Generalized Additive Models

The Model

The GLM is:

$$g(\mu) = ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_kx_k$$

The generalization to the GAM is:

$$g(\mu) = ß_0 + f_1(x_1) + f_2(x_2) + \ldots + f_k(x_k)$$

where the functions $f_i(x_i)$ are arbitrary functions defined by the data.

Simple Additive Models

- $Y = ß_0 + f_1(x_1) + f_2(x_2) + \ldots + f_k(x_k) + e$

- e is independent of x_i and $E(e) = 0$
- Y is continuous
- $var(Y) = sigma^2$, for all observations (homoscedasticity)

Assumptions

- Statistical independence of observations
- Variance function is specified correctly
- Correct link function
- Specific observations do not influence fit

Smoothers Available in Splus

- Loess (locally weighted regression)
- Splines (regression, cubic, natural)

Properties of Smoothers

- Most smoothers are local, in the sense that they use adjacent data points (neighbourhoods) to estimate "predicted" values for each data point

- One must balance bias with precision
  - The choice of how much smoothing to do is key in this decision process

- This is no different than when you specify different parametric forms for an explanatory variable, in that one is trying to specify the correct functional form.
- NB: A linear variable is equivalent to an infinitely smoothed function!

- The functions $f_i(x_k)$ can be very general and can include:
  - splines or LOESS
    - » interactions such as $lo(x_1,x_2)$ can be fit and these produce smooth two-dimensional surfaces in three dimensions
  - parametric forms, such as $E(Y) = ß_0 + ßx$, where x can be continuous, ordinal, nominal

Estimation in GAMs

- Estimation is through a combination of backfitting and iteratively reweighted least squares

- The method is not maximum likelihood but is based on similar types of principals.

- The functions f_i and $ß_k$ are determined empirically according to the data and the assumed model.

- The deviance is calculated from the model, just as in GLMs.

Fitting

- Usual linear model is fit with least squares and there is an exact solution (no iterations).
- *Backfitting algorithm* used for GAMs, and it requires >1 iteration.

Backfitting Algorithm

$Y = ß_0 + f_1(x_1) + f_2(x_2) + ... + f_k(x_k) + e$

1) Set $ß_0 = \text{mean}(Y)$

2) Initialize $f_j(x_j) = f_j(x_j)^0$

3) Iterate and cycle over the k variables

   $f_j = S_j(Y - ß_0 - \sum_{k \neq j} f_k \mid x_j)$

   until the f_j do not change

Goodness-of-fit

- Deviance is defined in the same way as in GLMs.
- The comparison of nested GAMs by substracting deviances does not necessarily follow a χ^2 distribution, even asymptotically.
- However, one can use the chi-square distribution as sort of a reference for assessing fits.

- However, approximately, E(Deviance) ~ residual df * phi.

- For nested models, $E(D_1, D_2) \sim df_1 - df_2$, implying that the χ^2 distribution on $df_1 - df_2$ degrees of freedom can be used.

- In practice, we use instead a **penalized version** of the deviance for comparing both nested and non-nested models.

- The penalty is proportional to the number of df used.

Aikaike Information Criterion

- AIC = Deviance + 2 * df_{model} * phi
- This statistic accounts for the number of degrees of freedom used by the smoothers. Usually, a lower AIC implies that the model fits better than another.
- There is no specific statistical test associated with comparing AICs.
- NB: must have the same number of observations in the two models.

Confidence Intervals

- Pointwise standard errors of the functions f_j are also calculated.
- Calculation of the confidence interval between two values of x is more difficult. For example, the 95% CI for the odds ratio between $x=x_1$ and $x=x_2$ in a model $\text{logit}(y) = ß_0 + f(x)$ must be obtained using the bootstrap.

Generalized Additive Model

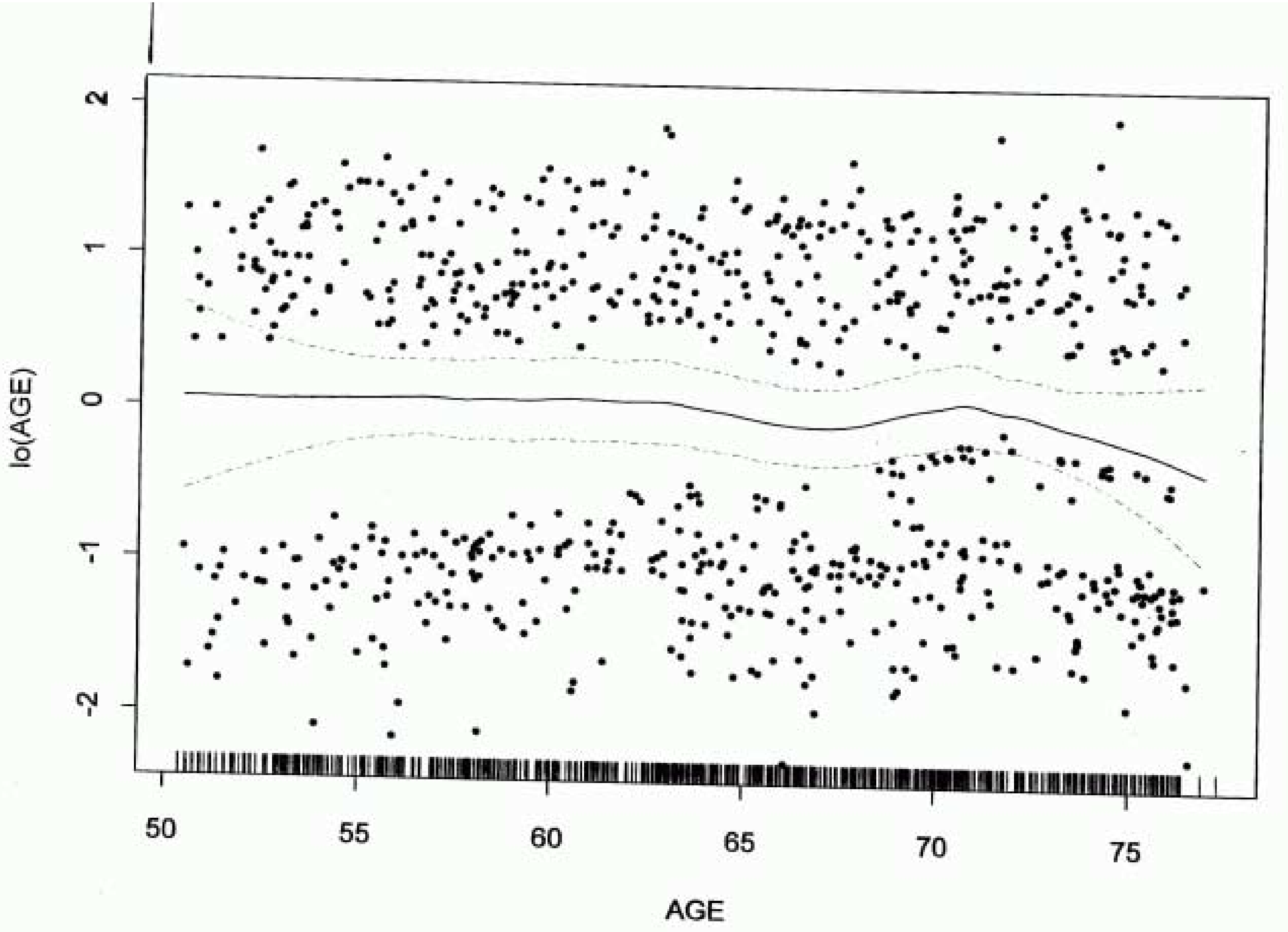
Age at menopause

```
> preg4.gam.1 _ gam(YVAR~lo(AGE)+lo(EDUCTN)+lo(AGE.FST)+lo(MG13.FIR)+lo(M18TO21.)
    +FAM.HIST+MG34.ALC+MG23.MAL,preg4,family=binomial,subset=(AGE>50),
    na.action=na.omit,x=T,y=T)
> summary(preg4.gam.1)

Call: gam(formula = YVAR ~ lo(AGE) + lo(EDUCTN) + lo(AGE.FST) +
       ….
Deviance Residuals:
       Min         1Q    Median        3Q      Max
 -2.192961 -0.9900211 0.4602939 0.9577022 2.138613

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 1007.108 on 726 degrees of freedom

Residual Deviance: 837.6898 on 704.0405 degrees of freedom

Number of Local Scoring Iterations: 4

DF for Terms and Chi-squares for Nonparametric Effects

             Df Npar Df Npar Chisq    P(Chi)
 (Intercept)  1
     lo(AGE)  1     2.4    2.53375 0.3475057
  lo(EDUCTN)  1     2.3    4.85548 0.1171361
 lo(AGE.FST)  1     2.9    4.53878 0.1975979
lo(MG13.FIR)  1     2.8    2.56048 0.4273571
lo(M18TO21.)  1     2.6   26.30268 0.0000045
    FAM.HIST  1
    MG34.ALC  2
    MG23.MAL  1
```

Age at menarche

<u>Npar Chisq</u> : Score test to evaluate the nonlinear contribution to the nonparametric functions.

Adjusted GAM Model



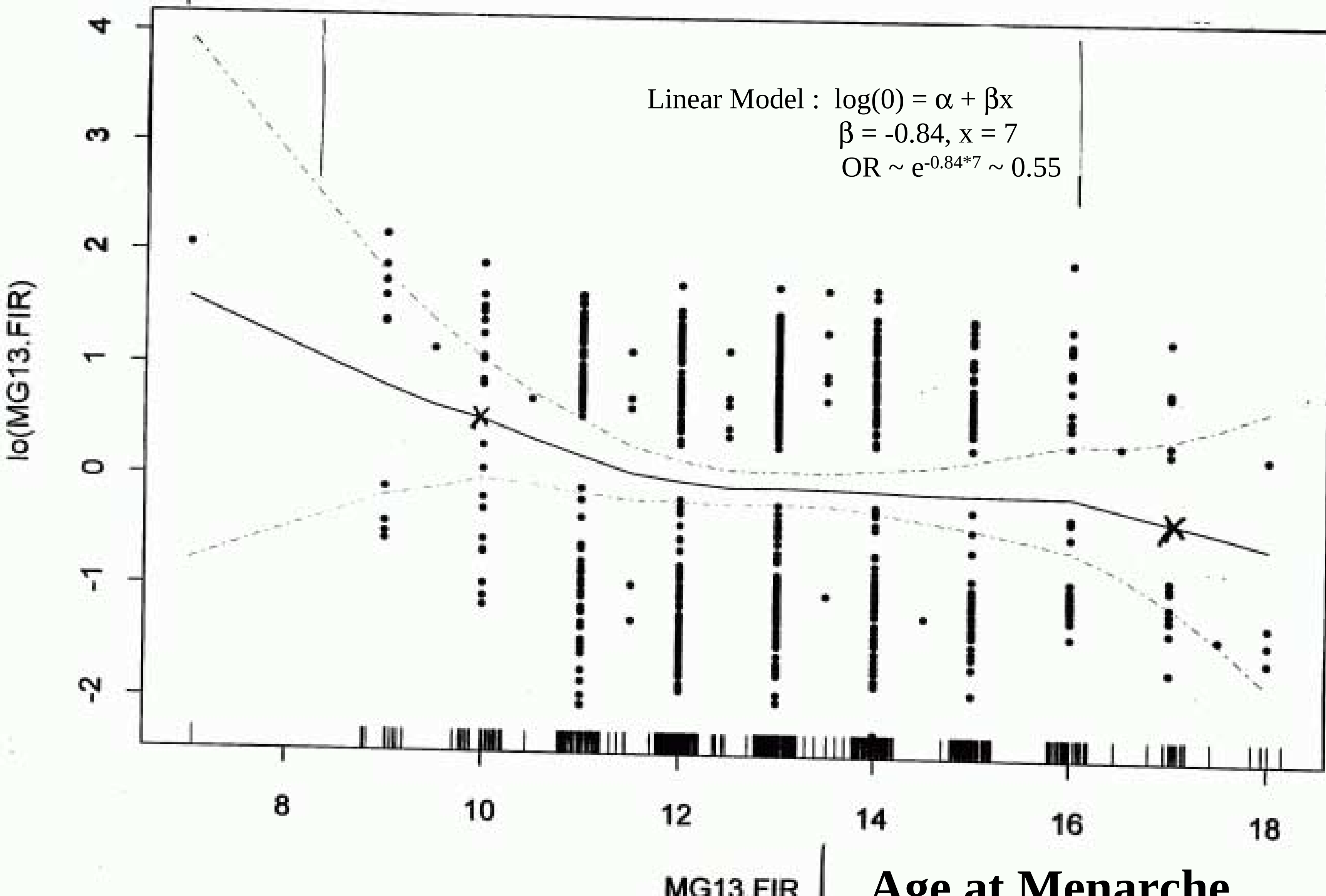
VII-43

Ordinate ~ log(odds scale) $\qquad \log(0) = \alpha + f(x)$

at $x = 10$, $f(x) \sim 0.5$

at $x = 17$, $f(x) \sim 0.1$

$\Delta f(x) \sim -0.6 \;\Rightarrow\;$ odds ratio $\sim e^{-0.6} \sim 0.55$



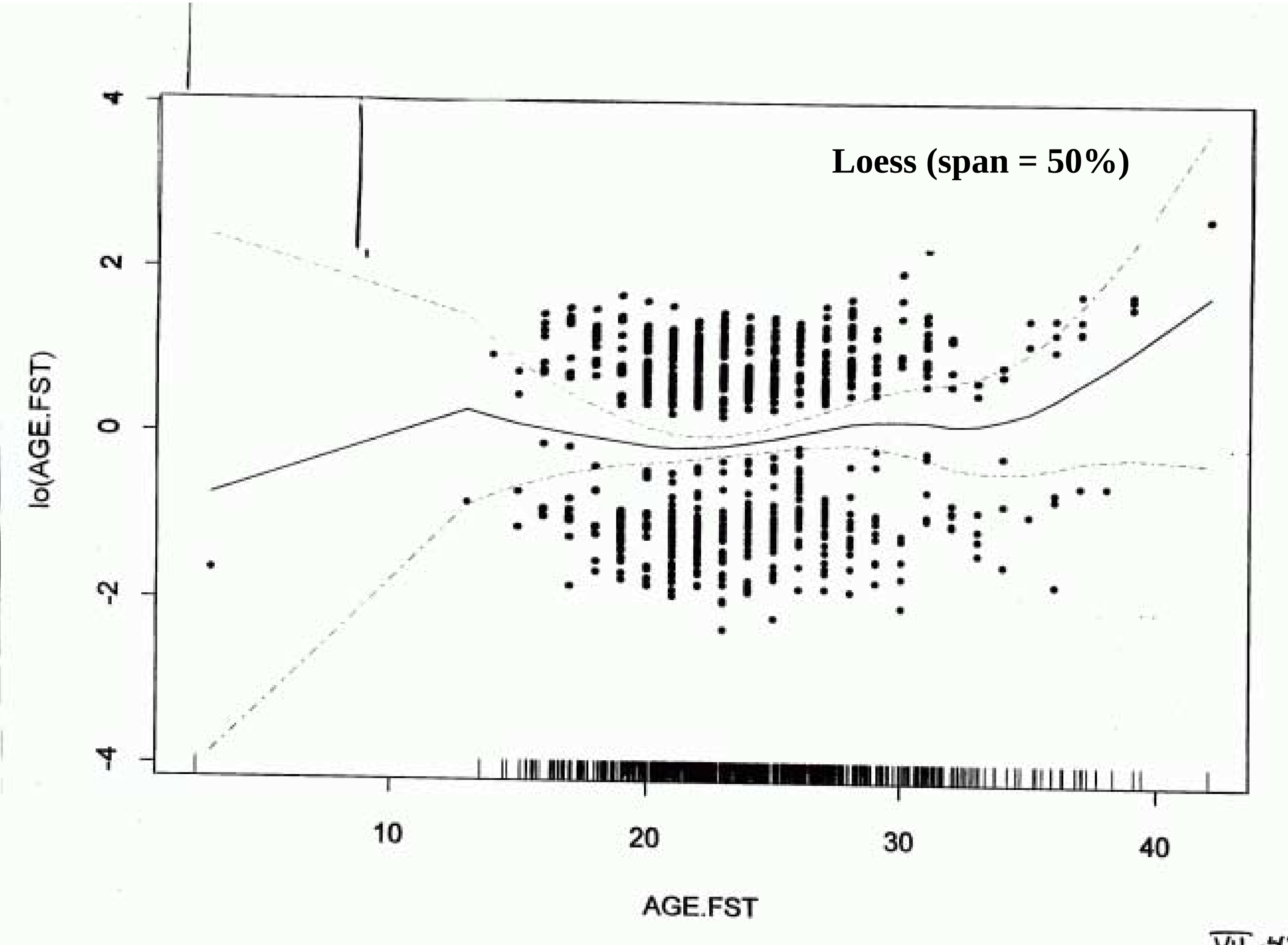
Linear Model : $\log(0) = \alpha + \beta x$

$\beta = -0.84$, $x = 7$

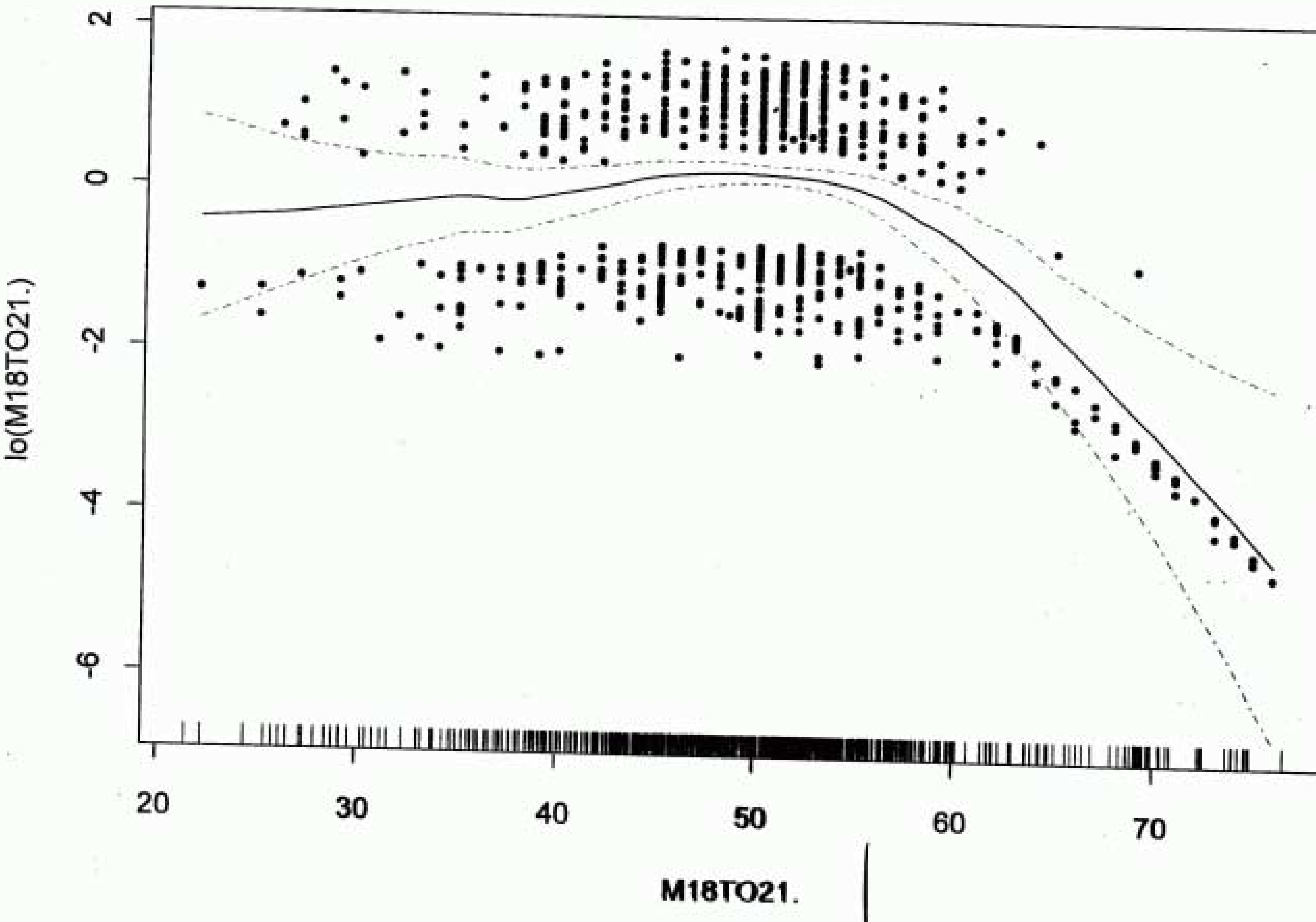
$OR \sim e^{-0.84*7} \sim 0.55$

Age at Menarche

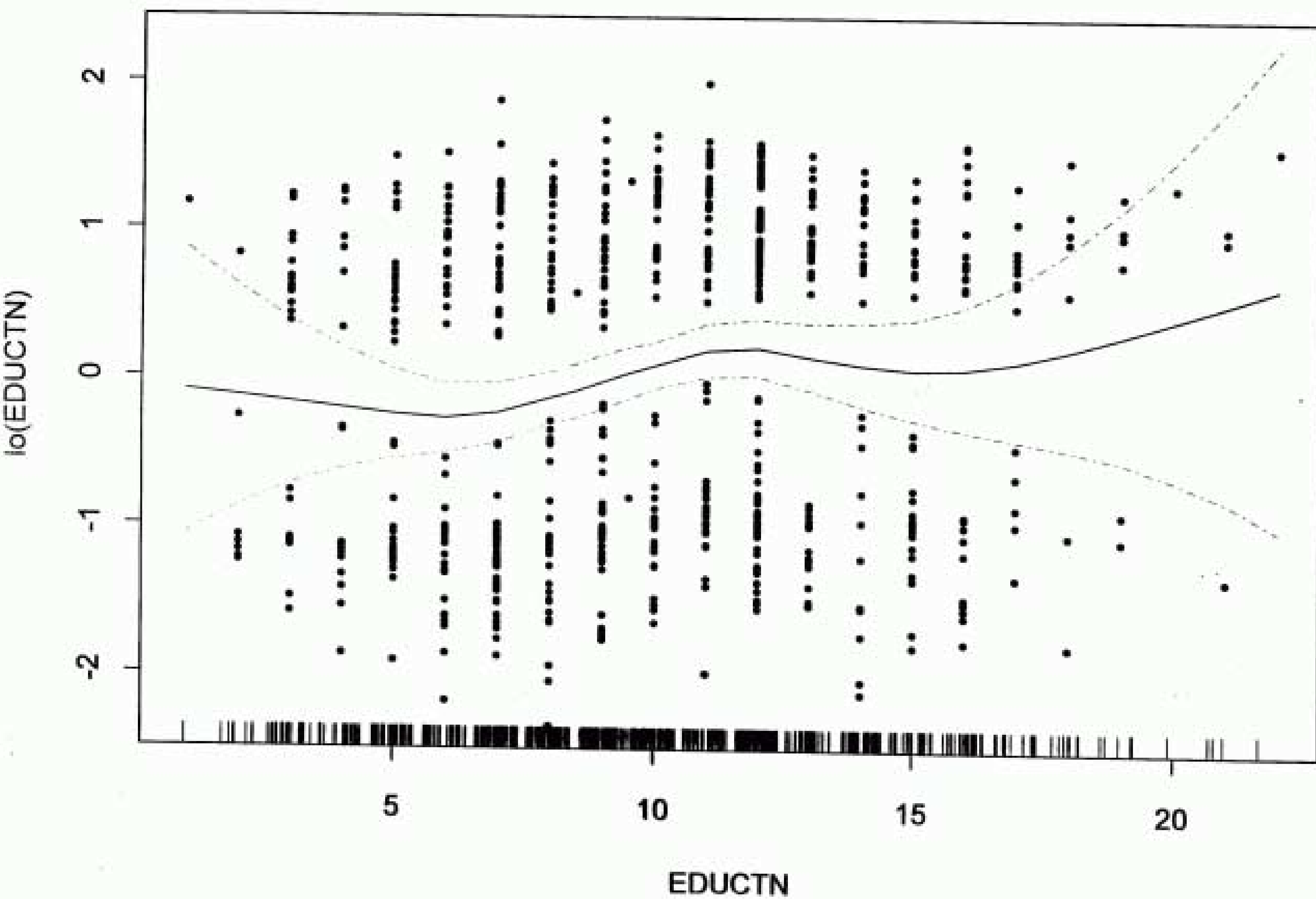
Age at 1st birth

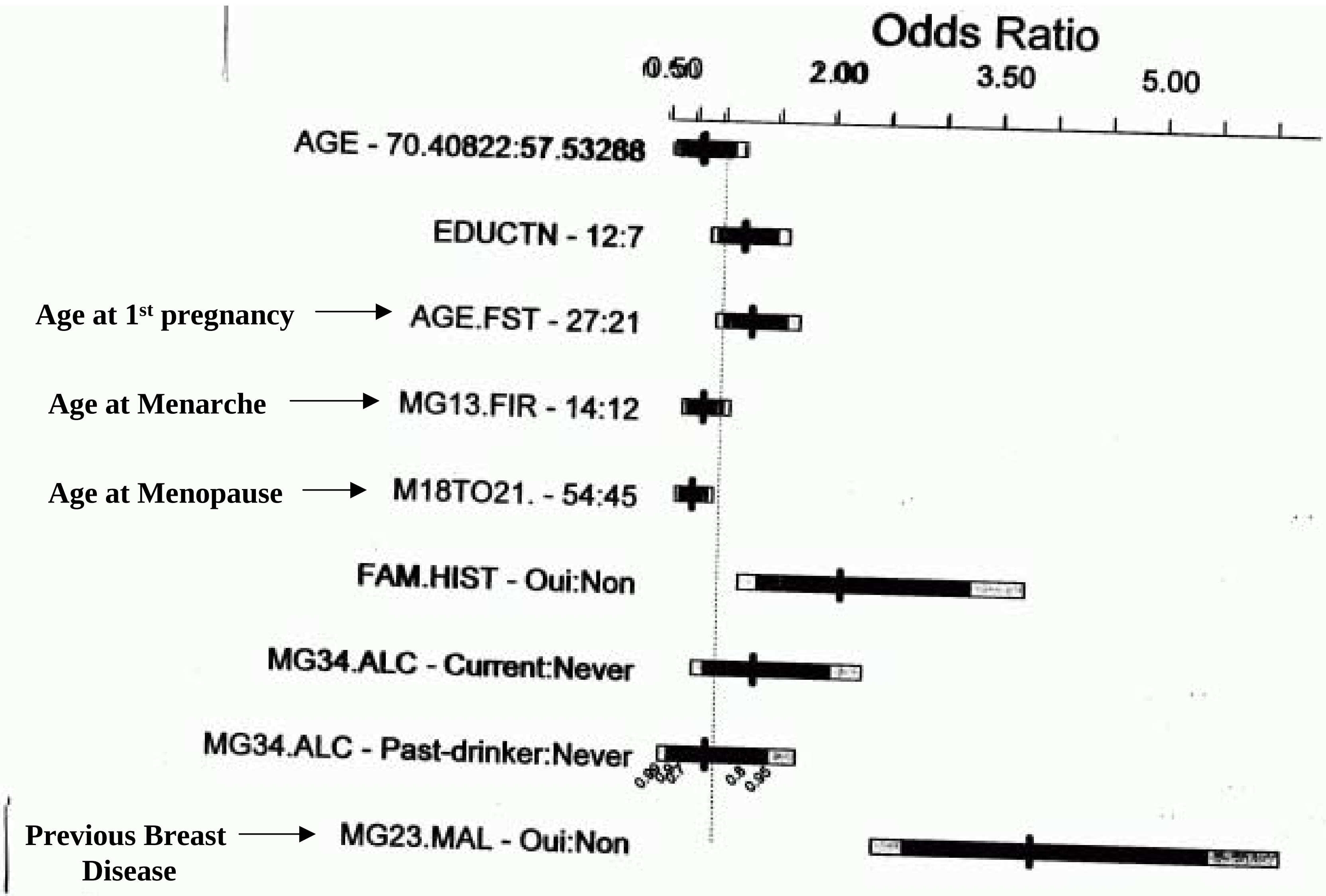


Age at Menopause



Education (in years)





Odds Ratio
0.50
2.00
3.50
5.00
AGE - 70.40822:57.53288
EDUCTN - 12:7
Age at 1st pregnancy
AGE.FST - 27:21
Age at Menarche
MG13.FIR - 14:12
Age at Menopause
M18TO21. - 54:45
FAM.HIST - Oui:Non
MG34.ALC - Current:Never
MG34.ALC - Past-drinker:Never
Previous Breast Disease
MG23.MAL - Oui:Non