Department of Statistics

COURSE STATS 330/762

Model answer for Assignment 4, 2012

Task 1. *Read in the data and make a data frame. Do the usual checks for typographical errors. Print out the first 10 lines of the data file. [5 marks]*

There is an error in the data – the very last line in the data file contained the value 2400 (should have been 24.0)

2 marks for reading, 2 marks for detecting the mistake, 1 mark for printing.

Task 2. Draw some plots that will shed light on the relationships (if any) between the response and the explanatory variables. Comment on what you find. [10 marks]

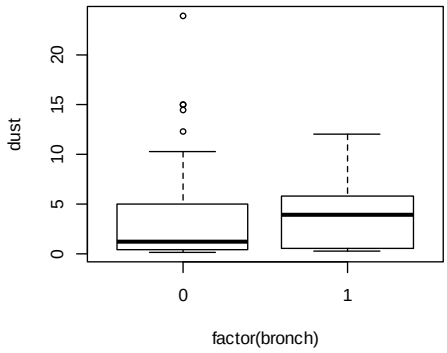
We can draw plots of the response versus the explanatory variables, and a scatterplot of dust versus years using the code

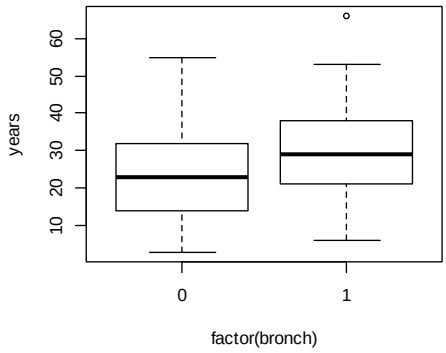
```
par(mfrow=c(2,2))
plot(dust~factor(bronch), data=dust.df)
plot(years~factor(bronch), data=dust.df)
plot(years~dust, data=dust.df)
barplot(smoke.percent, names.arg=c("Non-smokers",
"Smokers"), ylab = "Proportion having bronchial reaction")
```

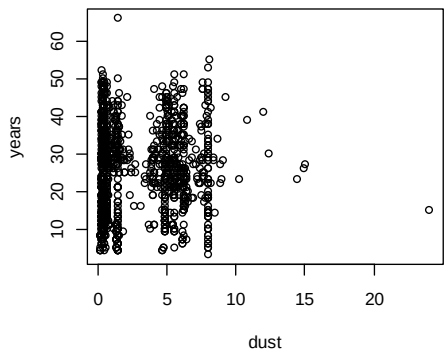
These plots (shown overleaf) clearly show that all three explanatory variables seem related to the response. A more refined plot can be made using trellis graphics. Let's divide the range of the dust and years variables up into say 5 groups, and work out the proportion in each dust/years/smoke group that have bronchial reactions. We can then plot the logits of these proportions against dust, conditioning on years and smoke:

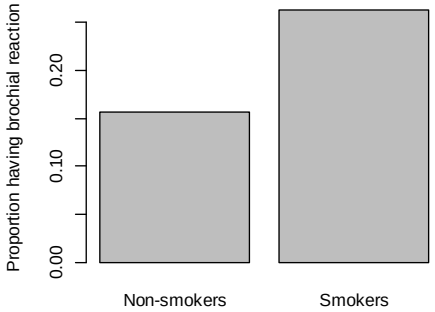
```
dust.cut = cut(dust.df$dust, c(0,
quantile(dust.df$dust,c(0.20, 0.4, 0.6, 0.8)),25))
years.cut = cut(dust.df$years, c(0,
quantile(dust.df$years,c(0.20, 0.4, 0.6, 0.8)),70))
        denom =table(dust.cut, years.cut,dust.df$smoke)
use = (dust.df$bronch==1)
r  =  as.vector(table(dust.cut[use], years.cut[use],
dust.df$smoke[use]))
n =  as.vector(table(dust.cut,  years.cut, dust.df$smoke))
```

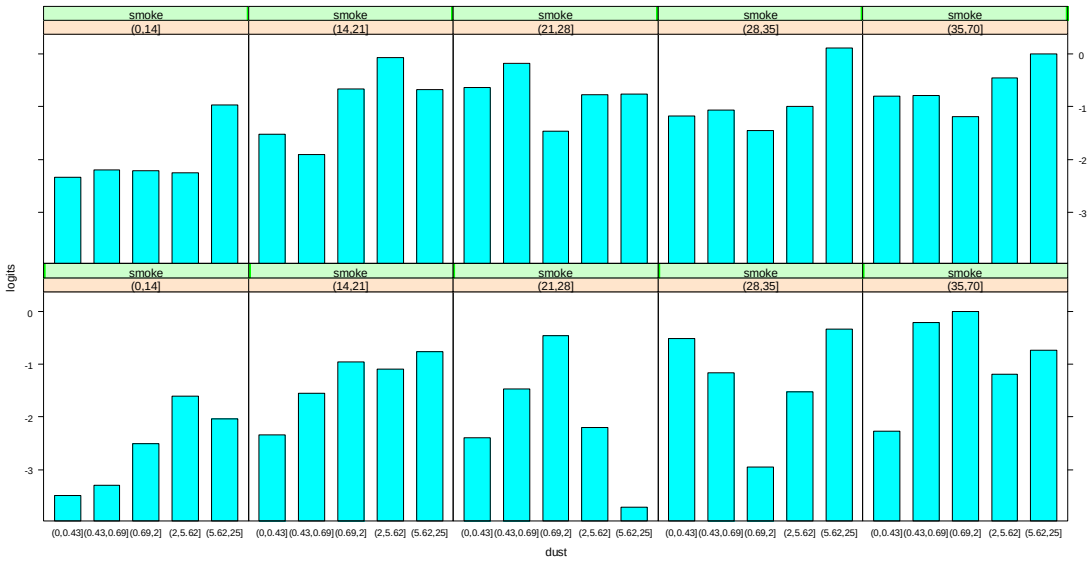
```
plot.data = data.frame(logits=log((r+0.5)/(n-
r+0.5)),expand.grid(dust=levels(dust.cut),
years=levels(years.cut), smoke=0:1))
library(lattice)
        barchart(logits~dust|years*smoke, data=plot.data)
```











From the trellis plot, it seems as though smoking has an effect, as the bars in the top row are usually higher. Also, the effect of dust is to increase the logits, although this effect is generally absent at high values of year for non-smokers (the bottom row).

It is not clear that there is a linear relationship between the logits and the continuous covariates, but this is something we can check by fitting models.

Since the logits in the bottom row seem to differ by non-constant amounts from those in the top row, there is a hint that smoking may interact with the other factors, although the heights of the bars are based on relatively small samples and may be unreliable.

Give 6 marks if the students have drawn a graph (and made suitable comments) that illustrates the relationship between each explanatory variable and the response. Give 4 marks if they have drawn a trellis or similar plot showing how the relationship between a variable and the response changes with the values of the other variables. Total 10.

Task 3. *Fit a logistic regression model to the data, carrying out the usual diagnostic checks. Does this identify any possible risk factors for bronchitis? What is the effect of smoking and dust exposure (both in duration and concentration) on the prevalence of a bronchial reaction? Is there an interaction between smoking and dust exposure? Are these conclusions dependent on the effect of one or two observations?*
Note: *to answer this question, it may be helpful to draw a graph of the probability of having a chronic bronchial reaction, versus concentration, and the same for years of exposure. [15 marks]*

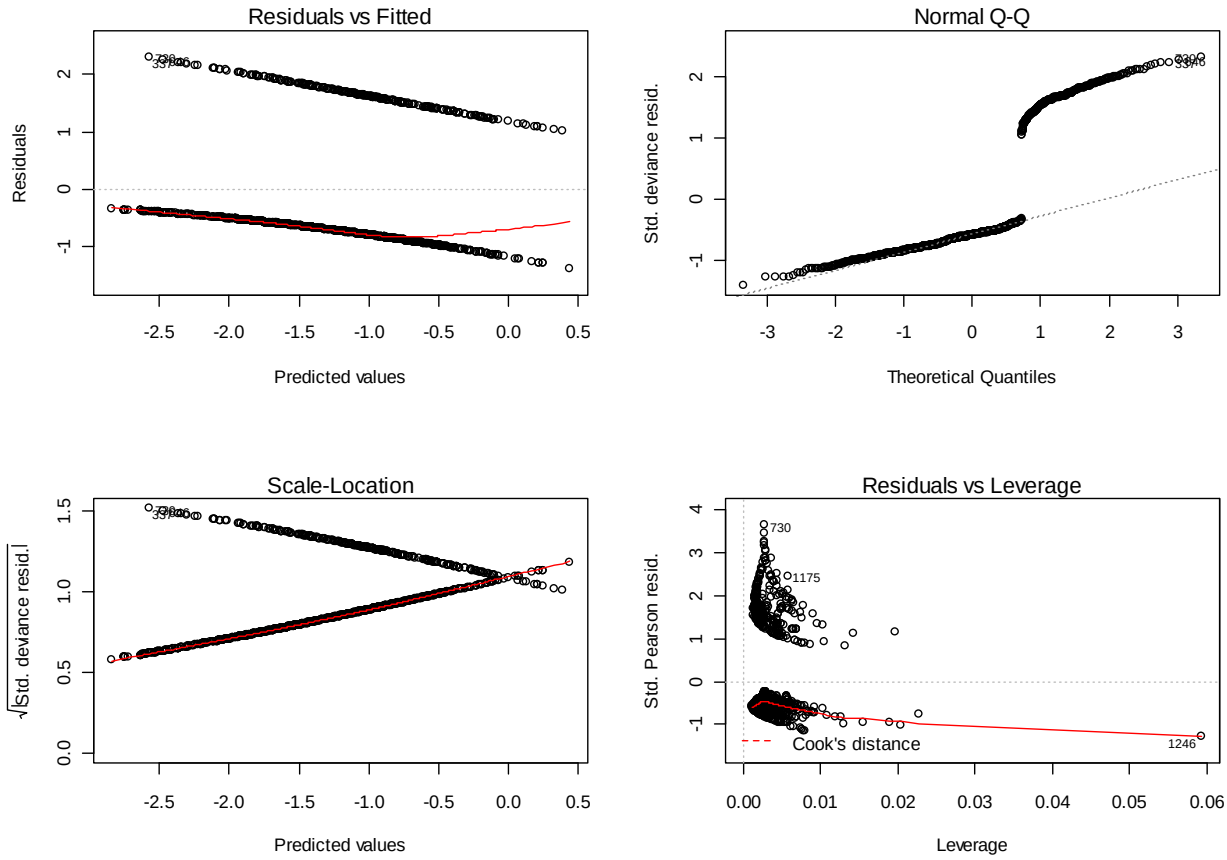
From the plots above, it seems pretty clear that all three explanatory variables have an effect on the response. Let us fit a logistic model

```
model1.glm = glm(bronch~ smoke + years +dust, data= dust.df,
family=binomial)
```

This gives the following plots and summary:

```
Call:
glm(formula = bronch ~ smoke + years + dust, family = binomial,
    data = dust.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3675  -0.7798  -0.5906  -0.3813   2.3022

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.047872   0.248570 -12.262  < 2e-16 ***
smoke        0.676844   0.174380   3.881 0.000104 ***
years        0.040155   0.006206   6.470 9.78e-11 ***
```

```
dust         0.091888   0.023243   3.953 7.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.8  on 1245  degrees of freedom
Residual deviance: 1278.3  on 1242  degrees of freedom
AIC: 1286.3

Number of Fisher Scoring iterations: 4
```



This confirms our impression that all the explanatory variables have an increasing relationship with the probability of a bronchial reaction (all p-values are small).

It seems as though point 1246 has very high leverage. However, removing the point and refitting does not result in much change in the coefficients. However, when we fit more complex models (see below) it causes numerical problems so we leave it out.

Do we need to include polynomial terms? We fit a polynomial model

```
model2.glm = glm(bronch~ smoke + poly(years,3) +  poly(dust,3), data=
dust.df, family=binomial, subset = -1246)
summary(model2.glm)
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.8975     0.1657 -11.448  < 2e-16 ***
smoke             0.6940     0.1776   3.908 9.29e-05 ***
poly(years, 3)1  20.1720     3.1473   6.409 1.46e-10 ***
poly(years, 3)2  -8.2453     2.9806  -2.766  0.00567 **
poly(years, 3)3  10.3699     3.4668   2.991  0.00278 **
poly(dust, 3)1   -0.9050     6.0909  -0.149  0.88188
poly(dust, 3)2  -36.1494    19.1526  -1.887  0.05910 .
poly(dust, 3)3  -23.5823    10.7930  -2.185  0.02889 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1356.3  on 1244  degrees of freedom
Residual deviance: 1250.5  on 1237  degrees of freedom
AIC: 1266.5
```

So it seems that we need polynomial terms. Note that the AIC for this model is lower than that for the linear model fitted above.

Should we include interaction terms (i.e. fit a separate polynomial for smokers and non-smokers?) We can fit this model by

```
model3.glm = glm(bronch~smoke*poly(years,3) + smoke*poly(dust,3),
  data=dust.df, family=binomial, subset=-1246)
```

and compare it to the “no interaction” model by

```
> anova(model3.glm,model2.glm, test="Chisq")
Analysis of Deviance Table

Model 1: bronch ~ smoke * poly(years, 3) + smoke * poly(dust, 3)
Model 2: bronch ~ smoke + poly(years, 3) + poly(dust, 3)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1231     1243.2
2      1237     1250.5 -6  -7.3346    0.291
```

Thus, it seems that there is no interaction. However, things are not quite so simple. We can fit the polynomial model to the smoking group

```
> use = (dust.df$smoke==1)
> use[1246]=FALSE
> model2.smoke.glm = glm(bronch~poly(years,3) + poly(dust,3), data =  dust.df,
+   family=binomial, subset=use)
```

The (abbreviated) summary is

```
Coefficients:
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.22678    0.09273 -13.229  < 2e-16 ***
poly(years, 3)1  20.52426    3.75245   5.470 4.51e-08 ***
poly(years, 3)2  -9.09965    3.48618  -2.610  0.00905 **
poly(years, 3)3  14.35724    4.56935   3.142  0.00168 **
poly(dust, 3)1    0.35460    6.39798   0.055  0.95580
poly(dust, 3)2  -39.09622   20.08198  -1.947  0.05155 .
poly(dust, 3)3  -25.92702   11.57060  -2.241  0.02504 *
```

These coefficients are quite similar to the coefficients of the no-interaction model. This is because there are many more smokers than non-smokers in the dataset (921 versus 325). Thus, the smokers dominate the non-smokers. The fit for the non-smokers is

```
> use = (dust.df$smoke==0)
> use[1246]=FALSE
> model2.nosmoke.glm = glm(bronch~poly(years,3) + poly(dust,3), data =
dust.df,
+   family=binomial, subset=use)

> summary(model2.nosmoke.glm)
```

(abbreviated output)

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.8605     0.2039  -9.123  < 2e-16 ***
poly(years, 3)1  22.2986     6.5823   3.388 0.000705 ***
poly(years, 3)2  -3.3334     6.2861  -0.530 0.595919
poly(years, 3)3   4.0139     4.5369   0.885 0.376310
poly(dust, 3)1   -8.8150    25.4205  -0.347 0.728766
poly(dust, 3)2  -31.8179    81.1743  -0.392 0.695080
poly(dust, 3)3  -17.0583    40.7056  -0.419 0.675168
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that for the non-smokers, there seems to be no effect of dust, and only a linear effect of years. This is supported by an AIC argument: the AIC of this model is 277.44, but the AIC of the model bronch~years fitted to the non-smokers is 268.84.

Thus, to summarise:

All 3 explanatory variables have an effect on the response. There is a polynomial effect for both dust and year for the smokers, but some evidence to suggest that the effect of dust is negligible for the non-smokers, and the effect of years is linear.

Give 5 marks for fitting the additive model (including diagnostics), 5 marks for fitting polynomials in dust and years, and 3 marks for doing the interaction test. Give 2 extra marks for sensible comments about the differences in the bronch-dust-years relationship between smokers and nonsmokers.

Task 4. *Do the explanatory variables have any ability to predict the prevalence of a bronchial reaction? [10 marks].*

We can compare the predictive ability of the variables by computing cross-validated and bootstrap prediction errors and the area under the ROC curve for different models.

The models are

```
model50.glm = glm(bronch~smoke*poly(years,3) +  smoke*poly(dust,3),
      data=dust.df, family=binomial, subset=-1246)


model51.glm = glm(bronch~smoke + poly(years,3) + poly(dust,3),
      data=dust.df,family=binomial, subset=-1246)

model52.glm = glm(bronch~smoke + years + dust, data=dust.df,
      family=binomial,subset=-1246)

model53.glm = glm(bronch~smoke + years, data=dust.df, family=binomial,
      subset=-1246)

model54.glm = glm(bronch~smoke + dust, data=dust.df, family=binomial,
      subset=-1246)

model55.glm = glm(bronch~smoke , data=dust.df, family=binomial,
      subset=-1246)
```

The AIC's percent correctly classified and areas under the ROC curve are

Model	AIC	% CORRECT	AREA
Model50.glm	1271.198	0.760	0.703
Model51.glm	1266.533	0.764	0.691
Model52.glm	1284.332	0.764	0.666
Model53.glm	1299.361	0.765	0.649
Model54.glm	1324.984	0.763	0.606
Model55.glm	1344.61	0.765	0.556

Thus, models 50.glm and 51.glm are the best predictors: both contain all the variables. Thus the variables taken as a group have reasonable predictive power. Smoke alone and smoke and dust alone are poor predictors. Smoke and dust alone are better, but the combination of all three is better still.

Give 5 marks if the students have identified a model having all the variables as being a good predictor, and 5 if they have commented on the AIC or the area under the ROC curve. Note that the earlier version of the ROC function gave slightly different values for the area under the ROC curve.

Total for assignment: 40 marks