KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

On focussed and less focussed model selection

Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0613

On Focussed and Less Focussed Model Selection

Gerda Claeskens[1]

[1] ORSTAT and University Centre for Statistics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

Abstract: Model selection usually provides models without specific concern about for which purpose the selected model will be used afterwards. The focussed information criterion, FIC, is developed to select a best model for inference for a given estimand. For example, in regression models the FIC can be used to select a model for the mean response for each individual subject in the study. This can be used to identify interesting subgroups in the data. Sometimes the FIC is considered too much focussed. We rather would want to select a model that performs well for a whole subgroup, or even for all of the subjects in the study. We explain how to make the focussed information criterion a little less focussed via weighting methods.

Keywords: Focussed Information Criterion; Model selection.

1 Description of the dataset

The dataset considered for discussion in this paper is from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984). It provides information to study diabetic retinopathy as a function of several measurements. The dataset consists of patient information for 343 women and 348 men. The binary outcome variable $Y = 0$ indicates whether there is no or only mild nonproliferate retinopathy on both of the eyes. An outcome value $Y = 1$ is obtained when there is moderate to severe nonproliferate retinopathy, or proliferate retinopathy for at least one of the eyes. Other variables are: x_1: the duration of diabetes in years; z_1: indicator for presence of macular edema in at least one eye; z_2: the percentage of glycosylated hemoglobin; z_3: the body mass index; z_4: pulse rate in beats per 30 seconds; z_5: sex, using 1 for male and 0 for female, z_6: indicator for presence of urine protein; and z_7: area of residence (1=urban, 2=rural).
We fit a logistic regression model to these data. In an earlier analysis of this dataset it is found that duration of diabetes is an important variable (see for example Claeskens, Croux and Van Kerckhoven, 2006), therefore, in this example, we include this variable in all of the models we consider,

as well as an intercept term. The model takes the form

$$\text{logit}P(Y=1) = \beta_0 + \beta_1 x_1 + \sum_{s\in S} \beta_s z_s,$$

where S is an index set, containing the variables contained in the corresponding model. For example, the biggest model considered has $S = \{1, 2, \ldots, 7\}$, containing all of the variables $z_1, \ldots, z_7$. The empty set $S = \emptyset$ corresponds to fitting the smallest model $\text{logit}P(Y=1) = \beta_0 + \beta_1 x_1$. The model selection questions we pose are which of the variables z_j are to be included in the model, thereby distinguishing between an individual model search, and a search for the subset of women separately.

2 Individual model searches by FIC

In this section we explain how to obtain a *very* focussed model selection, subject specific. This follows the approach as in Claeskens and Hjort (2003). We start with defining the focus parameter. For ease of explanation we use the logistic regression model as in the example in the previous section. As a focus parameter we take the linear predictor

$$\mu = x^t\beta + z^t\gamma = \text{ logit}\{E(Y|x,z)\}.$$

It is important to note that this focus parameter changes for each different set of covariate values (x, z). We define the design matrices X and Z, of dimension $n \times p$ and $n \times q$ respectively. In the above example we have $p = 1$ and $q = 7$. The ith row of X consists of $(1, x_{1i}, \ldots, x_{p-1,i})$, and the ith row of Z consists of $(z_{1i}, \ldots, z_{qi})$, for $i = 1, \ldots, n$. Here we assume that the intercept is contained in all models, of course this can be changed. The Fisher information matrix in the largest model is crucial for the FIC, and in particular the lower right submatrix of the inverse Fisher information matrix, which we denote by K_n. For logistic regression

$$K_n = n\{Z^tV(I - X(X^tVX)^{-1}X^tV)Z\}^{-1},$$

where $V = \text{diag}\{p_i(1-p_i)\}$, with $p_i = P(Y_i = 1|x_i, z_i)$ according to the logit model including all variables. For practical use, we insert estimators for the unknown parameters, obtained from the full model. A third ingredient is a vector

$$\omega = Z^tVX(X^tVX)^{-1}x - z,$$

also clearly depending on the covariate values (x, z). This, together with estimators $\hat{\gamma}$ obtained in the largest model, is all that is needed to obtain the value of the focussed information criterion, for the purpose of model selection at the exact covariate position (x, z). The value of FIC is obtained for each model indexed by a set S separately. Each time we select the

corresponding rows and columns of the matrix K_n^{-1}. For example, if $S = \{2, 5\}$, we select from K_n^{-1} the 2×2 submatrix containing entries at the 2nd and 5th rows and 2nd and 5th columns. Algebraically this is denoted by means of a projection matrix π_S, which is of dimension $|S| \times q$ and selects from a vector v only those components v_j for which $j \in S$, which we denote by $\pi_S v = v_S$. We also used the notation $|S|$ for the number of components in S. With this we define $K_{n,S} = (\pi_S K_n^{-1} \pi_S^t)^{-1}$, and $G_{n,S} = \pi_S^t K_{n,S} \pi_S K_n^{-1}$, and finally

$$\mathrm{FIC}(S; x, z) = n\omega^t (I_q - G_{n,S})\hat{\gamma}\hat{\gamma}^t (I_q - G_{n,S})^t \omega + 2\omega^t \pi_S^t K_{n,S} \pi_S \omega.$$

The FIC decomposes into a squared bias estimator and a variance estimator (times 2). In cases where the estimated squared bias component happens to be negative, we replace it by zero.
Let S_1, S_2,... be the set of index sets for the model search. We obtain a list of corresponding $\mathrm{FIC}(S_1; x, z)$, $\mathrm{FIC}(S_2; x, z)$,.... The best model is that one with the smallest value of FIC.

3 Individual model selection for the dataset

The WESDR dataset is used for an individual model search. We perform the model search as described above, where the covariate positions chosen correspond to the subset of 343 women in the study. This means, for each (x_i, z_i) where $i = 1, \ldots, 343$ we compute all $2^7 = 128$ models in an all subsets model search. The best model according to the FIC, is that model for which the corresponding FIC value is the smallest of the 128 FIC values computed for this person. This dataset is quite rich in the sense that it here is not obvious that one model is best for all different subjects. In fact, the picture in Figure 1 shows that 39 models were selected at least once. Three model stand out: they got selected 95, 46 and 35 times respectively. The order of the model numbers is irrelevant. These models include only variable z_1, no extra variables, and only variable z_6 respectively.
A study of the 39 selected models reveals that z_1, presence of macular edema was chosen in 62.1% of the 343 subject specific models, z_3 pulse was chosen 20.1%, hemoglobin level z_6 came third with 14.9%. Table 1 gives the full results for all seven variables, for the subsets of men and women separately.
Let us now take a closer look at the most selected individual model, the model with extra variable z_1. Figure 2 summarises the search result for this model across all female individuals in the study. We already know that the model received 95 times rank 1. A closer look reveals that it received 25 times rank 2, 23 times rank 3 and for 230 out of the 343 subjects, corresponding to 67%, it received a rank number no higher than 10 (out of 128).

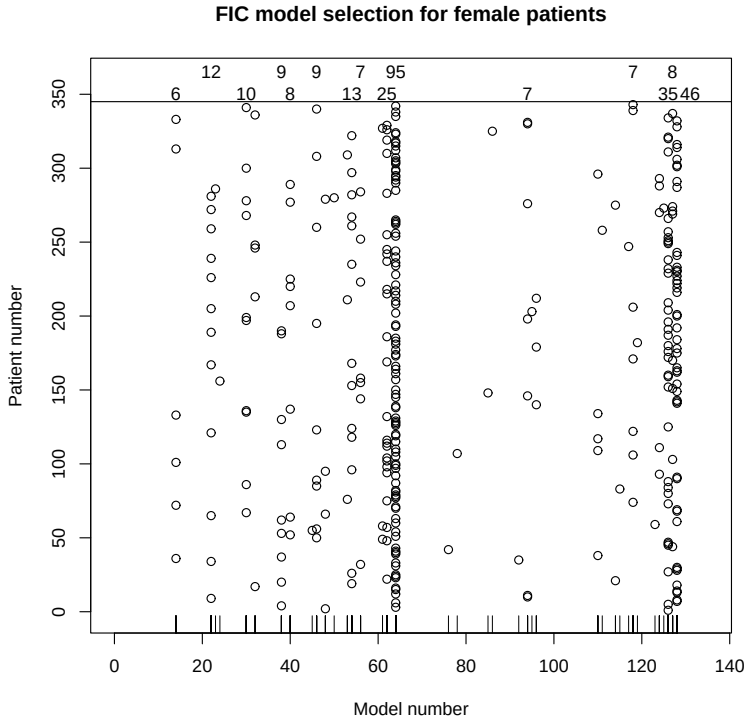

FIGURE 1. For each subject in the study a separate model search has been performed amongst $2^7 = 128$ different models. The graph shows the model number (from 1 to 128) and indicates for which subjects that particular model was selected. The numbers on top are the most frequent model counts. 39 models got selected at least once.

TABLE 1. Percentage of times that a variable has been selected by the individual model searches.

Variable	Women	Men
z_1: Edema	0.621	0.693
z_6: Urine protein	0.446	0.382
z_4: Pulse	0.201	0.216
z_2: Hemoglobin	0.149	0.158
z_3: Body mass index	0.131	0.118
z_7: Residence	0.070	0.083
z_5: Sex	0.035	0.034

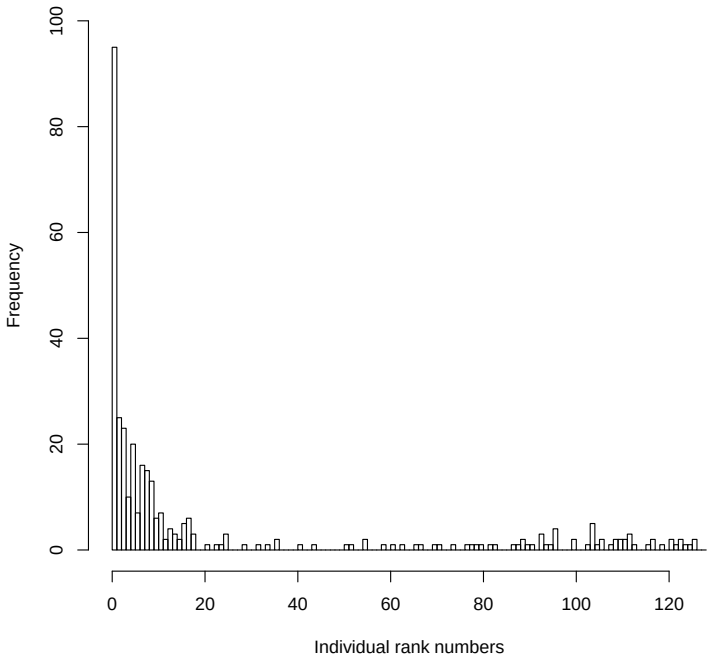

FIGURE 2. Across all 343 female patients in the study, we count the rank number of the most often individually selected model according to a subjectwise FIC, including extra variable z_1, amongst all possible subset models. A frequency histogram is shown here. This model received 95 times rank 1, 25 times rank 2, etc.

Figures 1 and 2 clearly illustrate that, at least for some datasets, it may be advantageous to perform a subjectwise model selection. There is no obvious overall winner.

4 Weighted FIC

An individual model search is sometimes not what one wants, rather, we wish to select a model that is focussed towards estimation of a certain focus parameter, but at the same time the model should perform well for all of the subjects. One strategy is to choose average values for the covariates and go through the model selection step for this hypothetical average subject. This would follow exactly the same steps as an individual model search, and

needs no further explanation. In this section we discuss how to construct a weighted version of the FIC. We start with a weighted average quadratic loss function on the scale of the linear predictor

$$\sum_{i=1}^{n} w(x_i, z_i)\{\hat{\mu}_S(x_i, z_i) - \mu_{\text{true}}(x_i, z_i)\}^2.$$

The weights $w(x_i, z_i)$ are user-specified. The expected value of this loss function can again be decomposed into a weighted squared bias and weighted variance term. The weighted focussed information criterion wFIC is an estimator hereof. To state the exact definition, denote W=diag$\{w(x_i, z_i)\}$, and let

$$\Omega_n = \frac{1}{n}\begin{pmatrix} X^tWX & X^tWZ \\ Z^tWX & Z^tWZ \end{pmatrix}, \quad J_n = \frac{1}{n}\begin{pmatrix} X^tVX & X^tVZ \\ Z^tVX & Z^tVZ \end{pmatrix}.$$

We denote by $J_{n,S}$ the corresponding submatrix for the model indexed by S. The matrix Ω_n only differs from the Fisher information matrix J_n by the use of the weight matrix W instead of the logistic weights V. Further, define for each index set S

$$F_{n,S} = \begin{pmatrix} (X^tVX)^{-1}X^tVZ(I_q - G_{n,S}) \\ -(I_q - G_{n,S}) \end{pmatrix}$$

and an extended projection matrix of dimension $(p + |S|) \times (p + q)$,

$$\widetilde{\pi}_S = \begin{pmatrix} I_p & 0_{p,q} \\ 0_{|S|,p} & \pi_S \end{pmatrix}.$$

The weighted FIC is now computed as follows:

$$w\text{FIC} = \text{trace}(\Omega_n \widetilde{\pi}_S^t J_{n,S}^{-1} \widetilde{\pi}_S) + \max\big\{\text{trace}\{\Omega_n F_{n,S}(n\hat{\gamma}\hat{\gamma}^t - \hat{K}_n)F_{n,S}^t\}, 0\big\}.$$

The construction with the truncation by zero avoids obtaining a negative bias-squared estimate. For more details and a justification of this derivation we refer to Claeskens and Hjort (2006).

5 Logistic variance weights and a connection to AIC

The weights in the weighted FIC are user defined. One particular type of weights leads not only to a simplification of the wFIC formula, but also to a connection to Akaike's (1974) information criterion. Let us take $w(x_i, z_i) = p_i(1 - p_i)$, which implies that $W = V$, and hence that $\Omega_n = J_n$ and $\text{trace}(\Omega_n \widetilde{\pi}_S^t J_{n,S}^{-1} \widetilde{\pi}_S) = p + |S|$, the number of parameters in the model indexed by S. For a positive estimated squared bias, we then get that, for this particular choice of weights,

$$w\text{FIC} = n\hat{\gamma}^t(K_n^{-1} - K_n^{-1}\pi_S^t K_{n,S}\pi_S K_n^{-1})\hat{\gamma} + 2|S| + p - q.$$

Let us compare this to Akaike's information criterion

$$\text{AIC}(S) = -2\sum_{i=1}^{n} \log f(y_i, x_i, z_i, \hat{\beta}, \hat{\gamma}) + 2(p + |S|).$$

The AIC values are obtained for each submodel S, including $S = \emptyset$, corresponding to not including any extra variables z_j. When subtracting the smallest model's AIC value from AIC(S), and performing a one-step Taylor expansion, we find that, for $n \to \infty$,

$$\text{AIC}(S) - \text{AIC}(\emptyset) \xrightarrow{d} -D^t K_n^{-1} \pi_S^t K_{n,S} \pi_S K_n^{-1} D + 2|S|.$$

See Claeskens and Hjort (2003, eq. (2.5)). The variable D in the expression above is in practice estimated via $\sqrt{n}\hat{\gamma}$. This immediately shows the connection to the weighted FIC, with weights equal to the logistic variances. Of course, the weighted FIC is able to incorporate other types of weights as well. In the next section, we illustrate this by example on the WESDR dataset, where we choose equal weighting for all female subjects in the dataset. An example of robust downweighting is shown on Hoffstedt's highway data in Section 7.

6 Weighted FIC model selection for the dataset

To obtain an overall best model according to FIC for the subset of women in the study, we compute for each subset S, wFIC(S) with weight vector $(1/n_F)I(\text{female})$ where the weights are indicator variables for women and n_F denotes the number of women in the dataset. It is important to note that the values of wFIC(S) are computed using the complete dataset, we are not splitting the dataset for model selection. Figure 3 shows the 20 smallest (best) wFIC values. There is not much difference between the first two values, which correspond to the models: (i) including variables z_1, z_4 and z_6 for the best model, and (ii) including variables z_1, z_4, z_6 and z_7 for the second best model. The best model corresponds to model number 54 in Figure 1, which was preferred 13 times as individually best model. This particular model is also the one chosen via a –not focussed– AIC model search.

7 Downweighting outlying observations, an example

As an illustration we use Hoffstedt's highway data, included in R's library(alr3), see also Weisberg (2005, Section 7.2). This dataset is used to explain the 1973 accident rate per million vehicle miles, as a function of several variables. There are 39 observations. In every model we include an intercept term and x_1, the length of the highway segment in miles. Other

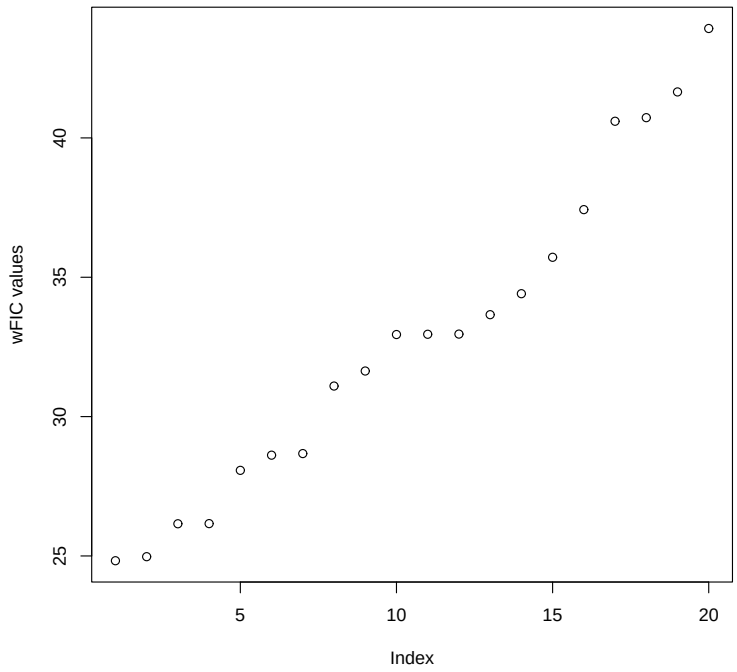

FIGURE 3. The 20 smallest wFIC values for the WESDR dataset where the weights are indicator variables for the subset of women. This is global model search.

variables are z_1: average daily traffic count in thousands, z_2: truck volume as a percent of the total volume, z_3: total number of lanes of traffic, z_4: number of access points per mile, z_5: number of signalised interchanges per mile, z_6: number of freeway-type interchanges per mile, z_7: speed limit in 1973, z_8: lane width, in feet, z_9: width of the outer shoulder on the roadway (in feet), and finally z_{10}: an indicator of the type of roadway or the source of funding for the road.

We first fit the full model using the method of M-estimation using Huber's psi function. This gives a set of residuals. Figure 4 shows the plot of these residuals against the fitted values. We identify five points where the absolute value of the residual is larger than 1.345. This particular cut-off value is also used by Ronchetti and Staudte (1994). We now define

$$w(x_i, z_i) = \begin{cases} 1 & \text{if } |\text{residual}_i| \leq 1.345 \\ 1.345/|\text{residual}_i| & \text{if } |\text{residual}_i| > 1.345, \end{cases}$$

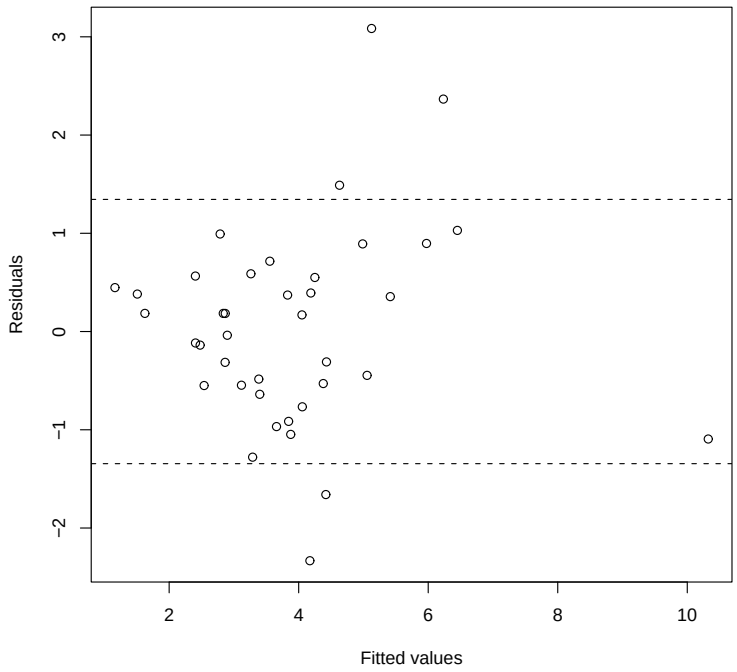

FIGURE 4. Residuals obtained after a robust full model fit, versus predicted values, for Hoffstedt's highway data.

and use these as weights to perform a global model search. The five identified residuals, outside the (-1345, 1.345) range in Figure 4 lead to weights 0.903, 0.568, 0.436, 0.577, and 0.811. All other observations get weight equal to one. This approach is effectively downweighting these 5 influential observations.

Figure 5 shows the 50 smallest wFIC values. The best value corresponds to the model including x_1 and z_4.

Based on robust C_p model selection, Ronchetti and Staudte (1994) support the model which includes, in addition to x_1, the variables z_5, z_6, z_7 and z_{10}, and also the model with additional variables z_2, z_3, z_4 and z_9. For this example, the model chosen by wFIC is more parsimonious.

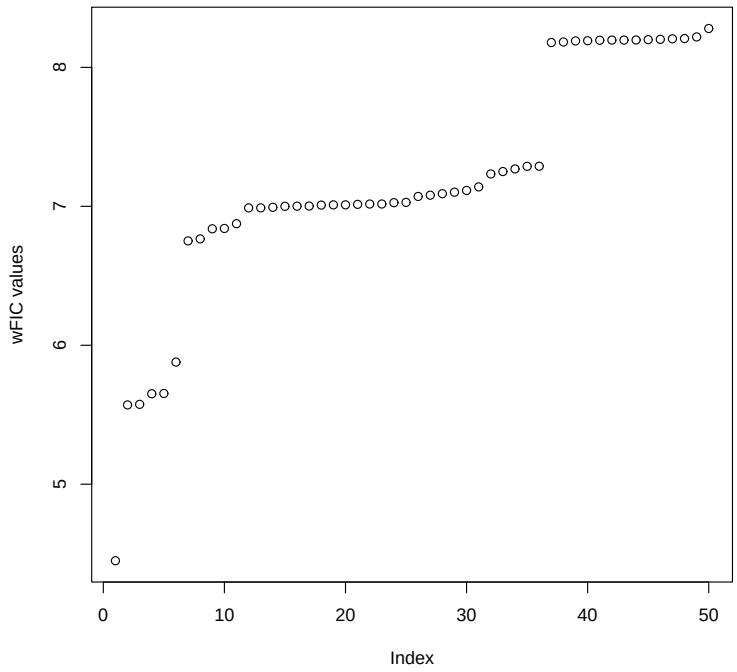

FIGURE 5. The 50 smallest wFIC values for Hoffstedt's highway data. The weights are based on Huber's psi function, downweighting 5 observations. This is global model search.

8 Remarks

It is worth investigating whether an individual model search is useful in identifying "outlying" observations. Outlying can here be understood in the sense of pointing to quite different models than the majority of the other observations.

When outlying observations are identified, they can be downweighted in a weighted model search. Also this approach needs some more investigation.

Claeskens, Croux and Van Kerckhoven (2006) construct a different type of (subjectwise) focussed information criterion. Instead of selecting a model with the goal of minimising the MSE, other FIC expressions are constructed which minimise the prediction error. This strategy is useful to direct model selection even more to a specific task.

References

Akaike, H. (1974). A new look at statistical model identification, *I.E.E.E. Transactions on Automatic Control* **19**, 716–723.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction focussed information criterion. *Biometrics*, to appear.

Claeskens, G., and Hjort, N. L. (2003). The focused information criterion [with discussion]. *Journal of the American Statistical Association*, **98**, 900–916.

Claeskens, G., and Hjort, N. L. (2006). Minimising average risk in regression models. *Submitted.*

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., and DeMets, D. L. , (1984). The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology*, **102**, 520–526.

Ronchetti, E., and Staudte, R. G (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association* **89**, 550–559.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edition. Wiley, New York.