A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 18

Incorporating Prior Knowledge via Bayesian Inference: Smoking and Lung Cancer

18.1 Introduction

At the beginning of the 20th century, the death toll due to lung cancer was on the rise and the search for possible causes began. For lung cancer in pit workers, animal experiments showed that the so-called 'Schneeberg lung disease' was induced by radiation. But this could not explain the increasing incidence of lung cancer in the general population. The identification of possible risk factors was a challenge for epidemiology and statistics, both disciplines being still in their infancy in the 1920s and 1930s.

The first modern controlled epidemiological study on the effect of smoking on lung cancer was performed by Franz Hermann Müller as part of his dissertation at the University of Cologne in 1939. The results were published a year later (**?**). Müller sent out questionnaires to the relatives of people who had recently died of lung cancer, asking about the smoking behavior and its intensity of the deceased relative. He also sent the questionnaire to healthy controls to obtain information about the smoking behavior in a control group, although it is not clear how this control group was defined. The number of lung cancer patients and healthy controls in five different groups (nonsmokers to extreme smokers) are given in Table 18.1.

Table 18.1: `Smoking_Mueller1940` data. Smoking and lung cancer case-control study by Müller (1940). The smoking intensities were defined by the number of cigarettes smoked daily: 1-15 (moderate), 16-25 (heavy), 26-35 (very heavy), and more than 35 (extreme).

		`Diagnosis`	
`Smoking`		Lung cancer	Healthy control
	Nonsmoker	3	14
	Moderate smoker	27	41
	Heavy smoker	13	22
	Very heavy smoker	18	5
	Extreme smoker	25	4

Four years later Erich Schöninger also wrote his dissertation on the association between smoking and lung cancer and, together with his supervisor Eberhard Schairer at the University of Jena, published his results on a case-control study (**?**) where he assessed the smoking behavior of lung cancer patients, patients diagnosed with other forms of cancer, and also a healthy control group. The data are given in Table 18.2.

Table 18.2: `Smoking_SchairerSchoeniger1944` data. Smoking and lung cancer case-control study by Schairer and Schöniger (1944). Cancer other than lung cancer omitted. The smoking intensities were defined by the number of cigarettes smoked daily: 1-5 (moderate), 6-10 (medium), 11-20 (heavy), and more than 20 (very heavy).

	Diagnosis	
`Smoking`	Lung cancer	Healthy control
Nonsmoker	3	43
Moderate smoker	11	98
Medium smoker	31	57
Heavy smoker	19	47
Very heavy smoker	29	25

Shortly after the war, a Dutch epidemiologist reported on a case-control study performed in Amsterdam (**?**) and found similar results as the two German studies; see Table 18.3.

Table 18.3: `Smoking_Wassink1945` data. Smoking and lung cancer case-control study by Wassink (1945). Smoking categories correspond to the categories used by Müller (1940).

	Diagnosis	
`Smoking`	Lung cancer	Healthy control
Nonsmoker	6	19
Moderate smoker	18	36
Heavy smoker	36	25
Very heavy smoker	74	20

In 1950 perhaps the most important, but not the first, case-control study showing an increasing risk of developing lung cancer with the amount of tobacco smoked, was published in Great Britain by Richard Doll and Austin Bradford Hill (**?**). We restrict discussion here to data obtained for males and the

data shown in Table 18.4 corresponds to the most recent amount of tobacco consumed regularly by smokers before disease onset (Table V in **?**).

Table 18.4: `Smoking_DollHill1950` data. Smoking and lung cancer case-control study (only males) by Doll and Hill (1950). The labels for the smoking categories give the number of cigarettes smoked every day.

		Diagnosis	
Smoking		Lung cancer	Other
	Nonsmoker	2	27
	1-	33	55
	5-	250	293
	15-	196	190
	25-	136	71
	50+	32	13

Although the design of the studies by **?** and **?**, especially the selection of their control groups, can be criticized (see **?**, for a detailed discussion) and the study by **?** was larger than the older studies and more detailed information on the smoking behavior was obtained by direct patient interviews, the information provided by the earlier studies was not taken into account by **?**. They cite **?** in their introduction, but did not compare their findings with his results. It is remarkable to see that both **?** and **?** extensively made use of the report by **?** and go as far as analyzing the merged data (Grafiek I, E, and F, in **?**). In an informal way, these authors wanted to use the already available information, in today's terms called 'prior knowledge', to make a stronger case with the new data. Formal statistical methods to incorporate prior knowledge into data analysis as part of the 'Bayesian' way of doing statistical analyses were developed in the second half of the last century, and we will focus on them in the present chapter.

18.2 Bayesian Inference

18.3 Analysis Using R

18.3.1 One-by-one Analysis

For the analysis of the four different case-control studies on smoking and lung cancer, we will (retrospectively, of course) update our knowledge with every new study. We begin with a re-analysis of the data described by **?**. Using an approximate permutation test introduced in Chapter **??** for the hypothesis of independence of the amount of tobacco smoked and group membership (lung cancer or healthy control), we get

```
R> library("coin")
R> set.seed(29)
R> independence_test(Smoking_Mueller1940, teststat = "quad",
+                    distribution = approximate(100000))
```

```
        Approximative General Independence Test

data:  Diagnosis by
         Smoking (Nonsmoker, Moderate smoker, Heavy smoker, Very heavy smoker, Extreme smoker)
chi-squared = 30, p-value <2e-16
```

and there is clearly a strong association between the number of cigarettes smoked and incidence of lung cancer. Because the amount of tobacco smoked is an ordered categorical variable, it is more appropriate to take this information into account, for example by means of a linear association test (see Chapter **??**). Nonsmokers receive a score of zero, and for the remaining groups we choose the mid-point of the intervals of daily cigarettes smoked that were used by **?** to define his groups:

```
R> ssc <- c(0, 1 + 14 / 2, 16 + 9 / 2, 26 + 9 / 2, 40)
R> independence_test(Smoking_Mueller1940, teststat = "quad",
+      scores = list(Smoking = ssc),
+      distribution = approximate(100000))
```

```
        Approximative General Independence Test

data:  Diagnosis by
         Smoking (Nonsmoker < Moderate smoker < Heavy smoker < Very heavy smoker < Extreme smoker)
chi-squared = 30, p-value <2e-16
```

The result shows that the data are in favor of an ordered alternative. The p-values obtained from approximate permutation tests are attractive because no distributional assumptions are required, but it is hard to derive estimates and confidence intervals for interpretable parameters from such tests. We will therefore now switch to logistic regression models as described in Chapter **??** to model the odds of lung cancer in the different smoking groups. Before we start, let us define a small function for computing odds (for intercept parameters) and odds ratios (for difference parameters) and corresponding confidence intervals from a logistic regression model:

```
R> eci <- function(model)
+      cbind("Odds (Ratio)" = exp(coef(model)),
+            exp(confint(model)))
```

We model the probability of developing lung cancer given the smoking behavior. Because our data was obtained from case-control studies where the groups (lung cancer patients and healthy controls) were defined first and only after that we observed data on the smoking behavior (in a so-called *choice-based sampling*), this may seem the wrong model to start with. However, the marginal distribution of the two groups only changes the intercept in such a logistic model and the effects of smoking can still be interpreted in the way we require (see **?**, for example). The formula for specifying a logistic regression model can be set up such that the response is a matrix with two columns for each

smoking group consisting of the number of lung cancer deaths and the number of healthy controls. Although smoking is an ordered factor, we first fit the model with treatment contrasts, i.e., we can interpret the exp of the regression coefficients as odds ratios between each smoking group and nonsmokers:

```
R> smoking <- ordered(rownames(Smoking_Mueller1940),
+                     levels = rownames(Smoking_Mueller1940))
R> contrasts(smoking) <- "contr.treatment"
R> eci(glm(Smoking_Mueller1940 ~ smoking, family = binomial()))
```

```
                       Odds (Ratio)  2.5 %  97.5 %
(Intercept)                   0.214 0.0494   0.656
smokingModerate smoker        3.073 0.8986  14.242
smokingHeavy smoker           2.758 0.7272  13.608
smokingVery heavy smoker     16.800 3.8256  98.205
smokingExtreme smoker        29.167 6.4728 180.002
```

We see that all but one of the odds ratios increase with the amount of tobacco smoked with a maximum of almost 30 for extreme smokers (more than 35 cigarettes per day). The likelihood confidence intervals are rather wide due to the limited sample size, but also the lower limit increases with smoking.

An alternative model formulation can help to compare each smoking group with the preceding group, the so-called split-coding (for this and other codings see `?`):

```
R> K <- diag(nlevels(smoking) - 1)
R> K[lower.tri(K)] <- 1
R> contrasts(smoking) <- rbind(0, K)
R> eci(glm(Smoking_Mueller1940 ~ smoking, family = binomial()))
```

```
            Odds (Ratio)  2.5 % 97.5 %
(Intercept)        0.214 0.0494  0.656
smoking1           3.073 0.8986 14.242
smoking2           0.897 0.3812  2.066
smoking3           6.092 1.9296 22.181
smoking4           1.736 0.4051  7.894
```

The two largest differences are between moderate smokers and nonsmokers (`smoking1`) and between very heavy and heavy smokers (`smoking3`). The latter group difference seems, at least judged by the confidence interval, to be larger than expected under a model with no effect of smoking.

For the analysis of the three remaining studies, we first perform permutation tests for the independence of smoking and the two groups (lung cancer and healthy controls) in males:

```
R> xSS44 <- as.table(Smoking_SchairerSchoeniger1944[,
+      c("Lung cancer", "Healthy control")])
R> ap <- approximate(100000)
R> pvalue(independence_test(xSS44,
+         teststat = "quad", distribution = ap))
```

```
[1] 0
99 percent confidence interval:
 0.0e+00 5.3e-05
```

```
R> pvalue(independence_test(Smoking_Wassink1945,
+         teststat = "quad", distribution = ap))
```

```
[1] 0
99 percent confidence interval:
 0.0e+00 5.3e-05
```

```
R> xDH50 <- as.table(Smoking_DollHill1950[,, "Male"])
R> pvalue(independence_test(xDH50,
+         teststat = "quad", distribution = ap))
```

```
[1] 0
99 percent confidence interval:
 0.0e+00 5.3e-05
```

All p-values indicate that the data are not well-described by the independence model.

18.3.2 Joint Bayesian Analysis

For a Bayesian analysis, we first merge the data from all four studies into one data frame. In doing so, we also merge the smoking groups in a way that we only have three groups left: nonsmokers, moderate smokers, and heavy smokers. These groups are chosen in a way that the number of daily cigarettes is comparable. We first merge the heavy, very heavy, and extreme smokers from **?**

```
R> (M <- rbind(Smoking_Mueller1940[1:2,],
+               colSums(Smoking_Mueller1940[3:5,])))
```

```
                Lung cancer Healthy control
Nonsmoker                 3              14
Moderate smoker          27              41
                         56              31
```

and proceed with the lung cancer patients and healthy controls from **?** in the same way

```
R> SS <- Smoking_SchairerSchoeniger1944[,
+      c("Lung cancer", "Healthy control")]
R> (SS <- rbind(SS[1,], colSums(SS[2:3,]), colSums(SS[4:5,])))
```

```
     Lung cancer Healthy control
[1,]           3              43
[2,]          42             155
[3,]          48              72
```

and finally perform the same exercise for the **?** and **?** data

```
R> (W <- rbind(Smoking_Wassink1945[1:2,],
+               colSums(Smoking_Wassink1945[3:4,])))
```

```
                Lung cancer Healthy control
Nonsmoker                 6              19
Moderate smoker          18              36
                        110              45
```

```
R> DH <- Smoking_DollHill1950[,, "Male"]
R> (DH <- rbind(DH[1,], colSums(DH[2:3,]), colSums(DH[4:6,])))
```

```
     Lung cancer Other
[1,]           2    27
[2,]         283   348
[3,]         364   274
```

The three new groups are now called nonsmokers, moderate smokers, and heavy smokers, and we set up a data frame that contains the number of people in each of the possible groups for all studies:

```
R> smk <- c("Nonsmoker", "Moderate smoker", "Heavy smoker")
R> x <- expand.grid(Smoking = ordered(smk, levels = smk),
+    Diagnosis = factor(c("Lung cancer", "Control")),
+    Study = c("Mueller1940", "SchairerSchoeniger1944",
+              "Wassink1945", "DollHill1950"))
R> x$weights <- c(as.vector(M), as.vector(SS),
+                 as.vector(W), as.vector(DH))
```

Before we fit logistic regression models using the data organized in such a way, we define the contrasts for the smoking ordered factor and expand the data in a way that each row corresponds to one person. This is necessary because the `weights` argument to the `glm` function must not be used to define case weights:

```
R> contrasts(x$Smoking) <- "contr.treatment"
R> x <- x[rep(1:nrow(x), x$weights),]
```

We now compute one logistic regression model for each study for later comparisons:

```
R> models <- lapply(levels(x$Study), function(s)
+      glm(Diagnosis ~ Smoking, data = x, family = binomial(),
+          subset = Study == s))
R> names(models) <- levels(x$Study)
```

In 1939, Müller was hardly in the position to come up with a reasonable prior for the odds ratios between moderate or heavy smokers and nonsmokers. So we also use a noninformative prior and just perform the maximum likelihood analysis:

```
R> eci(models[["Mueller1940"]])
```

```
                      Odds (Ratio)  2.5 % 97.5 %
(Intercept)                  0.214 0.0494  0.656
SmokingModerate smoker       3.073 0.8986 14.242
SmokingHeavy smoker          8.430 2.5199 38.641
```

Four years later, the maximum likelihood results obtained for the **?** data

```
R> eci(models[["SchairerSchoeniger1944"]])
```

```
                      Odds (Ratio)  2.5 % 97.5 %
(Intercept)                 0.0698 0.0169  0.191
SmokingModerate smoker      3.8839 1.3284 16.569
SmokingHeavy smoker         9.5556 3.2417 40.975
```

could have been improved by using a normal prior for the difference in log odds whose distribution is the distribution of the maximum likelihood estimator obtained for Müller's data. At least approximately, we can compute posterior 90% credibility intervals and the posterior mode from the Schairer and Schöniger data by analyzing both data sets simultaneously. We should, however, keep in mind that the odds of developing lung cancer for nonsmokers is not really interesting for our analysis and that the four studies may very well

differ with respect to this intercept parameter. Consequently, we don't want to specify a prior for the intercept. One way to implement such a strategy is to exclude the intercept term from the joint model while allowing a separate intercept for each of the studies:

```
R> mM40_SS44 <- glm(Diagnosis ~ 0 + Study + Smoking, data = x,
+      family = binomial(),
+      subset = Study %in% c("Mueller1940",
+                            "SchairerSchoeniger1944"))
R> eci(mM40_SS44)
```

```
                          Odds (Ratio)  2.5 % 97.5 %
StudyMueller1940                0.1955 0.0732  0.438
StudySchairerSchoeniger1944     0.0753 0.0284  0.166
SmokingModerate smoker          3.5212 1.5441  9.538
SmokingHeavy smoker             9.0121 3.9572 24.398
```

We observe two important differences between the maximum likelihood and Bayesian results for the Schairer and Schöniger data: In the Bayesian analysis, the estimated odds ratio for moderate smokers is closer to the smaller value obtained from Müller's data and, more important, the credibility intervals are much narrower and, one has to say, more realistic now. An odds ratio as large as 40 is hardly something one would expect to see in practice.

If Wassink had been aware of Bayesian statistics, he could have used the posterior distribution of the parameters from our model `mM40_SS44` as a prior distribution for analyzing his data. The maximum likelihood results for his data

```
R> eci(models[["Wassink1945"]])
```

```
                     Odds (Ratio) 2.5 % 97.5 %
(Intercept)             0.316 0.115  0.747
SmokingModerate smoker  1.583 0.558  4.965
SmokingHeavy smoker     7.741 3.054 22.421
```

would have changed to

```
R> mM40_SS44_W45 <- glm(Diagnosis ~ 0 + Study + Smoking,
+      data = x, family = binomial(),
+      subset = Study %in% c("Mueller1940",
+                            "SchairerSchoeniger1944",
+                            "Wassink1945"))
R> eci(mM40_SS44_W45)
```

```
                          Odds (Ratio)  2.5 % 97.5 %
StudyMueller1940                0.2250 0.1096  0.428
StudySchairerSchoeniger1944     0.0878 0.0436  0.163
StudyWassink1945                0.2603 0.1298  0.482
SmokingModerate smoker          2.7570 1.4554  5.629
SmokingHeavy smoker             8.3795 4.5061 16.862
```

The rather small odds ratios obtained from the model fitted to the Wassink data only are now closer to the estimates obtained from the two previous studies and the variability, as given by the credibility intervals, is much smaller.

Now, finally, the model for the Doll and Hill data reports rather large odds ratios with wide confidence intervals:

```
R> eci(models[["DollHill1950"]])
```

```
                     Odds (Ratio)  2.5 %  97.5 %
(Intercept)                0.0741 0.0119   0.247
SmokingModerate smoker    10.9784 3.2545  68.434
SmokingHeavy smoker       17.9343 5.3168 111.793
```

With a (now rather strong) prior defined by the three earlier studies, we get from the joint model for all four studies

```
R> m_all <- glm(Diagnosis ~ 0 + Study + Smoking, data = x,
+               family = binomial())
R> eci(m_all)
```

```
                           Odds (Ratio)  2.5 % 97.5 %
StudyMueller1940                 0.1772 0.0911  0.323
StudySchairerSchoeniger1944      0.0665 0.0349  0.118
StudyWassink1945                 0.2200 0.1160  0.390
StudyDollHill1950                0.1629 0.0874  0.282
SmokingModerate smoker           4.5131 2.5918  8.451
SmokingHeavy smoker              8.8971 5.1298 16.605
```

In 1950, the joint evidence based on such an analysis with an odds ratio between 2.6 and 8.5 for moderate smokers and between 5.1 and 16.6 for heavy smokers compared to nonsmokers, would have made a much stronger case than any of the single studies alone. It is interesting to see that with this strong prior for the Doll and Hill study, we also get relatively large odds ratios when comparing heavy to moderate smokers (see row labeled `Smoking2`):

```
R> K <- diag(nlevels(x$Smoking) - 1)
R> K[lower.tri(K)] <- 1
R> contrasts(x$Smoking) <- rbind(0, K)
R> eci(glm(Diagnosis ~ 0 + Study + Smoking, data = x,
+          family = binomial()))
```

```
                           Odds (Ratio)  2.5 % 97.5 %
StudyMueller1940                 0.1772 0.0911  0.323
StudySchairerSchoeniger1944      0.0665 0.0349  0.118
StudyWassink1945                 0.2200 0.1160  0.390
StudyDollHill1950                0.1629 0.0874  0.282
Smoking1                         4.5131 2.5918  8.451
Smoking2                         1.9714 1.6384  2.374
```

18.3.3 A Comparison with Meta Analysis

One may ask how the Bayesian approach of progressively updating the estimates considered here differs from a classical meta analysis described in Chapter **??**. We first reshape the data into a form suitable for such an analysis

```
R> y <- xtabs(~ Study + Smoking + Diagnosis, data = x)
R> ntrtM <- margin.table(y, 1:2)[,"Moderate smoker"]
R> nctrl <- margin.table(y, 1:2)[,"Nonsmoker"]
R> ptrtM <- y[,"Moderate smoker","Lung cancer"]
R> pctrl <- y[,"Nonsmoker","Lung cancer"]
R> ntrtH <- margin.table(y, 1:2)[,"Heavy smoker"]
R> ptrtH <- y[,"Heavy smoker","Lung cancer"]
```

and then compute joint odds ratios and confidence intervals for moderate and heavy smokers compared to nonsmokers:

```
R> library("rmeta")
R> meta.MH(ntrt = ntrtM, nctrl = nctrl,
+           ptrt = ptrtM, pctrl = pctrl)
```

```
Fixed effects ( Mantel-Haenszel ) Meta-Analysis
Call: meta.MH(ntrt = ntrtM, nctrl = nctrl, ptrt = ptrtM, pctrl = pctrl)
Mantel-Haenszel OR =3.89    95% CI ( 2.13, 7.09 )
Test for heterogeneity: X^2( 3 ) = 4.77 ( p-value 0.1897 )
```

```
R> meta.MH(ntrt = ntrtH, nctrl = nctrl,
+           ptrt = ptrtH, pctrl = pctrl)
```

```
Fixed effects ( Mantel-Haenszel ) Meta-Analysis
Call: meta.MH(ntrt = ntrtH, nctrl = nctrl, ptrt = ptrtH, pctrl = pctrl)
Mantel-Haenszel OR =10.25    95% CI ( 5.59, 18.8 )
Test for heterogeneity: X^2( 3 ) = 0.99 ( p-value 0.8042 )
```

For moderate smokers, the effect is a little weaker compared with the results reported on earlier and for heavy smokers, the meta analysis identifies a stronger effect for heavy smokers. Nevertheless, the differences between the two rather different approaches are negligible and the conclusions would have been the same.

18.4 Summary of Findings

We have seen that, using a Bayesian approach to incorporate prior knowledge into a model, the odds of developing lung cancer increase with increased amounts of smoking. Of course, our analysis here is very simplistic, because we ignored that also pipe and cigar smokers were present in the data, we merged the data based on a very rough assessment of the number of cigarettes smoked per day, ignored whether or not the smokers inhaled the smoke into their lungs, or if nonsmokers were subject to passive-smoking, as we call it today. Most importantly, we must not misinterpret findings from case-control studies as casual and, in fact, none of the authors cited here did so. The debate on whether smoking, and which kind of smoking, actually *causes* lung cancer was initiated by the publications cited in this chapter and many famous statisticians took part in the debate, for example, Sir Ronald Fisher (**?**), took the view that the inference of causation was premature. In retrospect this was one issue (perhaps the only one) where Fisher was mistaken.

18.5 Final Comments

There remain a few hard-line opponents of Bayesian inference (just a few) who reject the method because of the use of subjective prior distributions which, these opponents feel, have no place in scientific investigations. And there are Bayesians who think that the only defense of using non-Bayesian methods is incompetence.

But for an increasing number of statisticians Bayesian inference is very attractive, because we can use the posterior distribution of the parameters to

draw conclusions from the data. Although this requires the specification of a prior distribution, we have seen in this chapter that, using data from previous experiments, priors can be defined in a reasonable way. It is not absolutely necessary to rely on rather complex numerical procedures to'estimate' a posterior distribution. When we are willing to cut some corners, we can implement simple Bayesian approaches using standard software. We should also keep in mind that the prior can be interpreted as a penalty on the parameters, and many penalization approaches therefore have an (often implicit) connection to the Bayesian way of doing statistics. Of course, just picking the prior that 'works best' is dangerous and almost surely inappropriate.

Exercises

Ex. 18.1 Produce a forest plot as introduced in Chapter **??** for the four smoking studies analyzed here.

Ex. 18.2 Produce a modified forest plot where one can see how the evidence for smoking being related to lung cancer evolved between 1940 and 1950.

Ex. 18.3 Use the **INLA** add-on package to perform a similar analysis by using the coefficients and their standard errors estimated from our initial logistic regression model `m[["Mueller1940"]]` as parameters of a normal prior for a logistic regression applied to the Schairer and Schöniger data. Compare the resulting credibility intervals for the two odds-ratios with the approximate results obtained in this chapter.