Logistic regression on the Framingham dataset

Estimating the 10-year risk of coronary heart disease

Leonardo Cerliani

The dataset

- 4238 Individuals, 15% at risk of CHD
- 43% males, 57 % females between 32 and 70 yrs (median age = 49)
- 15 recorded variables

  education, smoking habits, blood pressure measurements and anomalies, blood sample indices (e.g. glucose), drug assumption

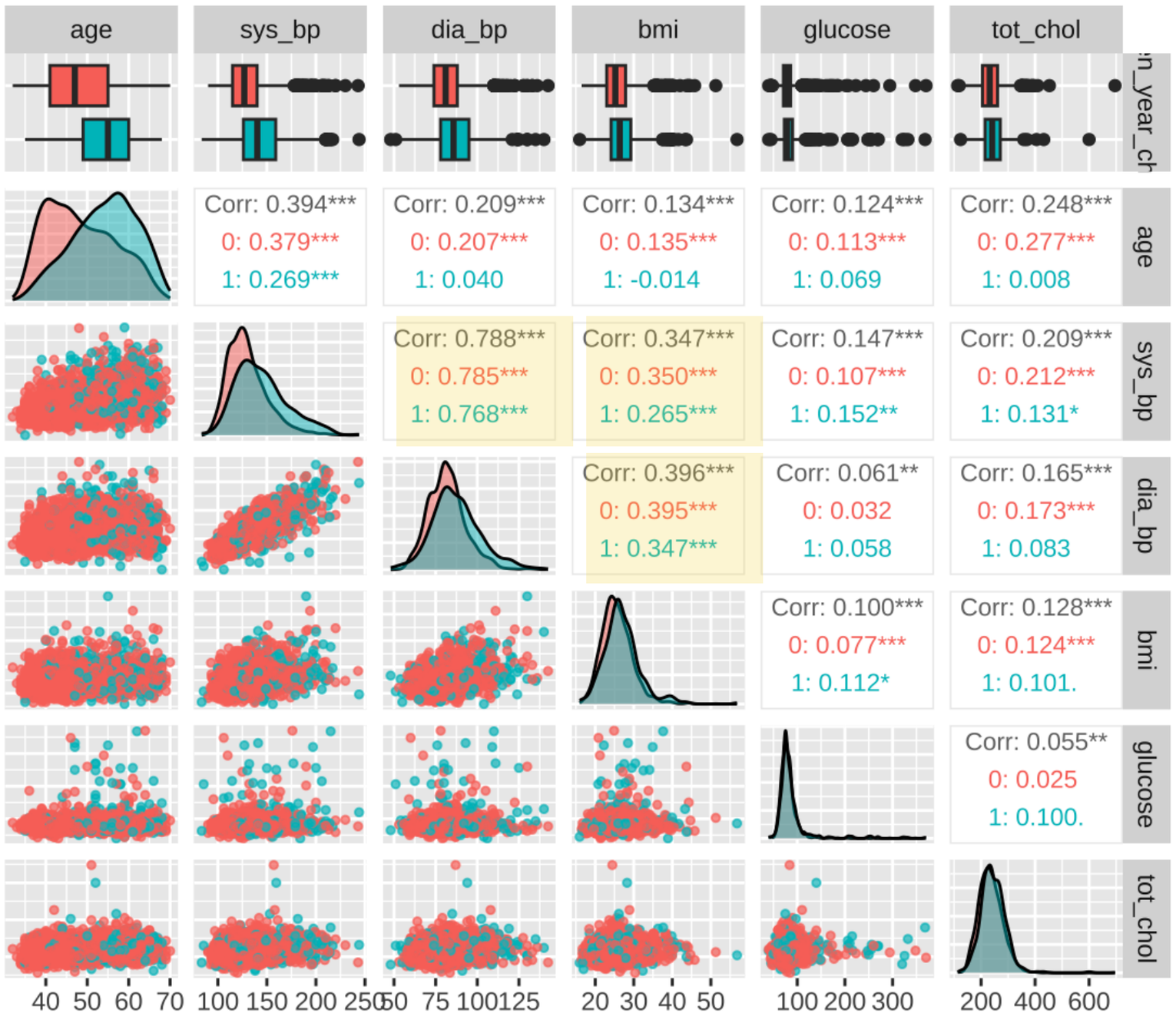
The dataset was split into a train (60%) and test (40%) set to assess the performance of the model on out-of-sample data

EDA numeric variables

Significant difference between people with and w/out risk of CHD

variable	t value	adj_pval
age	12.0815563	1.01e-31
sys_bp	11.9732840	1.73e-31
dia_bp	8.2215602	8.85e-16
glucose	6.5653995	1.28e-10
bmi	5.6308260	3.22e-08
tot_chol	4.0326391	7.59e-05
cigs_per_day	2.1141771	3.96e-02
heart_rate	0.8512159	3.95e-01

NB: p values corrected for multiple comparisons

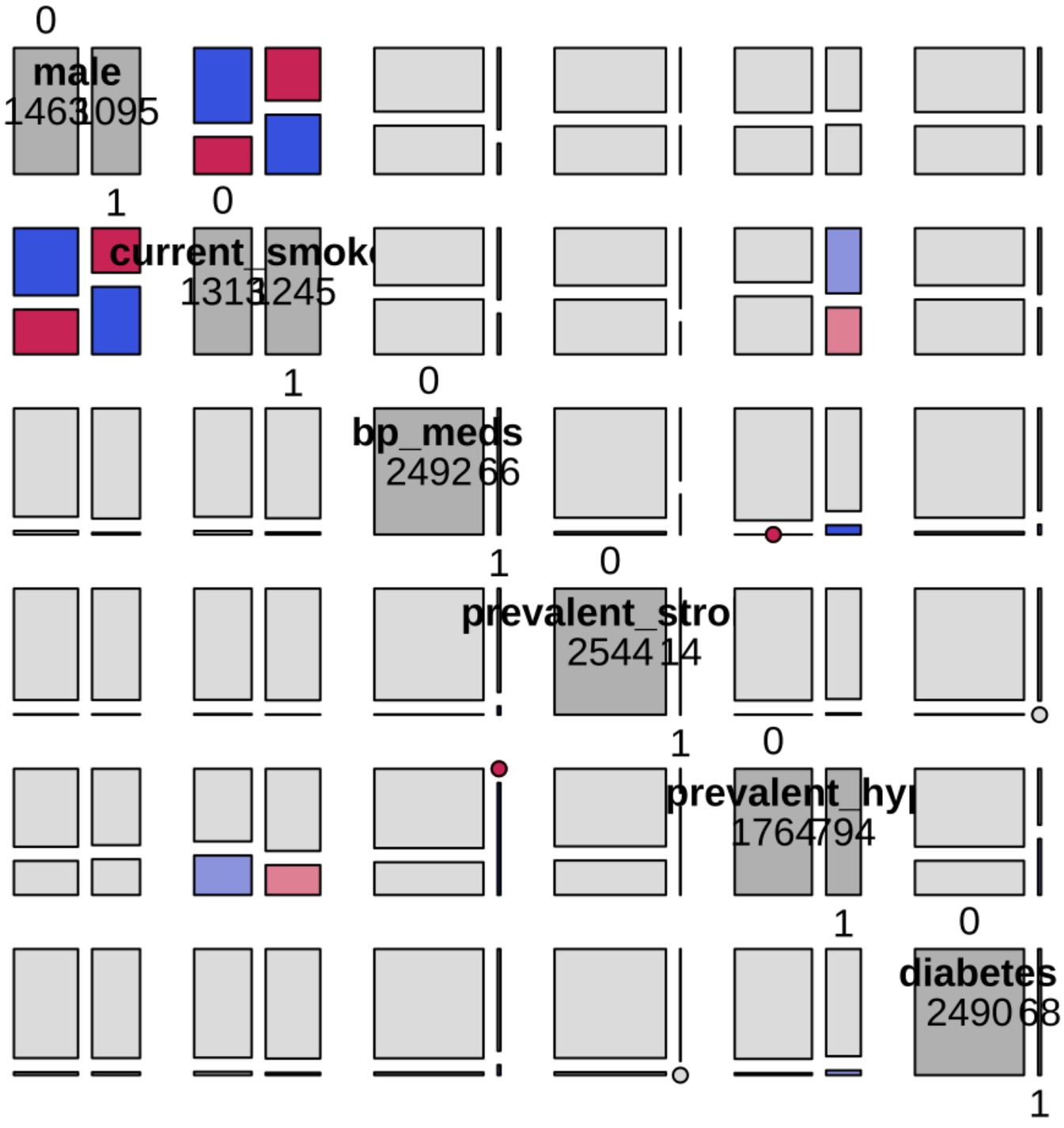


EDA categorical variables

Significant difference between people with and w/out risk of CHD

variable	chi square	pval
age	16.40	5.14e-05
hypertension	106.32	6.29e-25
10s cigarettes	14.22	6.6e-03
current smoker	1.28	0.26
bd medications	30.41	3.5e-08
education	31.41	6.97e-07

NB: p values corrected for multiple comparisons



Interaction of numeric and categorical variables

- There were **widespread interactions**
- Some are proxy - e.g hypertension and medication
- We decided to include them nevertheless and **evaluate variable deletion** based on coefficient estimates and **VIF**

Outliers and NA assessment

- Only **sys_bp** had extreme values - sys_bp > 200 were removed

- NAs in **glucose** and **cholesterol** replaced with the median of the hypertension group (highly correlated)

- NAs in **bp medications** were discarded given the sensitivity of this variable

Interesting questions

- How well is CHD risk predicted by **generic factors** alone - such as age and gender
- How much we can increase prediction using **more specialized measures** - such as hypertension and medication

Analytic strategy

- Backwards stepwise regression
  - **maximal model** : all 15 variables
  - **full model** : all selected variables
  - **final model** : after progressive variable removal

- At each step evaluate
  - AIC (goodness of fit)
  - VIF (collinearity)
  - significance of the variables
  - AUC (sensitivity / specificity)
  - pseudo R^2 (explained variance)

Full Model

AIC	max VIF	AUC	R^2
1885	2.06	0.705	0.17

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.8924235  0.7224553 -10.924  < 2e-16 ***
## male1            0.3996399  0.1306947   3.058 0.002230 **
## prevalent_hyp1   0.3564960  0.1628251   2.189 0.028565 *
## bp_meds1         0.4381388  0.2883448   1.519 0.128638
## education2      -0.3319065  0.1512202  -2.195 0.028174 *
## education3      -0.3409724  0.1858949  -1.834 0.066621 .
## education4      -0.0373354  0.1959238  -0.191 0.848870
## age              0.0613141  0.0080512   7.616 2.63e-14 ***
## sys_bp           0.0118211  0.0034909   3.386 0.000709 ***
## bmi              0.0137041  0.0146029   0.938 0.348011
## glucose          0.0075754  0.0021169   3.579 0.000345 ***
## tot_chol         0.0002287  0.0014069   0.163 0.870891
## cigs_per_day     0.0191629  0.0052321   3.663 0.000250 ***
```

Remove EVs with no significant effect

AIC	max VIF	AUC	R^2
1884	1.83	0.711	0.16

```
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.027822   0.562910 -14.261  < 2e-16 ***
## male1           0.434934   0.127776   3.404 0.000664 ***
## prevalent_hyp1  0.387263   0.160552   2.412 0.015862 *
## age             0.064891   0.007775   8.346  < 2e-16 ***
## sys_bp          0.013408   0.003393   3.952 7.76e-05 ***
## glucose         0.007572   0.002106   3.595 0.000324 ***
## cigs_per_day    0.018188   0.005197   3.500 0.000465 ***
```

- AUC increases and VIF decreases
- Significance of the main predictors increases

Remove prevalent hypertension

AIC	max VIF	AUC	R^2
1887	1.20	0.713	0.16

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.643579   0.504565 -17.131  < 2e-16 ***
## male1         0.449590   0.127385   3.529 0.000417 ***
## age           0.065845   0.007745   8.502  < 2e-16 ***
## sys_bp        0.018684   0.002597   7.194 6.29e-13 ***
## glucose       0.007547   0.002112   3.573 0.000353 ***
## cigs_per_day  0.017858   0.005184   3.445 0.000571 ***
```

- AUC increases and VIF decreases
- Significance of the main predictors increases

Final model

AIC	max VIF	AUC	R^2
1887	1.20	0.713	0.16

```
##                    exp(Est.)   2.5%   97.5%   z val.       p    VIF
## ---------------- ----------- ------ ------- -------- ------ ------
## (Intercept)             0.00   0.00    0.00   -17.14   0.00
## male1                   1.57   1.22    2.01     3.52   0.00   1.16
## age                     1.07   1.05    1.08     8.50   0.00   1.17
## sys_bp                  1.02   1.01    1.02     7.19   0.00   1.14
## glucose                 1.01   1.00    1.01     3.56   0.00   1.01
## tens_cigs               1.19   1.08    1.32     3.44   0.00   1.20
## ------------------------------------------------------------------
```

- Simple: only 5 easily retrievable measures
- Best goodness of fit and predictability out of all models
- Accuracy = 85.6%, however Sensitivity = 0.7 (too many false negative) with a threshold for binary prediction = 0.5

Final vs Automatic Stepwise (AIC)

Final model

AIC	max VIF	AUC	R^2
1887	1.20	0.713	0.16

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.643579   0.504565 -17.131  < 2e-16 ***
## male1         0.449590   0.127385   3.529 0.000417 ***
## age           0.065845   0.007745   8.502  < 2e-16 ***
## sys_bp        0.018684   0.002597   7.194 6.29e-13 ***
## glucose       0.007547   0.002112   3.573 0.000353 ***
## cigs_per_day  0.017858   0.005184   3.445 0.000571 ***
```

Automatic Stepwise

AIC	max VIF	AUC	R^2
1888	1.85	0.706	0.17

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.655170   0.587371 -13.033  < 2e-16 ***
## male1              0.387895   0.129655   2.992 0.002774 **
## age                0.060750   0.007983   7.610 2.74e-14 ***
## education2        -0.335970   0.150240  -2.236 0.025337 *
## education3        -0.352595   0.184679  -1.909 0.056232 .
## education4        -0.038528   0.195294  -0.197 0.843606
## cigs_per_day       0.019113   0.005227   3.657 0.000256 ***
## prevalent_stroke1  0.969730   0.601408   1.612 0.106868
## prevalent_hyp1     0.372800   0.161896   2.303 0.021295 *
## sys_bp             0.013282   0.003412   3.893 9.92e-05 ***
## glucose            0.007787   0.002106   3.698 0.000217 ***
```

- Same main predictors as our full model
- Additional Education and Hypertension do NOT survive correction for multiple comparison
- More complex, less contribution of the main predictors

Final model + categorical interactions

Final model

AIC	max VIF	AUC	R^2
1887	1.20	0.713	0.16

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.643579   0.504565 -17.131  < 2e-16 ***
## male1         0.449590   0.127385   3.529 0.000417 ***
## age           0.065845   0.007745   8.502  < 2e-16 ***
## sys_bp        0.018684   0.002597   7.194 6.29e-13 ***
## glucose       0.007547   0.002112   3.573 0.000353 ***
## cigs_per_day  0.017858   0.005184   3.445 0.000571 ***
```

Categorical Interactions

AIC	max VIF	AUC	R^2
1889	4.18	0.707	0.16

```
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -8.071834   0.575043 -14.037  < 2e-16 ***
## male1                    0.411137   0.177391   2.318 0.020466 *
## age                      0.064531   0.007790   8.284  < 2e-16 ***
## sys_bp                   0.013450   0.003398   3.959 7.54e-05 ***
## glucose                  0.007486   0.002101   3.563 0.000367 ***
## male0:smoker1            0.386439   0.213094   1.813 0.069760 .
## male1:smoker1            0.560270   0.206397   2.715 0.006637 **
## smoker0:hypertension     0.472103   0.200758   2.352 0.018693 *
## smoker1:hypertension     0.287916   0.201998   1.425 0.154058
```

- High collinearity between sex, current smoker and hypertension
- Most of the interactions are not significant
- More complex, less contribution of the main predictors

Model age with a sigmoid

Final model

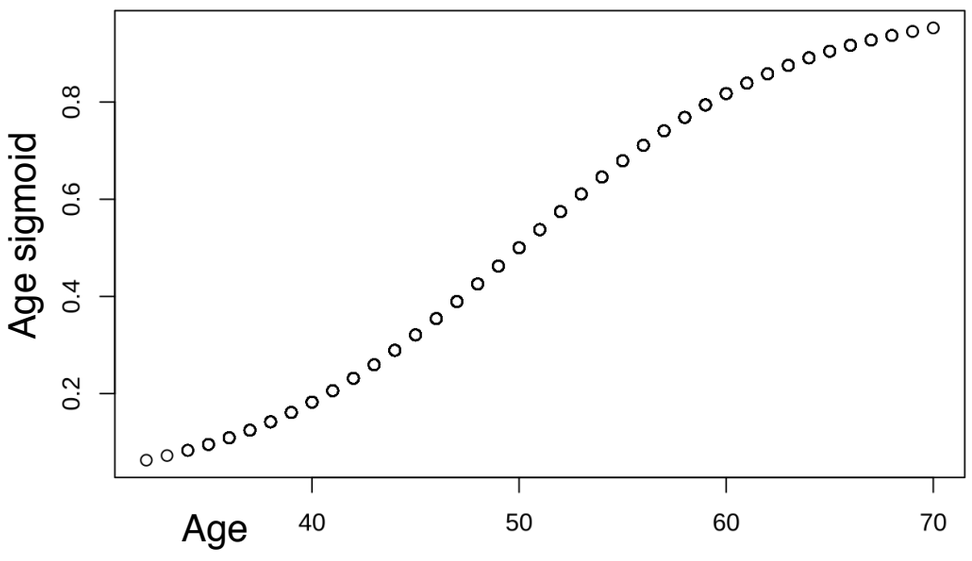
AIC	max VIF	AUC	R^2
1887	1.20	0.713	0.16

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.643579   0.504565 -17.131  < 2e-16 ***
## male1         0.449590   0.127385   3.529 0.000417 ***
## age           0.065845   0.007745   8.502  < 2e-16 ***
## sys_bp        0.018684   0.002597   7.194 6.29e-13 ***
## glucose       0.007547   0.002112   3.573 0.000353 ***
## cigs_per_day  0.017858   0.005184   3.445 0.000571 ***
```

Age sigmoid

AIC	max VIF	AUC	R^2
1886	1.20	0.712	0.16

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.433353   0.401903 -16.007  < 2e-16 ***
## male1         0.449367   0.127351   3.529 0.000418 ***
## age_sigmoid   2.197337   0.258184   8.511  < 2e-16 ***
## sys_bp        0.018622   0.002598   7.167 7.64e-13 ***
## glucose       0.007502   0.002109   3.556 0.000376 ***
## cigs_per_day  0.017897   0.005184   3.452 0.000556 ***
```



- Hypothesis: risk of CHD increases faster after 40 yrs old
- Coefficients and significance are virtually identical, no benefit

Final vs Maximal Model

Final model

AIC	max VIF	AUC	R^2
1887	1.20	0.713	0.16

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.643579   0.504565 -17.131  < 2e-16 ***
## male1         0.449590   0.127385   3.529 0.000417 ***
## age           0.065845   0.007745   8.502  < 2e-16 ***
## sys_bp        0.018684   0.002597   7.194 6.29e-13 ***
## glucose       0.007547   0.002112   3.573 0.000353 ***
## cigs_per_day  0.017858   0.005184   3.445 0.000571 ***
```

Maximal model

AIC	max VIF	AUC	R^2
1890	3.64	0.703	0.17

```
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.4497344  0.8561698  -8.701  < 2e-16 ***
## male1             0.3800302  0.1329243   2.859  0.00425 **
## age               0.0608765  0.0083022   7.333 2.26e-13 ***
## education2       -0.3243521  0.1517706  -2.137  0.03259 *
## education3       -0.3327557  0.1860997  -1.788  0.07377 .
## education4       -0.0502180  0.1969542  -0.255  0.79874
## current_smoker1   0.1717930  0.1922979   0.893  0.37166
## cigs_per_day      0.0149836  0.0078998   1.897  0.05787 .
## bp_meds1          0.3755894  0.2932894   1.281  0.20033
## prevalent_stroke1 0.8294602  0.6158164   1.347  0.17800
## prevalent_hyp1    0.3669952  0.1660561   2.210  0.02710 *
## diabetes1         0.3126526  0.3897172   0.802  0.42241
## tot_chol          0.0004310  0.0014095   0.306  0.75980
## sys_bp            0.0121563  0.0046474   2.616  0.00890 **
## dia_bp            0.0004736  0.0078561   0.060  0.95193
## bmi               0.0141265  0.0151335   0.933  0.35058
## heart_rate       -0.0068913  0.0050928  -1.353  0.17601
## glucose           0.0063839  0.0028683   2.226  0.02604 *
```

- Misses cigarettes per day and barely captures glucose
- Significance of sex is highly decreased
- High collinearity (as expected) especially between sys_bp and dia_bp

Final model performance

- **male** : males have **57%** greater odds of CHD risk than females
- **cigarettes** : **19%** greater odds for any additional 10 per day
- **age** : the odds of CHD increase **6.8%** every year
- **systolic blood pressure** : **1.8%** greater odds for each mmHg

Lowering the threshold for binary prediction to **0.1** yields:

Sensitivity 83%	TP = 197	FP = 757	Specificity 43%
	FN = 38	TN = 588	

Samples balanced for risk

AIC	max VIF	AUC	R^2
922	1.20	0.738	0.22

```
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.238967   0.720868 -10.042  < 2e-16 ***
## male1         0.319629   0.169510   1.886  0.05935 .
## age           0.074519   0.010682   6.976 3.04e-12 ***
## sys_bp        0.015708   0.003688   4.259 2.06e-05 ***
## glucose       0.010113   0.003923   2.578  0.00994 **
## cigs_per_day  0.030757   0.007706   3.991 6.57e-05 ***
```

Sex is NOT anymore significant

Lowering the threshold for binary prediction to **0.4** yields:

Sensitivity 81%

TP = 191

FN = 44

FP = 120

TN = 115

Specificity 48%

Conclusions

- In this situation, **the aim is to maximise sensitivity** - i.e. **minimizing false negatives**

- **The final model correctly predicts risk of CHD in 10 years in 83% of the people who are actually at risk** (i.e. sensitivity) when the threshold for binary classification is set to 0.1

- **This model is simple**, containing only easily retrievable variables: **sex, age, systolic blood pressure, glucose and # cigarettes per day**

- **Assumption of blood pressure medicaments *does not* appear to decrease the odds of CHD**

- The explained variance is low (17% max). **Other unexplored variables** such as alcohol consumption, stress, wealth, **might improve fit and performance**

- **Sex** has probably a relatively low impact, as revealed by a balanced sample