432 Class 11 Slides

github.com/THOMASELOVE/432-2018

2018-02-20

Setup

```
library(skimr)
library(pROC)
library(ROCR)
library(rms) # note: also loads Hmisc
library(simputation)
library(broom)
library(tidyverse)
```

Today's Materials

- Logistic Regression and the Framingham Study (part 2)
- Performing Linear Regression with `ols`
- Hormone Therapy and Baseline LDL in the `HERS` trial

The HERS trial is described in Vittinghoff et al., especially Chapter 4.

Logistic Regression and Framingham

Data Ingest, Cleanup (from Class 10)

```
fram <- read.csv("data/fram_new.csv") %>% tbl_df
set.seed(432001)
fram1 <- fram %>%
    impute_pmm(educ + cigs_day + heart_r ~ age + smoker) %>%
    impute_rlm(bmi + tot_chol ~ sex + age + sbp + heart_r) %>%
    impute_pmm(bp_meds ~ hx_htn + bmi + tot_chol) %>%
    impute_rlm(glucose ~ hx_dm + bmi + tot_chol + age) %>%
    mutate(ed_f = fct_recode(factor(educ),
                  "1_Some_HS" = "1", "2_HS_grad" = "2",
                  "3_Some_Col" = "3", "4_Col_grad" = "4"))
fram2 <- fram1 %>%
    select(subj, sex, age, smoker, cigs_day, bp_meds,
           hx_stroke, hx_htn, hx_dm, ed_f, tot_chol,
           sbp, dbp, bmi, heart_r, glucose, CHD_10)
```

The Models We've Fit (predicting CHD_10)

```
m_01 <- glm(CHD_10 ~ hx_htn, data = fram2,
            family = binomial)

d <- datadist(fram2)
options(datadist = "d")
m_01_lrm <- lrm(CHD_10 ~ hx_htn, data = fram2, x = T, y = T)

m_02 <- glm(CHD_10 ~ hx_htn + tot_chol,
            data = fram2, family = binomial)
m_02_lrm <- lrm(CHD_10 ~ hx_htn + tot_chol, data = fram2,
                x = TRUE, y = TRUE)
```

Assessing Predictive Quality: Discrimination

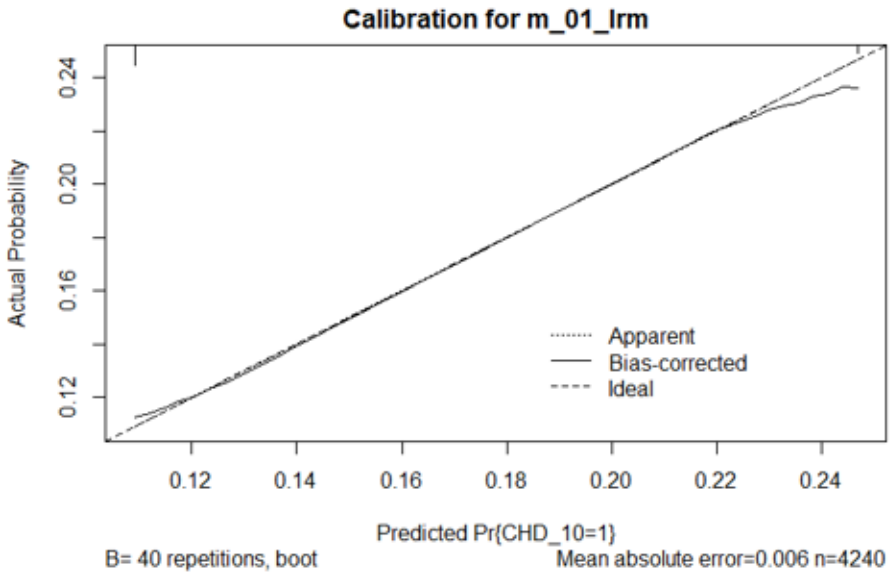
Key measures: C statistic, Nagelkerke R^2

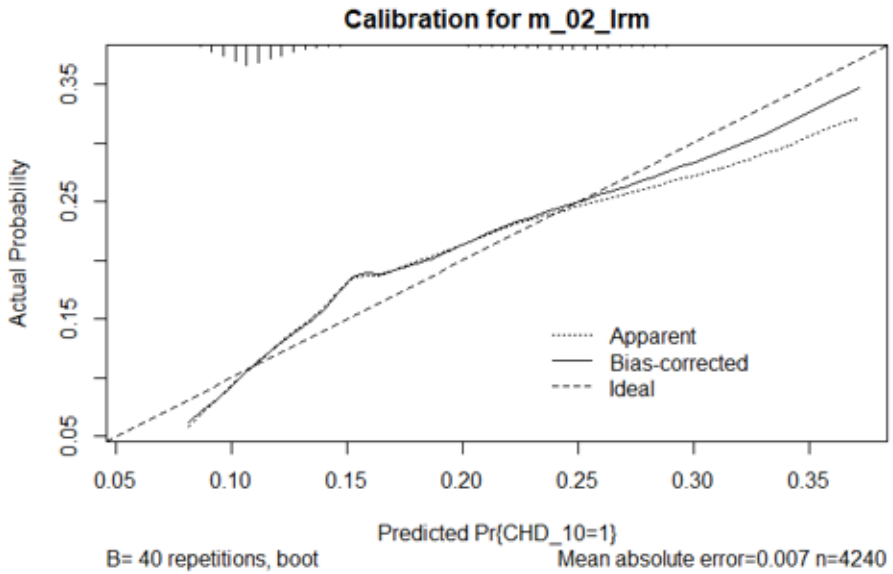
Model	C statistic	Nagelkerke R^2
`m_01_lrm`	0.614	0.051
`m_02_lrm`	0.640	0.055

and we could use `validate(model)` to address how well these results might hold up in new data.

Assessing Predictive Quality: Calibration Curves

```
plot(calibrate(m_01_lrm), main = "Calibration for m_01_lrm")
plot(calibrate(m_02_lrm), main = "Calibration for m_02_lrm")
```





Assessing Predictive Quality: Goodness of Fit Test

This uses the le Cessie-van Houwelingen-Copas-Hosmer unweighted sum of squares test statistic. to produce (using up just one degree of freedom) a global goodness of fit test. It's available through `residuals` applied to a `lrm` fit, with `type = "gof"`).

The essential components of a logistic regression fit are:

1. The logit transformation is the correct function linking the covariates with the conditional mean,
2. The linear predictor is correct (we don't need to include additional variables, transformations of predictors or interaction terms), and
3. The variance follows a Bernoulli distribution.

See Hosmer et al. 1997

The Omnibus Goodness of Fit Test

As in any omnibus test, a significant result here is difficult to interpret, but it means that something somewhere in the model is probably wrong.

- Harrell: I focus on directed tests such as allowing all continuous variables to have nonlinear effects or allowing selected interactions, and finding out how important the complex model terms are.

```
round(residuals(m_01_lrm, type = "gof"),3)
round(residuals(m_02_lrm, type = "gof"),3)
```

```
        Sum of squared   Expected
            errors       value|H0       SD           Z       P
Model 1    528.985        528.985    0.000  -2268.981   0.000
Model 2    527.948        527.291    0.331      1.986   0.047
```

Looking better in `m_02_lrm` but still some work to do.

Goal 3. Kitchen Sink Model for CHD_10

Focus on model with `lrm` first!

```
m_03 <- glm(CHD_10 ~ hx_htn + tot_chol + sex + age +
                     smoker + cigs_day + bp_meds +
                     hx_stroke + hx_dm + ed_f + sbp + dbp +
                     bmi + heart_r + glucose,
                   data = fram2, family = binomial)

d <- datadist(fram2)
options(datadist = "d")
m_03_lrm <- lrm(CHD_10 ~ hx_htn + tot_chol + sex + age +
                     smoker + cigs_day + bp_meds +
                     hx_stroke + hx_dm + ed_f + sbp + dbp +
                     bmi + heart_r + glucose,
                   data = fram2, x = TRUE, y = TRUE)
```

m_03_lrm (first section of output)

```
> m_03_lrm
Logistic Regression Model

 lrm(formula = CHD_10 ~ hx_htn + tot_chol + sex + age + smoker +
     cigs_day + bp_meds + hx_stroke + hx_dm + ed_f + sbp + dbp +
     bmi + heart_r + glucose, data = fram2, x = TRUE, y = TRUE)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs            4240    LR chi2     405.40    R2        0.159   C         0.733
  0             3596    d.f.            17    g         1.016   Dxy       0.466
  1              644    Pr(> chi2) <0.0001    gr        2.763   gamma     0.466
 max |deriv| 6e-10                           gp        0.120   tau-a     0.120
                                             Brier     0.115
```

m_03_lrm (second section of output)

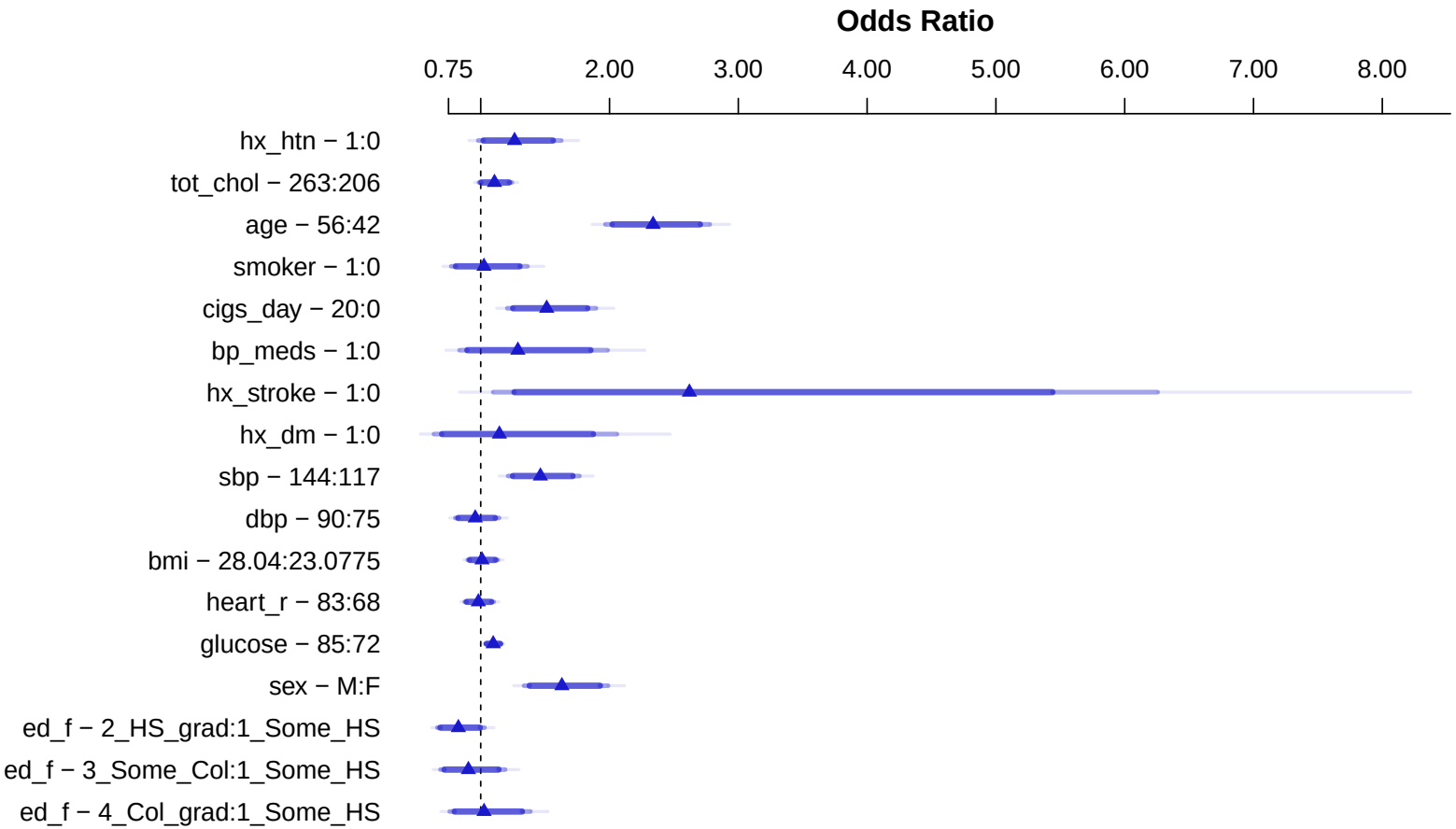
	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	-7.9981	0.6583	-12.15	<0.0001
hx_htn	0.2331	0.1287	1.81	0.0700
tot_chol	0.0018	0.0010	1.73	0.0842
sex=M	0.4886	0.1012	4.83	<0.0001
age	0.0607	0.0063	9.67	<0.0001
smoker	0.0248	0.1451	0.17	0.8642
cigs_day	0.0207	0.0057	3.60	0.0003
bp_meds	0.2534	0.2206	1.15	0.2506
hx_stroke	0.9633	0.4439	2.17	0.0300
hx_dm	0.1353	0.2989	0.45	0.6507
ed_f=2_HS_grad	-0.1906	0.1120	-1.70	0.0889
ed_f=3_Some_Col	-0.1005	0.1397	-0.72	0.4719
ed_f=4_Col_grad	0.0255	0.1533	0.17	0.8679
sbp	0.0141	0.0035	3.98	<0.0001
dbp	-0.0029	0.0060	-0.48	0.6294
bmi	0.0019	0.0118	0.16	0.8712
heart_r	-0.0012	0.0039	-0.32	0.7524
glucose	0.0071	0.0022	3.28	0.0010

Validating our Summary Statistics

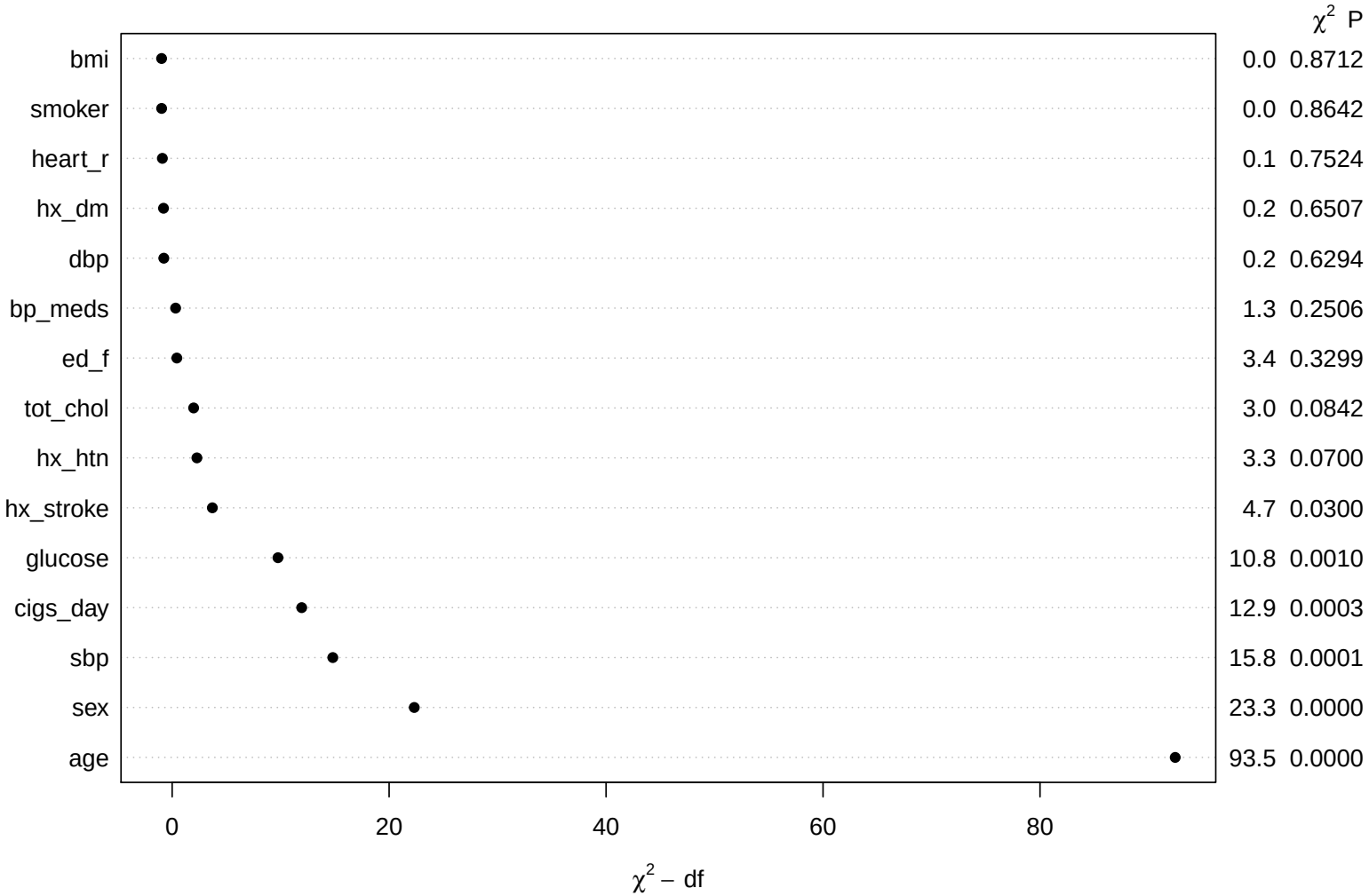
```
set.seed(432020) # probably better to set a seed
validate(m_03_lrm)[1:4,] # to fit things in the slide
```

```
          index.orig  training       test    optimism
Dxy        0.4658670 0.4690634  0.4575484 0.011515041
R2         0.1590194 0.1624443  0.1526612 0.009783165
Intercept  0.0000000 0.0000000 -0.0466690 0.046668999
Slope      1.0000000 1.0000000  0.9635322 0.036467806
          index.corrected  n
Dxy             0.4543520 40
R2              0.1492362 40
Intercept      -0.0466690 40
Slope           0.9635322 40
```

plot(summary(m_03_lrm))

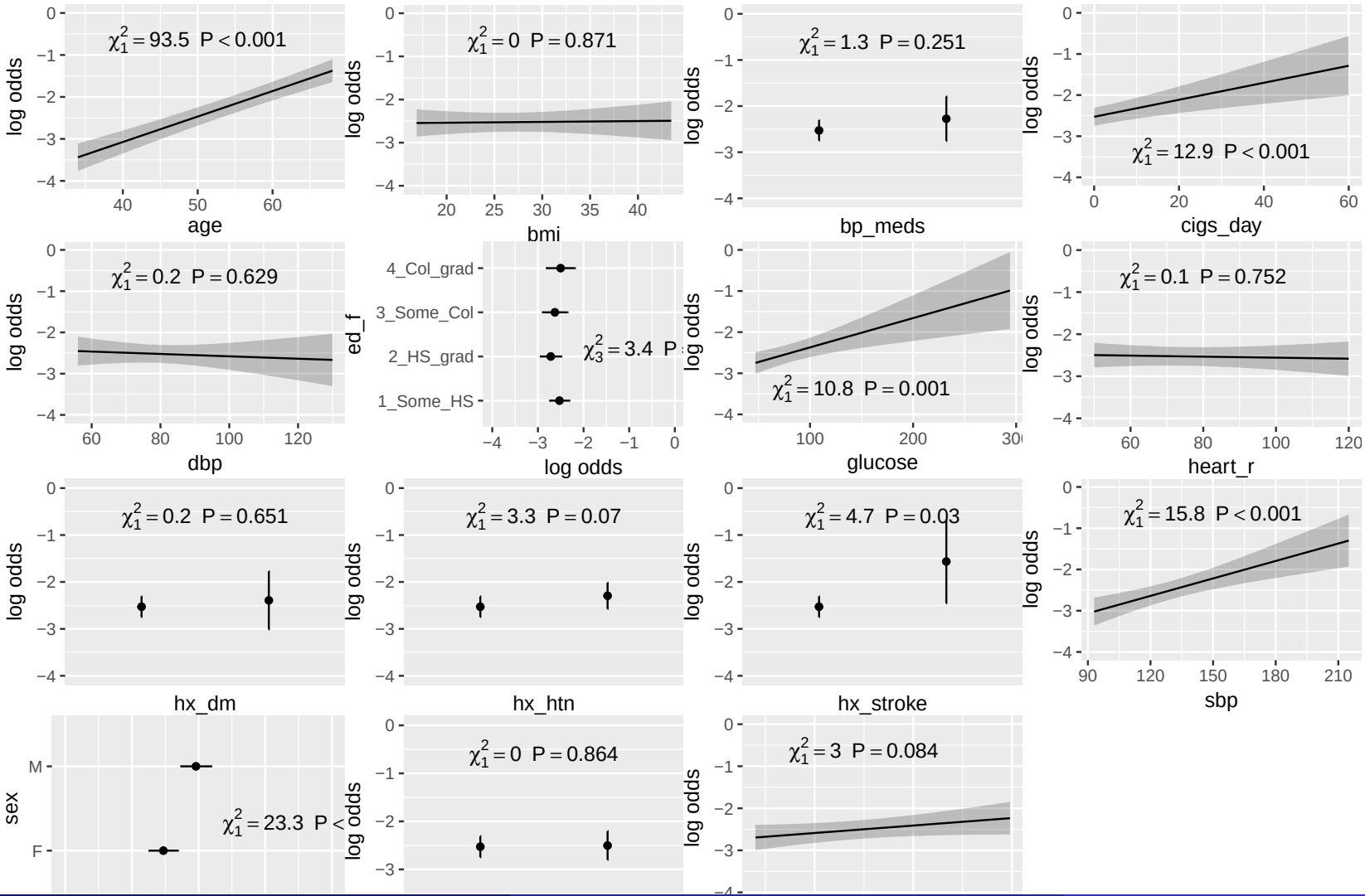


plot(anova(m_03_lrm))



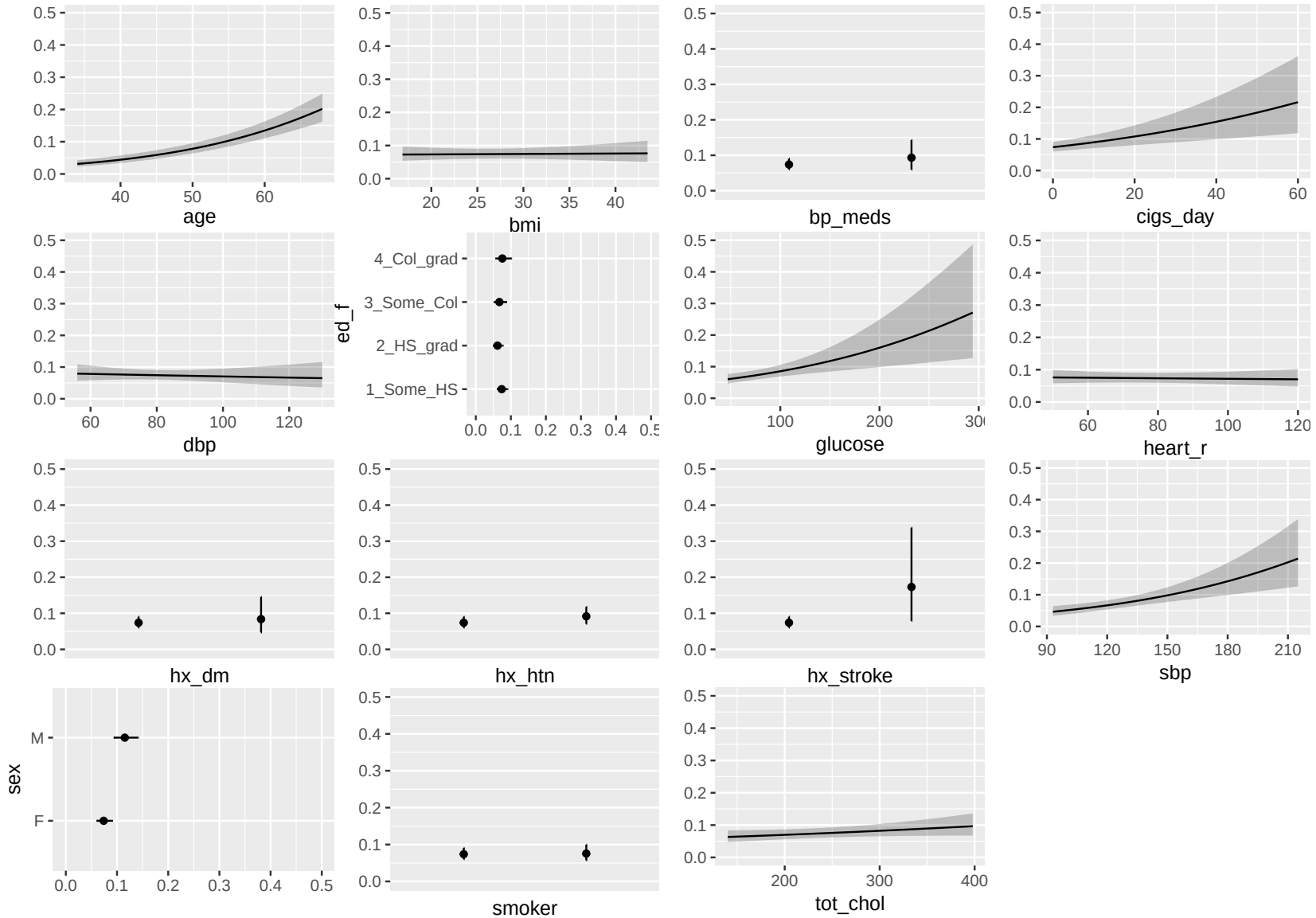
Can we see the prediction results?

```
ggplot(Predict(m_03_lrm),
       anova = anova(m_03_lrm), pval = TRUE)
```



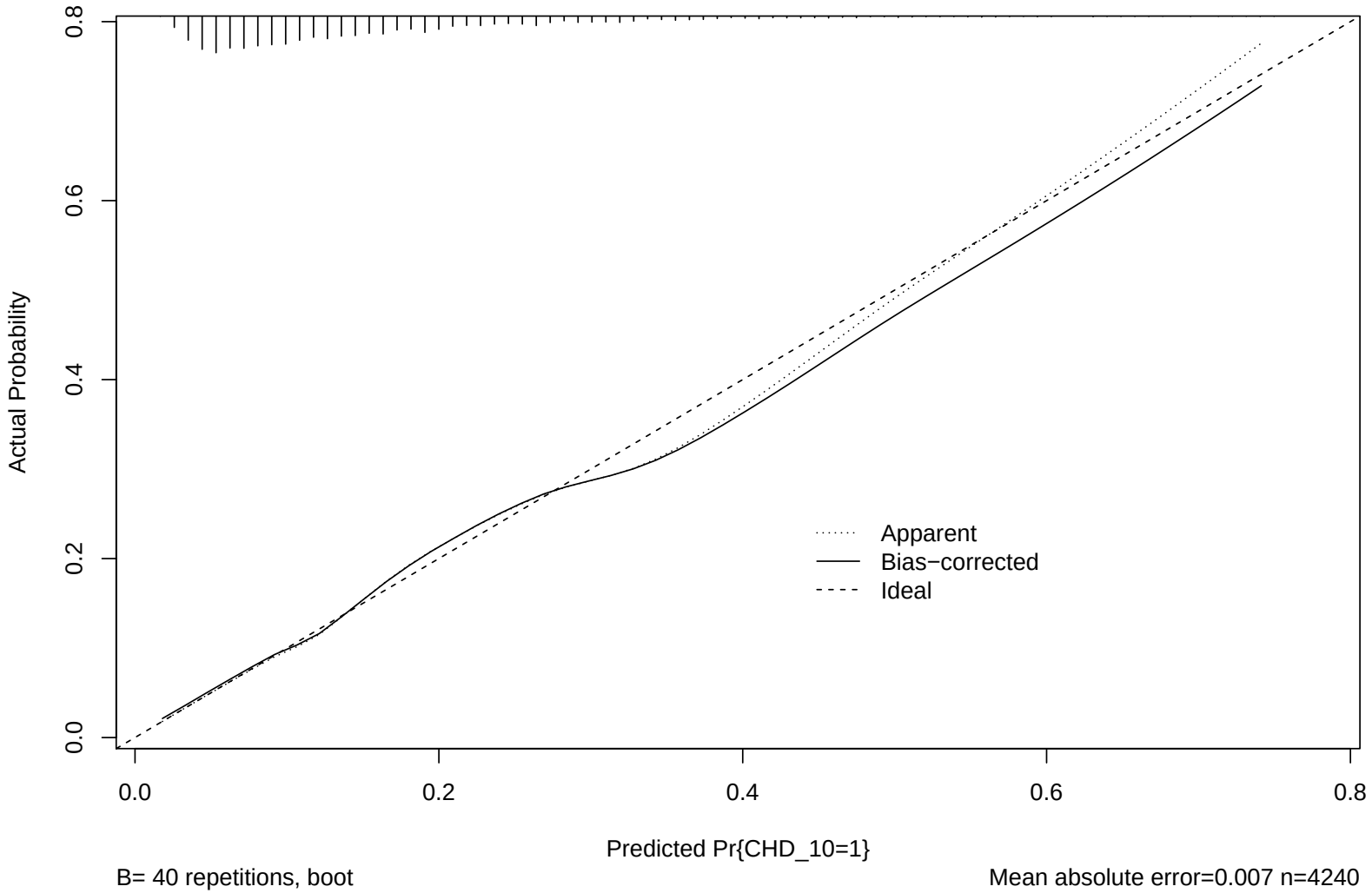
What about on a better scale?

```
ggplot(Predict(m_03_lrm, fun = plogis))
```



Calibration of mod_03_lrm

```
set.seed(432029); plot(calibrate(m_03_lrm))
```



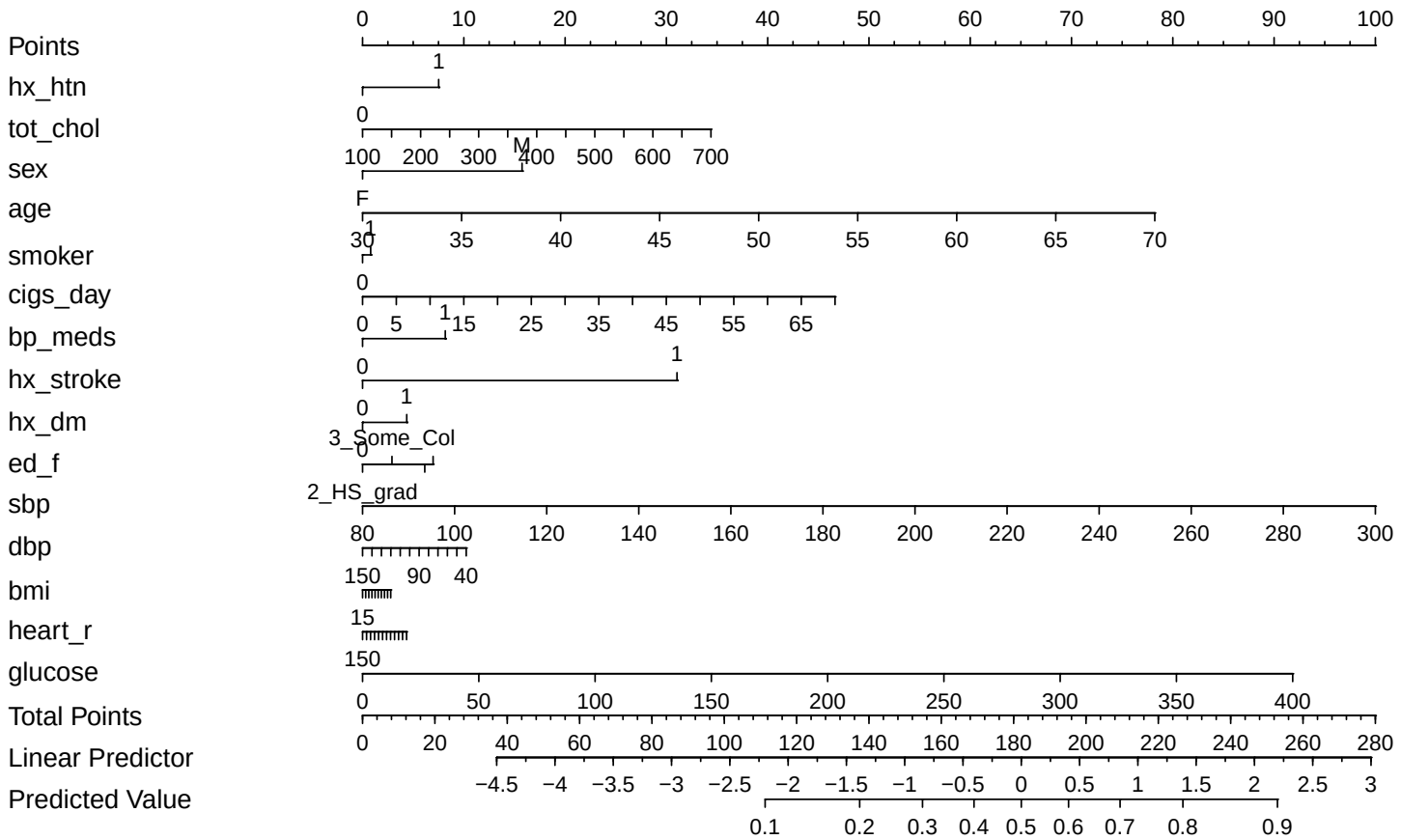
Goodness of fit test?

```
round(residuals(m_03_lrm, type = "gof"),3)
```

```
Sum of squared errors     Expected value|H0
              488.329                488.026
                   SD                      Z
                1.407                  0.215
                    P
                0.830
```

Nomogram of `mod_03_lrm`

```
plot(nomogram(m_03_lrm, fun = plogis))
```



Comparing our Three Nested Models

```
anova(m_01, m_02, m_03)
```

```
Analysis of Deviance Table

Model 1: CHD_10 ~ hx_htn
Model 2: CHD_10 ~ hx_htn + tot_chol
Model 3: CHD_10 ~ hx_htn + tot_chol + sex + age + smoker + cig
    bp_meds + hx_stroke + hx_dm + ed_f + sbp + dbp + bmi + hea
    glucose
  Resid. Df Resid. Dev Df Deviance
1      4238     3486.9
2      4237     3475.5  1   11.411
3      4222     3206.8 15  268.682
```

Model 2 vs. Model 3 at a glance

```
glance(m_02)
```

```
  null.deviance df.null    logLik      AIC      BIC
1      3612.209    4239 -1737.748 3481.495 3500.552
  deviance df.residual
1 3475.495        4237
```

```
glance(m_03)
```

```
  null.deviance df.null    logLik      AIC      BIC
1      3612.209    4239 -1603.407 3242.813 3357.155
  deviance df.residual
1 3206.813        4222
```