# STA2201H1S
# END OF TERM TEST

## April 11, 2014, 2.00 – 5.00 pm

**Answer all four questions in examination booklets.**
**Each question is worth 25 marks.**

---

1. The data shown in Figure 1 is the "Challenger Data" given in Chapter 1 of SM. It shows the number of O-rings damaged in each of 23 shuttle launches, and the temperature of the launch. O-rings are rubber insulating rings that plug the joints in the fuel system. There are 6 O-rings on each launch rocket, and there was some evidence from bench testing that O-ring damage was associated with lower temperatures, causing the O-rings to be less flexible. This data was discussed prior to the launch of the Challenger on 28 January 1986, because the predicted temperature for the launch date was 31°F, and there was uncertainty about whether or not O-ring damage was associated with temperature.

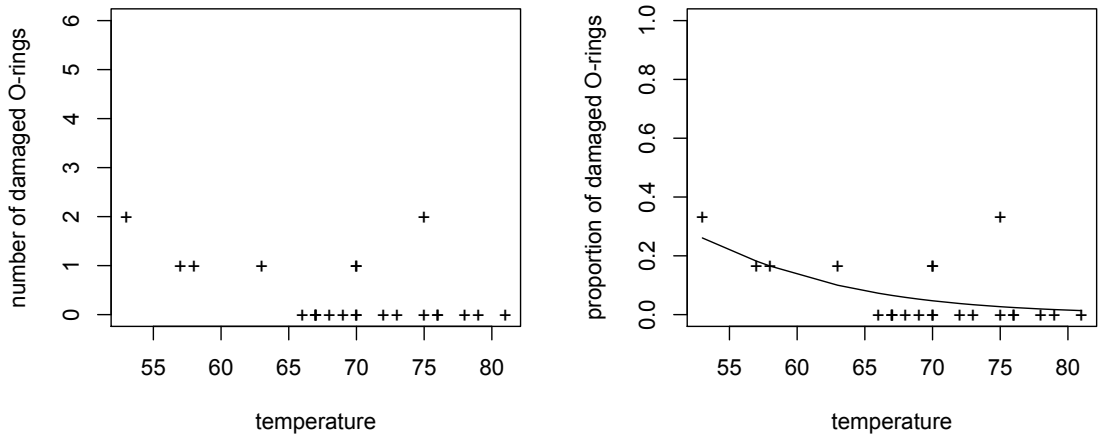


Figure 1: Left panel shows number of damaged O-rings, as a function of temperature. Right panel shows proportion of damaged O-rings, and fitted model `shuttle.glm`.

(a) Based on the output from `shuttle.glm`:

i. Writing $R_i$ for the number of damaged O-rings on launch $i$, give an expression for the model for $R_i$ used in analyzing this data. Give both the form of the linear predictor, and the probability density function. What independence assumptions are implicit in the model used in this question?

ii. what is the estimated effect of a decrease of 1°F on the probability of O-ring damage? Is there evidence of over-dispersion in the data, relative to the model fitted? Explain.

iii. What is the predicted number of damaged O-rings at 31°F? Describe briefly how you might estimate a standard error for that prediction.

(b) The model `shuttle.glm2` includes pressure as a covariate. What is the $p$-value for a log-likelihood ratio test of the effect of pressure on the probability of O-ring damage?

(c) The original analysis of the data in 1986 omitted all the launches with no damaged O-rings, on the grounds that these observations did not provide any evidence about the link between damage and temperature. If these points were omitted, what would be the apparent relationship between O-ring damage and temperature?

```
> library(SMPracticals)
Loading required package: ellipse
> data(shuttle)
> head(shuttle)

   m r temperature pressure
1  6 0          66       50
2  6 1          70       50
3  6 0          69       50
4  6 0          68       50
5  6 0          67       50

> shuttle.glm <- glm(cbind(r,m-r) ~ temperature, family = binomial,
+ data = shuttle)

> summary(shuttle.glm)

Call:
glm(formula = cbind(r, m - r) ~ temperature, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95227  -0.78299  -0.54117  -0.04379   2.65152

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 18.086  on 21  degrees of freedom
AIC: 35.647

Number of Fisher Scoring iterations: 5

> betahat <- coef(shuttle.glm)
> exp(betahat[1] +31*betahat[2])/(1+exp(betahat[1] +31*betahat[2]))
(Intercept)
  0.8177744

> shuttle.glm2 <- glm(cbind(r,m-r)~temperature + pressure, family = binomial
+ data = shuttle)

> anova(shuttle.glm2,shuttle.glm)
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ temperature + pressure
Model 2: cbind(r, m - r) ~ temperature
  Resid. Df Resid. Dev Df Deviance
1        20     16.546
2        21     18.086 -1  -1.5407

> pchisq(1.5407, df=1, lower.tail = F)
[1] 0.2145136
```

(a) i. The model is

$$\prod_{i=1}^{23} p(r_i;\pi_i) = \prod_{i=1}^{23} \binom{6}{r_i} \pi_i^{r_i}(1-\pi_i)^{6-r_i}, \quad i = 1,\ldots,23; 0 \le r_i \le 6,$$

with

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 temp_i.$$

This model assumes that the 23 individual launches are independent, and also that the damage to each of the six seals on a given launch are independent Bernoulli's, with the same probability. The second independence assumption seems more suspect.

ii. This question wasn't worded very well, because this cannot be assessed directly without using $\hat{\beta}_0$; the answer "a 1 degree (F) increase in temperature is associated with a decrease in the log-odds of O-ring damage of 0.1156, or a decrease in the odds of damage of $\exp(-0.1156)$" was quite acceptable, but more detailed answers were accepted as well. There is not evidence in the `R` output of over-dispersion; the residual deviance is 18 on 21 degrees of freedom.

iii. The predicted number of failures at 31°F is $6*\exp(5.08-0.1156*31)/\{1+\exp(5.08-0.1156*31)\} = 6*0.82 = 4.92$. The estimated standard error could be computed using the delta-method, or we could compute a confidence interval for $\hat{\beta}_0 + \hat{\beta}_1 * 31$

using $\text{var}(\hat{\beta}_0+\hat{\beta}_1*31) = \text{var}(\hat{\beta}_0)+31^2\text{var}(\hat{\beta}_1)+2*31\text{cov}(\hat{\beta}_0,\hat{\beta}_1)$, the entries of which are all available using `vcov`, and convert the endpoints of this confidence interval to endpoints for a confidence interval for $\hat{\pi}(31)$. We could also use `predict.glm, se=TRUE`.

(b) The $p$-value is 0.2145.

(c) After deleting the points with 0 failures, the plot of number of failures against temperature shows either no trend (if a linear model is fit), or a quadratic trend (if the model is expanded to include this). It omits most of the evidence that low temperature is associated with higher probability of failure, mainly because of the single launch at 75°F with 2 failures.

This example is discussed in SM, and in many other books, including Faraway, and Maindonald and Braun. The original statistical discussion was given in Fowlkes, et al in *JASA*. The launch that took place at 31°F on 28 January is the "Challenger disaster"; there are spectacular images available via Wikipedia. The data analysis undertaken before the launch did apparently ignore the launches with zero failures. Tufte (www.tufte.com) has also blamed poor communication caused in part by the use of Powerpoint Slides for the poor decisions made before the launch. The three books above, and the associated `R` files, all seem to use slightly different data sets. The original data is available on-line in the report of the presidential commission.

2. The gamma density is given by

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)} y^{\nu-1}\left(\frac{\nu}{\mu}\right)^{\nu}\exp(-\nu y/\mu), \tag{1}$$

where $\text{E}(y)=\mu$ and $\text{var}(Y)=\nu^{-1}\mu^2$.

(a) Show that this density has the form of a generalized linear model, and identify the canonical parameter $\theta$ and the scale parameter $\phi$ in terms of $\mu$ and $\nu$.

(b) Give an expression for the maximum likelihood estimate of $\nu$, based on an i.i.d. sample of size $n$ from (1), using the notation $\psi(\nu)$ for the *digamma* function, $d\log\Gamma(\nu)/d\nu$.

(c) Suppose now the sample $(y_1,\ldots,y_n)$ is independent, but $\text{E}(y_j)=\mu_j$, where $1/\mu_j = x_j^{\text{T}}\beta = \eta_j$, with $\beta=(\beta_1,\ldots,\beta_p)$, $p<n$. Explain the motivation for the estimator

$$\tilde{\phi} = \tilde{\nu}^{-1} = \frac{1}{n-p}\sum_{j=1}^{n}\frac{(y_j-\mu_j(\hat{\beta}))^2}{\mu_j(\hat{\beta})}.$$

(a) The general form for a GLM is

$$f(y;\theta,\phi) = \exp\{\frac{y\theta+b(\theta)}{\phi}+c(y,\phi)\},$$

and the gamma density is

$$f(y;\mu,\nu) = \exp\{\frac{-\frac{1}{\mu}y}{\frac{1}{\nu}}+\frac{\log(\frac{\nu}{\mu})}{\frac{1}{\nu}}-n\log\Gamma(\nu)+(\nu-1)\log y,$$

giving

$$\theta = -\frac{1}{\mu}, \quad b(\theta) = -\log\mu = -\log(-1/\theta), \quad \phi = \frac{1}{\nu}, c(y,\phi) = \nu\log\nu - n\log\Gamma(\nu) + (\nu-1)\log y.$$

(b) The log-likelihood function based on a sample of size $n$ is

$$\ell(\mu,\nu;y) = \sum_{j=1}^{n}(\nu\log y_j - \nu\log\mu - \nu y_j/\mu) - n\log\Gamma(\nu) + n\nu\log\nu,$$

leading to $\hat{\mu} = \bar{y}$, and thence to

$$\ell_{\mathrm{p}}(\nu) = n\nu\log\nu - n\log\Gamma(\nu) + \nu\Sigma(\log y_j - \log\bar{y} - 1),$$

and thence to

$$\log\hat{\nu} - \psi(\hat{\nu}) = \log\bar{y} - \Sigma\log y_j/n.$$

(c) This can be justified as an approximation to the maximum likelihood estimator, as in the HW, with an adjustment for degrees of freedom, or, can be motivated by noting that

$$\mathrm{E}\sum_{j=1}^{n}\frac{(y_j-\mu_j)^2}{\mu_j^2} = \sum_{j=1}^{n}\frac{\mathrm{var}(y_j)}{\mu_j^2} = n\nu^{-1},$$

and estimating $\mu_j$ by $\hat{\mu}_j$, and adjusting the divisor accordingly.
The justification "it's the Pearson $\chi^2$ estimator of $\phi$ proposed in the textbook" was also acceptable.

3. (See Figure 2): An often-used data-set in statistics textbooks is the `ozone` data, which gives daily average readings of ozone (O3) in Los Angeles, along with data on meteorological variables. A linear model of ozone (`O3`) on temperature (`temp`), inversion base height (`ibh`) and inversion base temperature (`ibt`) gave the following:

```
Call:
lm(formula = O3 ~ temp + ibh + ibt, data = ozone)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3224  -3.1913  -0.2591   2.9635  13.2860

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.7279822  1.6216623  -4.765 2.84e-06 ***
temp         0.3804408  0.0401582   9.474  < 2e-16 ***
ibh         -0.0011862  0.0002567  -4.621 5.52e-06 ***
ibt         -0.0058215  0.0101793  -0.572    0.568
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.748 on 326 degrees of freedom
Multiple R-squared:  0.652,     Adjusted R-squared:  0.6488
F-statistic: 203.6 on 3 and 326 DF,  p-value: < 2.2e-16
```

The plot of residuals against fitted values had some anomalies, so a more flexible model was fit:

$$O3 = s(temp) + s(ibh) + s(ibt) + error$$

leading to

```
> summary(ammgcv)

Family: gaussian
Link function: identity

Formula:
O3 ~ s(temp) + s(ibh) + s(ibt)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7758     0.2382   49.44   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
          edf Ref.df      F  p-value
s(temp) 3.386  4.259 20.553 7.69e-16 ***
s(ibh)  4.174  5.076  7.338 1.38e-06 ***
s(ibt)  2.112  2.731  1.612    0.187
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

R-sq.(adj) =  0.708   Deviance explained = 71.7\%
GCV score = 19.346  Scale est. = 18.72     n = 330
```

(a) How many extra degrees of freedom are used fitting the second model, compared to the linear regression?

(b) In the plot of `s(temp)` from `mgcv` it appears that there may be a change in the slope at about `temp` = 65. How might you use `mgcv` to test for the presence of a change point?

(c) In the plots of the smooth functions of temperature and of `ibt`, why do the standard errors increase at the ends of the range of values? This is not the case however for the smooth function of `ibh`: what is a possible explanation?

(a) $1 + 3.4 + 4.2 + 2.1 = 10.7 \approx 11$ degrees of freedom were used in the smooth fit, and 4 were used in the linear fit, of a difference of $6.7 \approx 7$.

(b) Could fit model A, with linear predictor `s(temp) + s(ibh) + s(ibt)` and model B, with linear predictor `temp + s(ibh) + s(ibt)`, and compare the residual deviances of the two models with an (approximate) $F$-statistic

$$\frac{\text{change in resid deviance}/(3.4-1)}{\text{resid deviance from B}/(\text{resid df})}.$$

Or, could fit a "hockey stick" model with linear predictor $\alpha_0 + \alpha_1 x, x < 65$ and $\beta_0 + \beta_1 x, x > 65$, with a constraint to make it continuous at $x = 65$.

(c) The point wise confidence intervals are wider at the endpoints because they are based on less data. With `ibt`, there is actually more data at the right hand endpoint, which suggests that the observed values were truncated, either due to instrument error or instrument limitations.

4. The abstract for a study[1] investigating biomarkers and mortality is reproduced in Figure 3. As explained in the editorial that accompanied the study, "Biomarkers are biological molecules found in blood, body fluids, or tissues that may signal an abnormal process, a condition, or a disease. Most current biomarkers are used to test an individuals risk of developing a specific condition. There are none that accurately assess whether a person is at risk of ill health generally, or likely to die soon from a disease."

   (a) Was this study a survey, an observational study, or an experiment? Explain.

   (b) The response variable was survival time (as measured by age), and the potential explanatory variables were levels of 106 different biomarkers, as well as other explanatory variables associated with mortality, including sex, cholesterol, smoking, prevalent cancer, prevalent heart disease and prevalent cancer. The model was fit with the Estonian data-base, and then validated in the Finnish data-base. In selecting biomarkers to fit the model, a cut-off of $p = 0.0005$ was used. Why did the researchers select such a small $p$-value to identify 'significant' biomarkers? Why are the other explanatory variables included in the final model?

   (c) The "biomarker summary score" for each individual in the study was computed as $\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{\beta}_4 x_{4i}$, where $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$ is the vector of standardized biomarker measures for individual $i$, and $\hat{\beta}$ is estimated by proportional hazards regression of failure age on $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$ and several other explanatory variables. Figure 4 shows a plot of this summary score, as a function of age. The authors say "The biomarker score was moderately correlated with age ($r = 0.38$), yet extreme biomarker score values were seen across all age groups. Excess mortality within 5 y of follow-up was observed for higher age, but in particular in combination with an elevated biomarker score". Summarize the information from the plot and text excerpt, in language suitable for explaining the main results in, for example, *The Varsity*.

   (d) The proportional hazards model takes the form

$$h(t; x) = h_0(t) \exp(x^{\mathrm{T}} \beta),$$

   where $h(t; x) = f(t; x)/\{1 - F(t; x)\}$ is the hazard function, or instantaneous failure rate, at time $t$ for an individual with covariates $x$. Show that

$$1 - F(t; x) = \{1 - F_0(t)\}^{\exp(x^{\mathrm{T}}\beta)}.$$

   What graphical check of the model does this result suggest?

1. This was a (prospective) observational study: subjects were enrolled in a cohort study, blood samples were taken at entry, and mortality assessed in a five-year window.

2. The researchers used a small $p$-value because they did 106 tests. Dividing the usual cut-off, 0.05, by 100, is an approximate correction so that the global level of significance is 0.05. This is known as the Bonferroni correction. The other explanatory variables are included in the model because they are highly predictive of death, and if omitted the impact of the biomarkers might be considerably exaggerated.

3. Because older people are more likely to die, it is important to check whether or not the new blood tests are simply reflecting the individuals' age. While there is evidence that the blood test 'biomarker score' does increase with age, there is additional evidence that high biomarker scores are associated with increased probability of death at any age.

4. (a)

$$H(t;x) = \int_0^t h(u;x)du = \int_0^t \frac{f(u;x)du}{1-F(u;x)} = -\log\{1-F(t;x)\} = H_0(t)\exp(x^{\mathrm{T}}\beta),$$

so we have

$$\log\{1-F(t;x)\} = \log\{1-F_0(t)\}\exp(x^{\mathrm{T}}\beta), \quad 1-F(t;x) = \{1-F_0(t)\}^{\exp(x^{\mathrm{T}}\beta)}.$$

(b) We could group observations according to their value of $x$ (if $x$ is continuous we could create groups of 'similar' $x$'s), and estimate the survivor function non-parametrically within each group. Plots of these survivor functions should not cross, if the proportional hazards assumption is valid. A plot of this type appears in SM, in Figure 10.22.

[1]Fischer K, Kettunen J, Würtz P, Haller T, Havulinna AS, et al. (2014) Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality. *PLoS Med* 11(2): e1001606. doi:10.1371/journal.pmed.1001606

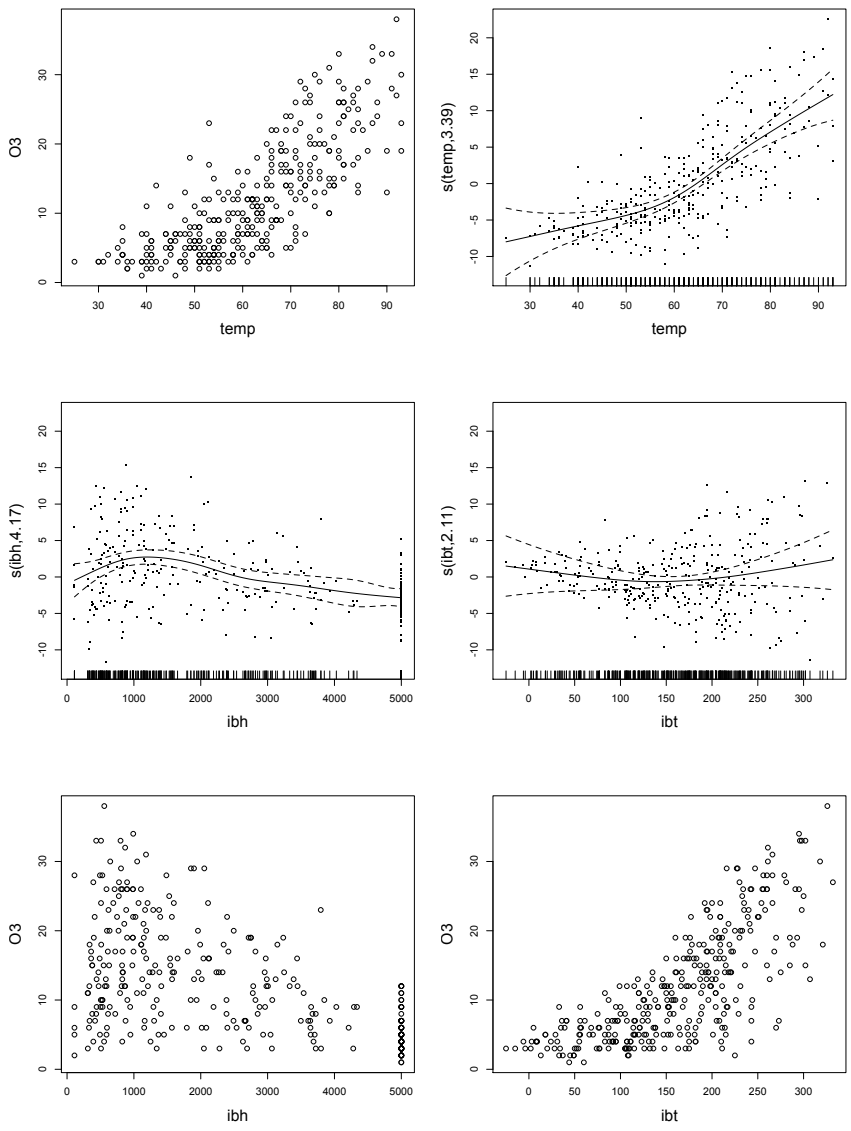


Figure 2: Ozone data plotted against temperature, inversion base height (ibh), and inversion base temperature (ibt), along with fitted smooth functions.

## Abstract

***Background:*** Early identification of ambulatory persons at high short-term risk of death could benefit targeted prevention. To identify biomarkers for all-cause mortality and enhance risk prediction, we conducted high-throughput profiling of blood specimens in two large population-based cohorts.

***Methods and Findings:*** 106 candidate biomarkers were quantified by nuclear magnetic resonance spectroscopy of non-fasting plasma samples from a random subset of the Estonian Biobank ($n = 9{,}842$; age range 18–103 y; 508 deaths during a median of 5.4 y of follow-up). Biomarkers for all-cause mortality were examined using stepwise proportional hazards models. Significant biomarkers were validated and incremental predictive utility assessed in a population-based cohort from Finland ($n = 7{,}503$; 176 deaths during 5 y of follow-up). Four circulating biomarkers predicted the risk of all-cause mortality among participants from the Estonian Biobank after adjusting for conventional risk factors: alpha-1-acid glycoprotein (hazard ratio [HR] 1.67 per 1–standard deviation increment, 95% CI 1.53–1.82, $p = 5 \times 10^{-31}$), albumin (HR 0.70, 95% CI 0.65–0.76, $p = 2 \times 10^{-18}$), very-low-density lipoprotein particle size (HR 0.69, 95% CI 0.62–0.77, $p = 3 \times 10^{-12}$), and citrate (HR 1.33, 95% CI 1.21–1.45, $p = 5 \times 10^{-10}$). All four biomarkers were predictive of cardiovascular mortality, as well as death from cancer and other nonvascular diseases. One in five participants in the Estonian Biobank cohort with a biomarker summary score within the highest percentile died during the first year of follow-up, indicating prominent systemic reflections of frailty. The biomarker associations all replicated in the Finnish validation cohort. Including the four biomarkers in a risk prediction score improved risk assessment for 5-y mortality (increase in C-statistics 0.031, $p = 0.01$; continuous reclassification improvement 26.3%, $p = 0.001$).

***Conclusions:*** Biomarker associations with cardiovascular, nonvascular, and cancer mortality suggest novel systemic connectivities across seemingly disparate morbidities. The biomarker profiling improved prediction of the short-term risk of death from all causes above established risk factors. Further investigations are needed to clarify the biological mechanisms and the utility of these biomarkers for guiding screening and prevention.

*Please see later in the article for the Editors' Summary.*


Figure 3: Abstract from Fischer et al, 2014.

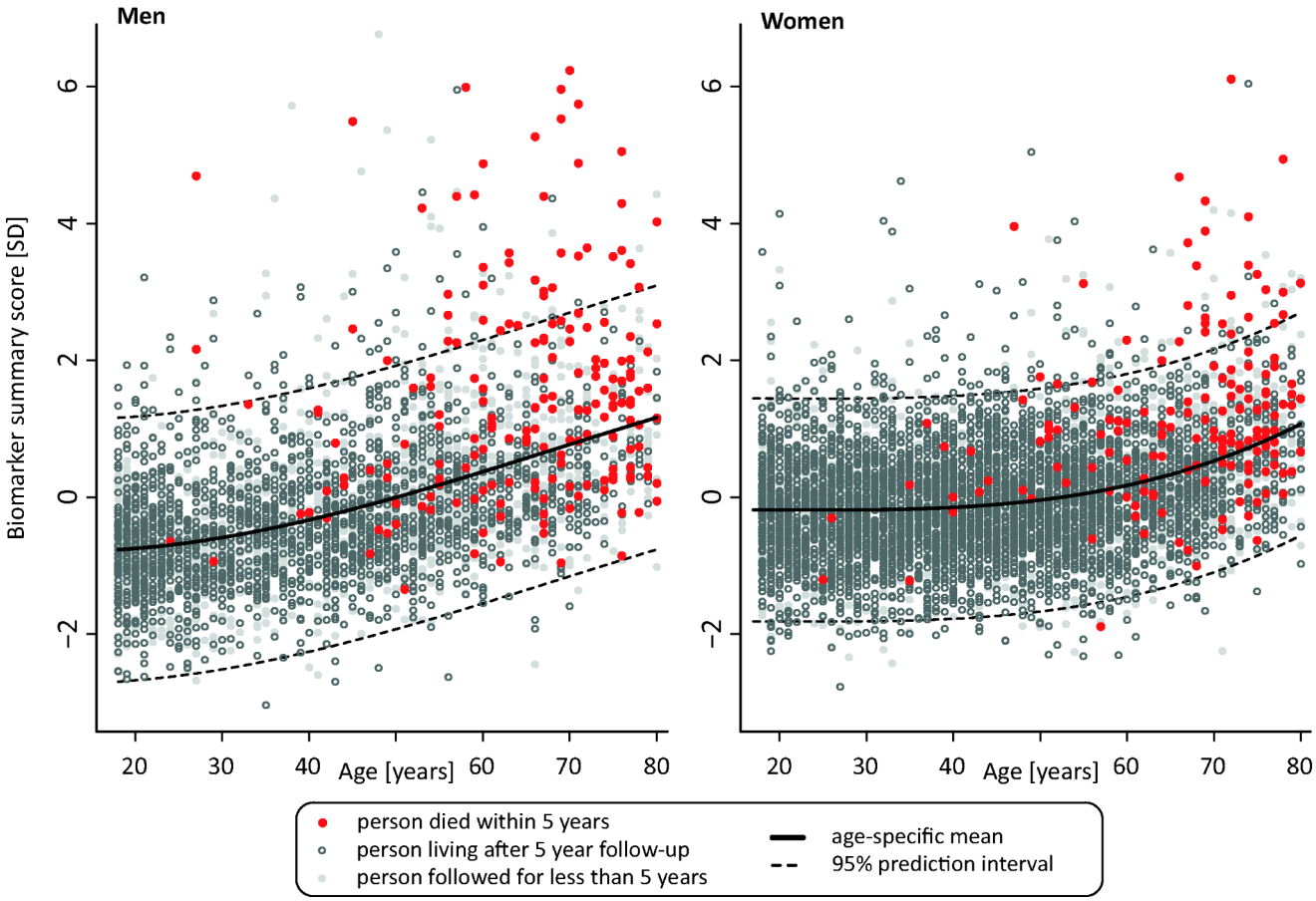


Figure 4: From Fischer et al, 2014. Scatter plot of age versus biomarker summary score for men and women from the Estonian Biobank cohort. The lines indicate a fit of age against the biomarker summary score, with dashed lines denoting 95% prediction intervals.