  -0.471 0.637394
## sexmale:empmedium      -0.52691    0.91296  -0.577 0.563841
## sexmale:empshort       -1.20455    0.79240  -1.520 0.128479
## sexmale:racewhite      -0.55447    0.82603  -0.671 0.502060
## smokesmoker:empmedium  -0.28293    0.63255  -0.447 0.654665
## smokesmoker:empshort   -0.06715    0.56769  -0.118 0.905844
## smokesmoker:racewhite   0.63627    0.55035   1.156 0.247628
## empmedium:racewhite     0.35292    0.63105   0.559 0.575987
## empshort:racewhite     -1.50410    0.71156  -2.114 0.034532 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 322.527  on 64  degrees of freedom
## Residual deviance:  15.667  on 38  degrees of freedom
## AIC: 176.34
##
## Number of Fisher Scoring iterations: 6
```

The residual deviance of the model is 15.667 on 38 degrees of freedom, the p-value is 0.9994913, this indicates a good fit, however, this does not mean it is a good model, it just fits the observations well. We can see that there are many predictors with insignificant (Wald test) p-values, this model is not parsimonious enough.

Since we want a parsimonious model, we start by using backward selection based on the AIC:

$$AIC = -2\log(L) + 2q,$$

any constant terms in the definition of log-likelihood can be ignored when comparing different models that will have the same constants. For this reason the AIC is equivalent to Deviance + $2q$ (q is the number of parameters). We drop the predictor that would decrease AIC the most until the AIC doesn't decrease anymore.

```
##
## Call:
## glm(formula = cbind(yes, no) ~ dust + sex + smoke + emp + race +
##     dust:sex + dust:smoke + sex:emp + sex:race + smoke:race +
##     emp:race, family = binomial, data = byssinosis)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.44400    1.00693  -4.413 1.02e-05 ***
## dustLow               -1.04632    0.70695  -1.480  0.13886
## dustMed               -0.38676    0.72797  -0.531  0.59522
## sexmale                2.56750    0.93469   2.747  0.00602 **
## smokesmoker            0.68495    0.32294   2.121  0.03392 *
## empmedium              0.03367    0.82742   0.041  0.96754
## empshort               1.07925    0.76121   1.418  0.15625
## racewhite              0.74496    0.77433   0.962  0.33601
## dustLow:sexmale       -1.36318    0.71152  -1.916  0.05538 .
## dustMed:sexmale       -1.97738    0.86195  -2.294  0.02179 *
## dustLow:smokesmoker   -0.60227    0.45943  -1.311  0.18989
## dustMed:smokesmoker   -1.18126    0.57295  -2.062  0.03924 *
## sexmale:empmedium     -0.27093    0.68070  -0.398  0.69062
## sexmale:empshort      -1.59038    0.67900  -2.342  0.01917 *
## sexmale:racewhite     -1.00659    0.67441  -1.493  0.13555
## smokesmoker:racewhite  0.59445    0.41176   1.444  0.14883
## empmedium:racewhite    0.07658    0.58649   0.131  0.89612
## empshort:racewhite    -1.15265    0.59713  -1.930  0.05357 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 322.527  on 64  degrees of freedom
## Residual deviance:  21.307  on 47  degrees of freedom
## AIC: 163.98
```

```
##
## Number of Fisher Scoring iterations: 5
```

Using the AIC to sequentially drop predictors gives the model:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1\text{LowDust} + \beta_2\text{MediumDust} + \beta_3\text{Male} + \beta_4\text{Smoking} + \beta_5\text{EmployMedium},$$
$$+\beta_6\text{EmployShort} + \beta_7\text{White} + \beta_8(\text{LowDust} \cdot \text{Male}) + \beta_9(\text{MediumDust} \cdot \text{Male}) + \beta_{10}(\text{LowDust} \cdot \text{Smoke})$$
$$+\beta_{11}(\text{MediumDust} \cdot \text{Smoke}) + \beta_{12}(\text{Male} \cdot \text{EmployMedium}) + \beta_{13}(\text{Male} \cdot \text{EmployShort}) + \beta_{14}(\text{Male} \cdot \text{White})$$
$$\beta_{15}(\text{Smoke} \cdot \text{White}) + \beta_{16}(\text{EmployMedium} \cdot \text{White}) + \beta_{17}(\text{EmployShort} \cdot \text{White})$$

This model's deviance has p-value 0.9995449, but it is still too large with too many non-significant predictors. Therefore, we now proceed to drop the least significant predictor one by one, using the p-value of the difference-in-deviance test as the criterion, i.e., if the increase in deviance from a model without a predictor is insignificant, we drop that predictor. We continue doing this until all predictors are significant at the 0.05 level.

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + race + dust:sex +
##     dust:smoke + sex:emp + sex:race + smoke:race + emp:race
##            Df Deviance    AIC    LRT Pr(>Chi)
## <none>          21.307 163.98
## dust:sex    2   28.178 166.85 6.8708  0.03221 *
## dust:smoke  2   26.061 164.74 4.7536  0.09285 .
## sex:emp     2   27.187 165.86 5.8801  0.05286 .
## sex:race    1   23.632 164.31 2.3250  0.12731
## smoke:race  1   23.408 164.08 2.1008  0.14723
## emp:race    2   26.513 165.19 5.2061  0.07405 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We drop Smoke · Race, refit the model with the remaining predictors, then look for the next predictor to drop.

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + race + dust:sex +
##     dust:smoke + sex:emp + sex:race + emp:race
##            Df Deviance    AIC    LRT Pr(>Chi)
## <none>          23.408 164.08
## dust:sex    2   30.834 167.51 7.4260  0.02440 *
## dust:smoke  2   27.240 163.92 3.8322  0.14718
## sex:emp     2   29.567 166.24 6.1591  0.04598 *
## sex:race    1   25.131 163.81 1.7225  0.18938
## emp:race    2   28.644 165.32 5.2363  0.07294 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We drop Male · Race, refit the model with the remaining predictors, then look for the next predictor to drop.

```
mdl3 <- glm(cbind(yes,no) ~ dust + sex + smoke + emp + race + dust:sex + dust:smoke + sex:emp + emp:race
drop1(mdl3,test="Chi")
```

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + race + dust:sex +
##     dust:smoke + sex:emp + emp:race
##            Df Deviance    AIC    LRT Pr(>Chi)
## <none>          25.131 163.81
## dust:sex    2   33.856 168.53 8.7258  0.01274 *
## dust:smoke  2   28.789 163.47 3.6584  0.16054
## sex:emp     2   30.320 165.00 5.1894  0.07467 .
## emp:race    2   28.665 163.34 3.5345  0.17081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We drop Employment · Race, refit the model with the remaining predictors, then look for the next predictor to drop.

```
mdl4 <- glm(cbind(yes,no) ~ dust + sex + smoke + emp + race + dust:sex + dust:smoke + sex:emp, family=bi
drop1(mdl4,test="Chi")
```

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + race + dust:sex +
##     dust:smoke + sex:emp
##            Df Deviance    AIC    LRT Pr(>Chi)
## <none>          28.665 163.34
## race        1   28.810 161.49 0.1452  0.70318
## dust:sex    2   36.475 167.15 7.8097  0.02014 *
## dust:smoke  2   32.311 162.99 3.6463  0.16152
## sex:emp     2   32.807 163.48 4.1416  0.12608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We drop Race, refit the model with the remaining predictors, then look for the next predictor to drop.

```
mdl5 <- glm(cbind(yes,no) ~ dust + sex + smoke + emp + dust:sex + dust:smoke + sex:emp, family=binomial
drop1(mdl5,test="Chi")
```

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + dust:sex + dust:smoke +
##     sex:emp
##            Df Deviance    AIC    LRT Pr(>Chi)
## <none>          28.810 161.49
## dust:sex    2   36.676 165.35 7.8659  0.01959 *
```

```
## dust:smoke  2   32.493 161.17 3.6826  0.15861
## sex:emp     2   33.076 161.75 4.2655  0.11851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We drop Dust · Smoke, refit the model with the remaining predictors, then look for the next predictor to drop.

```
mdl6 <- glm(cbind(yes,no) ~ dust + sex + smoke + emp + dust:sex + sex:emp, family=binomial,byssinosis)
drop1(mdl6,test="Chi")
```

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + dust:sex + sex:emp
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       32.493 161.17
## smoke     1   43.736 170.41 11.2433 0.0007991 ***
## dust:sex  2   42.119 166.79  9.6267 0.0081205 **
## sex:emp   2   36.019 160.69  3.5260 0.1715282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We drop Employment · Male, refit the model with the remaining predictors, then look for the next predictor to drop.

```
mdl7 <- glm(cbind(yes,no) ~ dust + sex + smoke + emp + dust:sex, family=binomial,byssinosis)
drop1(mdl7,test="Chi")
```

```
## Single term deletions
##
## Model:
## cbind(yes, no) ~ dust + sex + smoke + emp + dust:sex
##          Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       36.019 160.69
## smoke     1   48.325 171.00 12.3065 0.0004514 ***
## emp       2   48.578 169.25 12.5588 0.0018746 **
## dust:sex  2   43.586 164.26  7.5676 0.0227363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the p-value of the difference-in-deviance test, we should not drop any other predictors, hence our final model is:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{LowDust} + \beta_2 \text{MediumDust} + \beta_3 \text{Male} + \beta_4 \text{Smoking} + \beta_5 \text{EmployMedium},$$
$$+\beta_6 \text{EmployShort} + \beta_7 (\text{LowDust} \cdot \text{Male}) + \beta_8 (\text{MediumDust} \cdot \text{Male})$$

2.

Our final model has the following summary:

```
summary(mdl7)
```

```
##
## Call:
## glm(formula = cbind(yes, no) ~ dust + sex + smoke + emp + dust:sex,
##     family = binomial, data = byssinosis)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.7730     0.6272  -4.422  9.8e-06 ***
## dustLow          -1.6207     0.6401  -2.532 0.011347 *
## dustMed          -1.2714     0.6571  -1.935 0.053020 .
## sexmale           0.9990     0.6106   1.636 0.101816
## smokesmoker       0.6578     0.1945   3.381 0.000722 ***
## empmedium        -0.1727     0.2458  -0.703 0.482352
## empshort         -0.6367     0.1837  -3.466 0.000528 ***
## dustLow:sexmale  -1.2576     0.6823  -1.843 0.065303 .
## dustMed:sexmale  -2.0058     0.8336  -2.406 0.016122 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 322.527  on 64  degrees of freedom
## Residual deviance:  36.019  on 56  degrees of freedom
## AIC: 160.69
##
## Number of Fisher Scoring iterations: 5
```

The residual deviance is 36.019 on 56 degrees of freedom, the p-value for this is 0.9825784, indicating good fit.

The Wald test for is significant for the main effects dustLow (dummy variable), smoking, and empshort and the interaction dustMed:sexmale. The Wald test is not significant for dustMed, sexmale, empmedium and the interaction dustLow:sexmale. However, we do not remove non-significant predictors because it is part of a categorical variable (e.g. empmedium) or because of the hierarchy principle (e.g. sexmale). An even more parsimonious model could be obtained by investigating whether some categories of employment length (emp) could be grouped together.

Our model fits well in terms of residual deviance, it contains all significant predictors (if we consider categorical variables as a whole and not just as separate dummy variables), and does not contain any predictors that isn't significant or is the main effect of a significant interaction (or is part of a categorical predictor that is significant). We cannot drop any other predictors because all candidate predictors are statistically significant. Thus, this model is suitable.

3.

All interpretations are associative and strictly non-causal.

Smokers are more likely to develop byssinosis than non-smokers, all else remaining equal, smokers have 0.6578 greater log-odds of developing byssinosis, this translates to 1.93054 times the odds of a non-smoker.

Compared to workers who has worked more than 20 years, workers who has worked for 10-20 years have 0.1727 smaller log-odds of developing byssinosis, this translates to 0.84139 times the odds of a worker who has worked for more than 20 years; workers who has worked for less than 10 years have 0.6367 smaller log-odds of developing byssinosis, this translates to 0.5290354 times the odds of a worker who has worked for more than 20 years. However, the association for workers who has worked for 10-20 years may not be reliable because the standard error for that predictor is relatively high (the p-value is 0.48).

All else remaining equal, men are more likely to develop byssinosis than females, the increase in log-odds is 0.9990 (2.715565 times the odds), however, the standard deviation for this predictor is also quite high, therefore, we should not read too much into the parameter for this predictor.

For females, being exposed to high dust in the work place is associated with a 1.6207 increase in the log-odds of developing byssinosis, compared to being exposed to low dust in the work place (the odds goes up 5.056629 times). Being exposed to a medium amount of dust in the work place is associated with a 0.3493 (1.6207-1.2714) increase in the log-odds of developing byssinosis (1.418075 times the odds), compared to being exposed to low dust in the work place.

For males, being exposed to high dust in the work place is associated with an additional 1.2576 (so the total increase in log-odds is 2.8783) increase in the log-odds of developing byssinosis (17.78401 times the odds), compared to being exposed to low dust in the work place. Being exposed to medium dust in the work place is associated with a decrease of 0.3989 ((-1.6207-1.2576)-(-1.2714-2.0058)) in the log-odds of developing byssinosis, compared to being exposed to low dust in the work place (the odds is 0.6710578 times of being exposed to low dust).

However, the p-values for dustLow:sexmale and dustMed are greater than 0.05, therefore, interpretations involving men being exposed to a medium amount of dust in the work place may not be very reliable.

4.

All else remaining equal, our model implies that smoking is associated with an increase in the log-odds of developing byssinosis of 0.6578. The standard error for the parameter of this predictor is 0.1945, so the 95% confidence interval for the increase in the log-odds is:

$$[0.6578 - 1.96 \times 0.1945, 0.6578 - 1.96 \times 0.1945] = [0.2764795, 1.039069],$$

the 95/ confidence interval for the odds can be obtained from exponentiating the end points of this confidence interval, giving:

$$[1.31848, 2.826584].$$

($0.975 = \Phi(1.96)$, $\hat{\boldsymbol{\beta}}$ is asymptotically normal.)

5.

One possible way to view our logistic regression model in terms of a latent continuous variable T is to let T have the standard logistic distribution with cumulative density function:

$$F_T(t) = \frac{1}{1 + \exp(-t)},$$

and for a worker with covariates $\mathbf{x}$, the worker suffers from byssinosis if

$$T < \beta_0 + \beta_1 \text{LowDust} + \beta_2 \text{MediumDust} + \beta_3 \text{Male} + \beta_4 \text{Smoking} + \beta_5 \text{EmployMedium},$$
$$+\beta_6 \text{EmployShort} + \beta_7 (\text{LowDust} \cdot \text{Male}) + \beta_8 (\text{MediumDust} \cdot \text{Male})$$

for some unknown parameter $\boldsymbol{\beta}$.

Then the probability of getting byssinosis is:

$$p_{\mathbf{x}} = P(T < \mathbf{x}^T \boldsymbol{\beta}) = F_T(\boldsymbol{\beta}) \iff$$

$$F_T^{-1}(p_{\mathbf{x}}) = \mathbf{x}^T \boldsymbol{\beta} \iff$$

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{LowDust} + \beta_2 \text{MediumDust} + \beta_3 \text{Male} + \beta_4 \text{Smoking} + \beta_5 \text{EmployMedium},$$
$$+\beta_6 \text{EmployShort} + \beta_7 (\text{LowDust} \cdot \text{Male}) + \beta_8 (\text{MediumDust} \cdot \text{Male}),$$

which is equivalent to our final model.