Generalized Linear Models

Applications

Generalized linear models (GLMs)

Generalized Linear Models provide a unified framework for modeling outcomes from the various family of distributions, including:

- Continuous outcomes (Gaussian $\rightarrow$ linear regression)
- Binary outcomes (Binomial $\rightarrow$ logistic regression)
- Count outcomes (Poisson $\rightarrow$ rate models)

Core idea

All GLMs share the same structure:

$$g(\mu_i) = X_i^\top \beta$$

Where:

- Y_i follows a distribution from the **exponential family**
- $\mu_i = E(Y_i \mid X_i)$
- $g(\cdot)$ is a **link function**
- $X_i^\top \beta$ is the linear predictor

Three components of a GLM

1. **Random component**

Distribution of Y (Normal, Binomial, Poisson)

2. **Systematic component**

Linear predictor $X^\top \beta$

3. **Link function**

Connects mean outcome to predictors

- Identity $\rightarrow$ Linear regression
- Logit $\rightarrow$ Logistic regression
- Log $\rightarrow$ Poisson regression

Why GLMs matter in health research

They allow us to model:

- Mean differences (QALY)
- Odds ratios (disease risk)
- Incidence rate ratios (person-time data)

Within one unified framework.

Outcome Type	Distribution	Link	Interpretation
Continuous	Gaussian	Identity	Mean difference
Binary	Binomial	Logit	Odds ratio
Count	Poisson	Log	Rate ratio

Linear regression: Ketamine for Chronic Pain (QALY)

Clinical Question

Does ketamine dosage and patient characteristics influence **Quality-Adjusted Life Years (QALY)**?

We model the expected outcome:

$$E(Y \mid X)$$

Where:

- $(Y = QALY)$
- $(X =)$ treatment characteristics + patient covariates

Model specification

Linear regression assumes:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

with:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Thus:

$$E(Y_i \mid X_i) = X_i^\top \beta$$

$$Var(Y_i \mid X_i) = \sigma^2$$

Interpretation of coefficients

For a continuous predictor (X_j):

$$\beta_j = \frac{\partial E(Y)}{\partial X_j}$$

- Expected change in QALY for a 1-unit increase in X_j, holding other variables constant.

For categorical predictors:

$$\beta_j = \text{Difference in mean QALY vs reference group}$$

Estimation principle

Parameters are estimated via **Ordinary Least Squares (OLS)**:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

OLS minimizes:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Assumptions

- Linearity
- Independence
- Homoscedasticity
- Normal residuals
- No strong multicollinearity

Questions

1. Load and inspect the `ketapain.csv` Dataset.

```r
keta <- read.csv("datasets/ketapain.csv")

head(keta)
```

```
  patientID   sexe age av_dose level_dose cum_dose cum_days perfusion     cost
1      6317 female  35   0.700   low dose    269.5        5         24  4689.95
2      2517 female  76   0.700   low dose    269.5        5         24  4689.95
3      1023 female  18   0.075   low dose     63.0       12          4 11165.30
4      3002 female  50  74.000   low dose    222.0        3         24  2829.20
5      7702   male  46  74.000   low dose    222.0        3         24  2829.20
6       439 female  54   0.500   low dose    105.0        3          6  2817.50
        qaly       mode
1 0.32938894    continu
2 0.69094614    continu
3 0.31734359 discontinu
4 0.04861697    continu
5 0.63407075    continu
6 0.43630836 discontinu
```

```r
str(keta)
```

```
'data.frame':   184 obs. of  11 variables:
 $ patientID : int  6317 2517 1023 3002 7702 439 1224 1616 2924 1317 ...
 $ sexe      : chr  "female" "female" "female" "female" ...
 $ age       : int  35 76 18 50 46 54 51 50 43 30 ...
 $ av_dose   : num  0.7 0.7 0.075 74 74 0.5 100 1.6 100 0.7 ...
 $ level_dose: chr  "low dose" "low dose" "low dose" "low dose" ...
 $ cum_dose  : num  270 270 63 222 222 ...
 $ cum_days  : int  5 5 12 3 3 3 3 5 3 5 ...
 $ perfusion : num  24 24 4 24 24 6 3 24 3 24 ...
 $ cost      : num  4690 4690 11165 2829 2829 ...
 $ qaly      : num  0.3294 0.6909 0.3173 0.0486 0.6341 ...
 $ mode      : chr  "continu" "continu" "discontinu" "continu" ...
```

Before modeling, check:

- Is `qaly` continuous?
- Any missing values?
- Correct variable types?

2. Perform some basic cleaning.

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

- Why convert to factors?

Because categorical variables must be treated as group comparisons, not numeric scales.

3. Fit the linear regression.

We estimate:

$$E(\text{QALY} \mid X) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \beta_3 \text{Average Dose} + \cdots$$

```
Call:
lm(formula = qaly ~ age + sexe + level_dose + cum_dose + cum_days +
    perfusion + mode, data = keta)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4744 -0.1313  0.0092  0.1148  0.4621

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.3580489  0.1467429   2.440   0.0157 *
age                     0.0013680  0.0012007   1.139   0.2561
sexemale                0.0306974  0.0350236   0.876   0.3820
level_doselow dose     -0.0529579  0.1371372  -0.386   0.6998
level_dosemidium level  0.0373005  0.1404256   0.266   0.7908
cum_dose               -0.0003105  0.0001972  -1.575   0.1171
cum_days               -0.0030772  0.0074231  -0.415   0.6790
perfusion               0.0045027  0.0027079   1.663   0.0981 .
```

```
modediscontinu          0.0627110  0.0582525   1.077   0.2832
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1922 on 175 degrees of freedom
Multiple R-squared:  0.04889,   Adjusted R-squared:  0.005413
F-statistic: 1.124 on 8 and 175 DF,  p-value: 0.349
```

4. Interpret key coefficients

```
        (Intercept)                 age             sexemale
       0.3580489014        0.0013679681        0.0306973972
level_doselow dose level_dosemidium level            cum_dose
      -0.0529579342        0.0373004898       -0.0003105442
           cum_days           perfusion       modediscontinu
      -0.0030772233        0.0045027387        0.0627109957
```

5. Check model assumptions

Residuals:

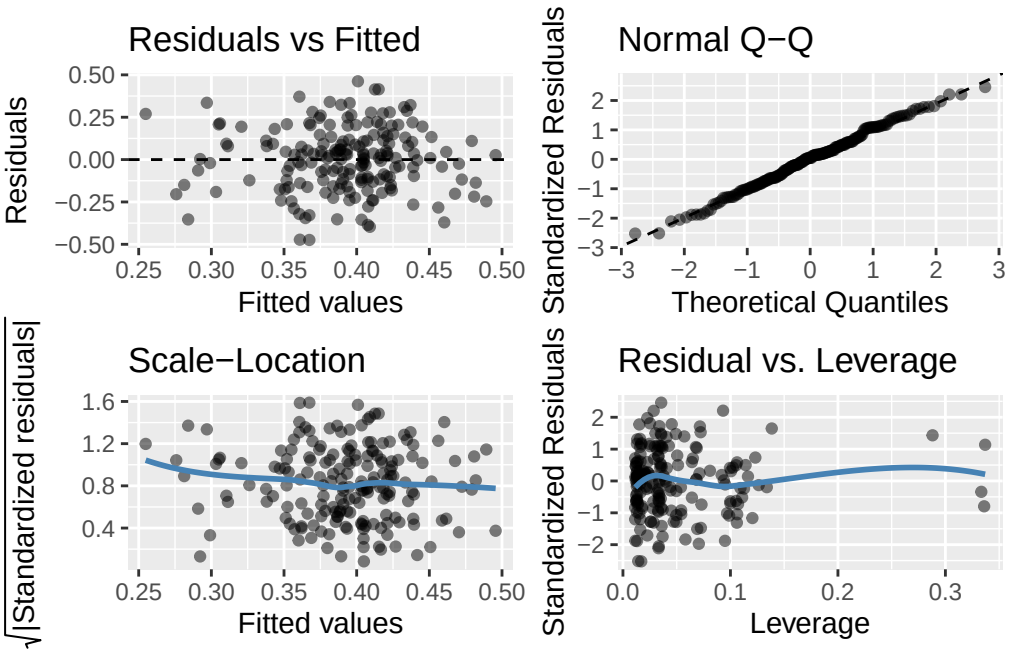
$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

We check:

- Residual vs fitted → Linearity and homoscedasticity
- QQ-plot → Normality
- Influence diagnostics

```
Loading required package: ggplot2
```

```
Warning: `fortify(<lm>)` was deprecated in ggplot2 4.0.0.
i Please use `broom::augment(<lm>)` instead.
i The deprecated feature was likely used in the ggplot2 package.
  Please report the issue at <https://github.com/tidyverse/ggplot2/issues>.
```



Interpretation of diagnostics

- Funnel shape → heteroscedasticity
- Systematic curve → non-linearity
- Heavy tails in QQ → non-normal residuals
- Extreme leverage → influential observations

6. Predict QALY for a new patient

Theoretical prediction:

$$\hat{Y}_{new} = x_{new}^{\top}\hat{\beta}$$

Prediction interval:

$$\hat{Y} * new \pm t * n - p\sqrt{\hat{\sigma}^2(1 + x_{new}^{\top}(X^{\top}X)^{-1}x_{new})}$$

Difference:

- Confidence interval → mean response
- Prediction interval → individual patient
- Helper functionfor predicting qaly:

```r
predict_qaly <- function(age, sexe, av_dose, level_dose,
                         cum_dose, cum_days, perfusion, mode) {

  newdata <- data.frame(
    age = age,
    sexe = factor(sexe, levels = levels(keta$sexe)),
    av_dose = av_dose,
    level_dose = factor(level_dose, levels = levels(keta$level_dose)),
    cum_dose = cum_dose,
    cum_days = cum_days,
    perfusion = perfusion,
    mode = factor(mode, levels = levels(keta$mode))
  )

  predict(fit_lm, newdata = newdata, interval = "prediction")
}
```

{r} #| echo: false predict_qaly(age = 50, sexe = “female”, av_dose = 0.7, level_dose = “low dose”, cum_dose = 269.5, cum_days = 5, perfusion = 24, mode = “continu”)

Logistic regression: Arthritis risk factors (`arthritis`)

Clinical Question

Which patient factors are associated with **arthritis** in a cohort dataset?

Outcome (binary):

$$Y = \begin{cases} 1 & \text{arthritis present} \\ 0 & \text{arthritis absent} \end{cases}$$

We want to model the **risk**:

$$P(Y = 1 \mid X)$$

with covariates age, gender, BMI, diabetes, smoking.

Why Logistic Regression?

A linear model can predict values outside ([0,1]). Logistic regression ensures predicted risks are valid probabilities by using a **link function**:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

with:

$$p_i = P(Y_i = 1 \mid X_i)$$

2Model Specification

$$\log\left(\frac{P(Y_i = 1 \mid X_i)}{1 - P(Y_i = 1 \mid X_i)}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

Equivalently, the predicted probability is:

$$p_i = \frac{1}{1 + \exp(-X_i^\top \beta)}$$

Interpretation of coefficients

Coefficients are in **log-odds**:

$$\beta_j = \text{change in log-odds per 1-unit increase in } X_j$$

Exponentiating gives the **Odds Ratio (OR)**:

$$OR_j = e^{\beta_j}$$

Interpretation:

- (OR>1): increases odds (higher risk)
- (OR<1): decreases odds (protective association)
- (OR=1): no association

For categorical predictors, ORs are relative to the **reference category**.

Estimation principle

Parameters are estimated by **Maximum Likelihood Estimation (MLE)**.

Likelihood:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

We typically use the log-likelihood and optimize numerically.

Inference

We test:

$$H_0 : \beta_j = 0$$

using Wald tests (z-statistics) or likelihood-based tests.

Confidence intervals for ORs are:

$$\exp(\hat{\beta}_j \pm 1.96 \cdot SE(\hat{\beta}_j))$$

Marginal Effects (Clinical interpretation)

Because ORs can be hard to interpret clinically, marginal effects translate to **probability changes**:

$$\frac{\partial p}{\partial x_j} = \beta_j, p(1-p)$$

So the probability impact depends on baseline risk p. That's why we often compute **average marginal effects**.

1. Load and inspect the dataset.

```
arthritis <- read.csv("datasets/arthritis.csv")

head(arthritis)
```

```
     id status heart.attack.relative gender age   bmi diabetes alcohol smoke
1 41475    Yes                    No Female  62 58.04       No      No    No
2 41477    Yes                    No   Male  71 30.05      Yes      No   Yes
3 41479     No                    No   Male  52 27.56       No     Yes    No
4 41481     No                    No   Male  21 23.34       No     Yes    No
5 41482     No                    No   Male  64 33.64       No     Yes   Yes
6 41483    Yes                    No   Male  66 44.06      Yes      No    No
  prehypertension vegetarian covered.health
1              No         No             No
2              No         No            Yes
3              No         No             No
4              No         No             No
5              No         No            Yes
6              No         No            Yes
```

```
str(arthritis)
```

```
'data.frame':   4856 obs. of  12 variables:
 $ id                   : int  41475 41477 41479 41481 41482 41483 41485 41486 41487 41489 .
 $ status               : chr  "Yes" "Yes" "No" "No" ...
 $ heart.attack.relative: chr  "No" "No" "No" "No" ...
 $ gender               : chr  "Female" "Male" "Male" "Male" ...
 $ age                  : int  62 71 52 21 64 66 30 61 27 40 ...
 $ bmi                  : num  58 30.1 27.6 23.3 33.6 ...
 $ diabetes             : chr  "No" "Yes" "No" "No" ...
 $ alcohol              : chr  "No" "No" "Yes" "Yes" ...
 $ smoke                : chr  "No" "Yes" "No" "No" ...
 $ prehypertension      : chr  "No" "No" "No" "No" ...
 $ vegetarian           : chr  "No" "No" "No" "No" ...
 $ covered.health       : chr  "No" "Yes" "No" "No" ...
```

2. Prepare the data.

Convert categorical variables to factors, and set the reference outcome category.

Why reference matters?

- It determines which group is the baseline for interpretation.
- Here we model the odds of **status = "Yes"** relative to **"No"** (depending on coding; see note below).

3. Fit a logistic regression model.

Predictors:

- age
- gender
- bmi
- diabetes
- smoke

```
Call:
glm(formula = status ~ age + gender + bmi + diabetes + smoke,
    family = binomial, data = arthritis)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.702551   0.235275 -24.238  < 2e-16 ***
age          0.060980   0.002369  25.746  < 2e-16 ***
genderMale  -0.602215   0.074183  -8.118 4.74e-16 ***
bmi          0.050995   0.005474   9.315  < 2e-16 ***
diabetesYes  0.221479   0.100630   2.201   0.0277 *
smokeYes     0.551452   0.073706   7.482 7.33e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5861.9  on 4855  degrees of freedom
Residual deviance: 4754.4  on 4850  degrees of freedom
AIC: 4766.4

Number of Fisher Scoring iterations: 5
```

4. Compute Odds Ratios (OR)

$$OR = e^{\beta}$$

```
(Intercept)         age  genderMale         bmi diabetesYes    smokeYes
0.003337442 1.062877476 0.547597410 1.052317555 1.247921256 1.735772165
```

5. Which variables increase the odds of arthritis?

Rule:

- OR > 1 increases odds
- OR < 1 decreases odds

```
(Intercept)         age  genderMale         bmi diabetesYes    smokeYes
0.003337442 1.062877476 0.547597410 1.052317555 1.247921256 1.735772165
```

6. Compute 95% confidence intervals (OR scale).

```
Waiting for profiling to be done...
```

```
                  2.5 %      97.5 %
(Intercept) 0.002093571 0.005266448
age         1.058002639 1.067873530
genderMale  0.473264698 0.633017443
bmi         1.041106201 1.063698236
diabetesYes 1.024189001 1.519654333
smokeYes    1.502786941 2.006308314
```

- If the CI includes 1 $\rightarrow$ evidence of association is weak (at 5% level)
- If CI entirely $> 1 \rightarrow$ increased odds
- If CI entirely $< 1 \rightarrow$ decreased odds

7. Compute predicted probability of arthritis.

Predicted risk:

$$\hat{p}_i = P(Y_i = 1 \mid X_i)$$

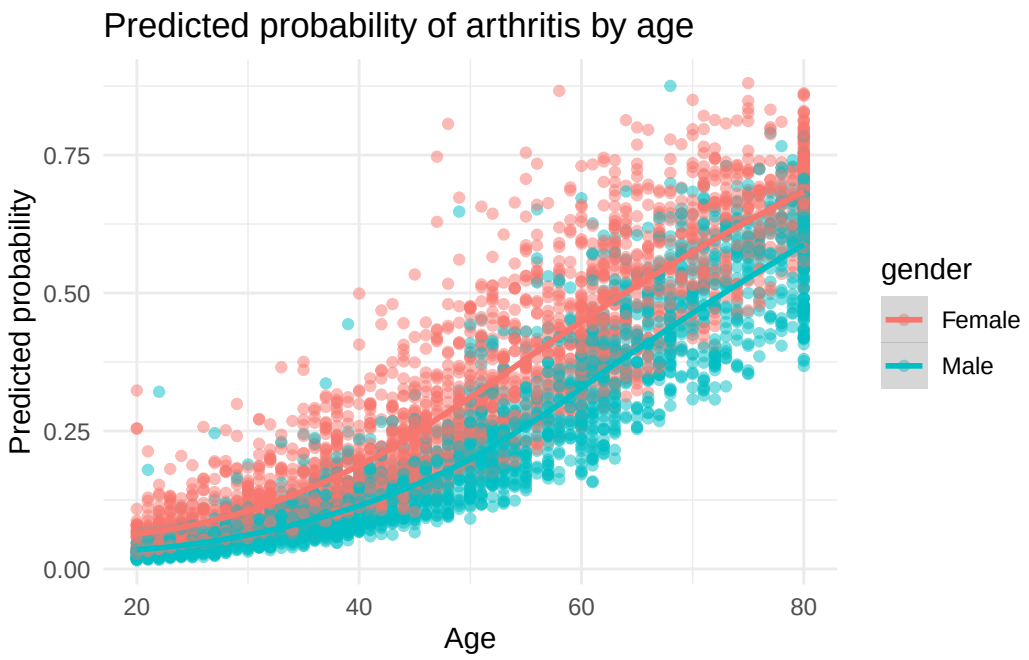
```
[1] 0.73845319 0.58179137 0.15079310 0.02116597 0.46626765 0.54691630
```

8. Visualize predicted probability vs age.

```
library(ggplot2)

ggplot(arthritis, aes(age, pred_prob, color = gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess") +
  labs(
    title = "Predicted probability of arthritis by age",
    x = "Age",
```

```
    y = "Predicted probability"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Teaching note:

- This is not a causal curve; it's model-based prediction given observed covariates.
- LOESS here is just for visualization, not the fitted logistic mean function.

9. Interpretation Questions (From ORs)

- Does age increase risk? (OR_age > 1?)
- Does BMI increase risk?
- Does smoking increase risk?
- Does gender affect risk?

10. Marginal effects of BMI and Age.

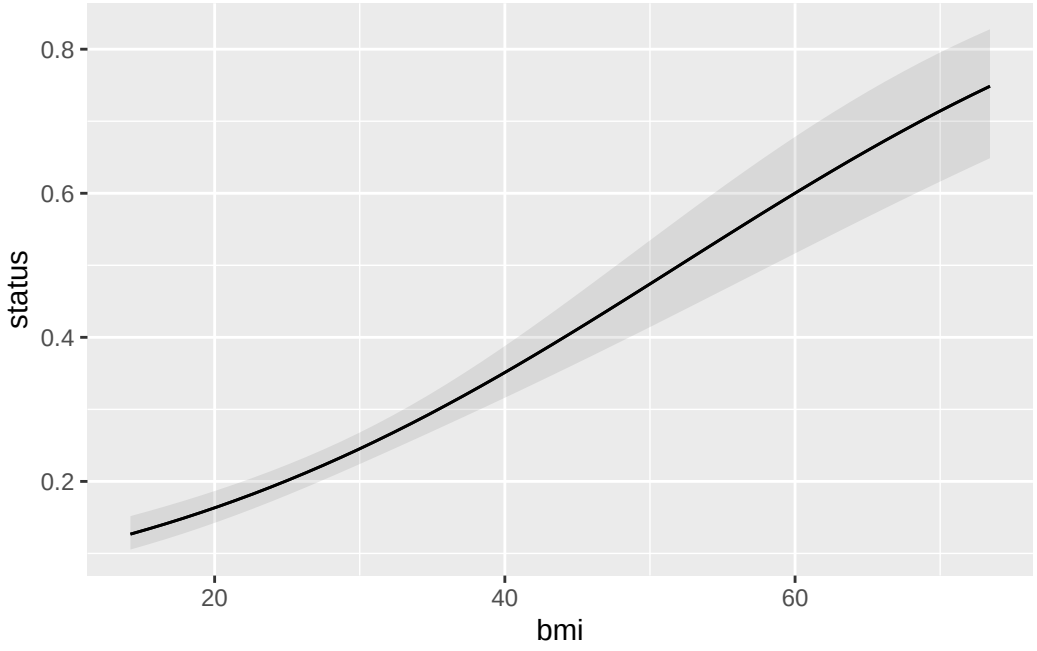
Coefficients are in log-odds so marginal effects express average probability change.

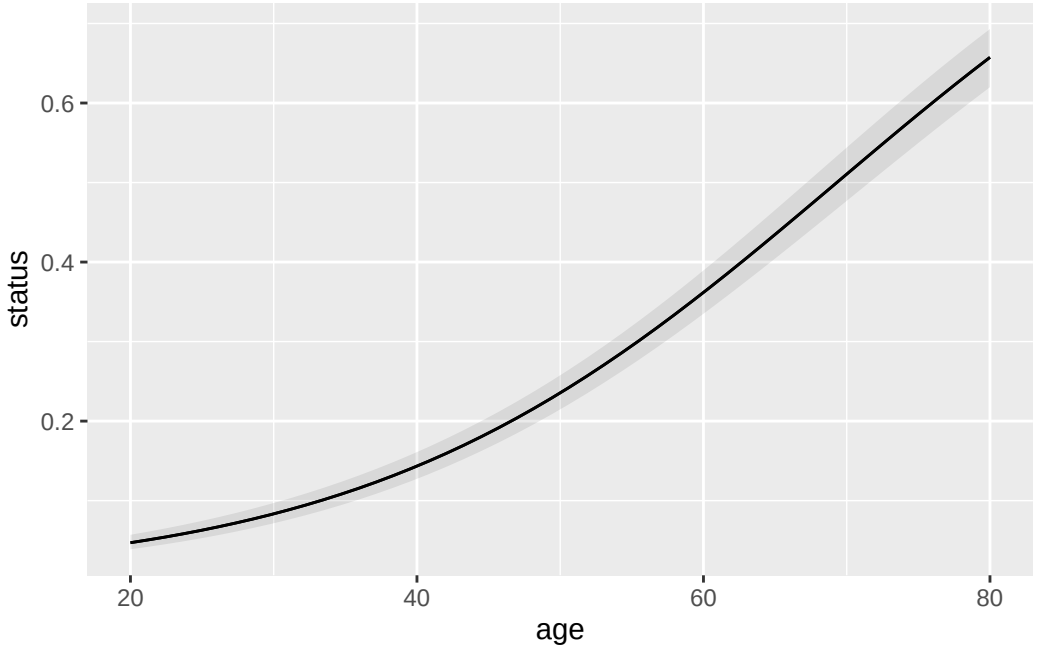
```r
library(marginaleffects)

mfx <- avg_slopes(fit)
mfx
```

```
     Term      Contrast Estimate Std. Error     z Pr(>|z|)     S    2.5 %
 age      dY/dX           0.00986   0.000282 34.95   <0.001 886.6  0.00931
 bmi      dY/dX           0.00824   0.000859  9.60   <0.001  70.0  0.00656
 diabetes Yes - No        0.03677   0.017112  2.15   0.0317   5.0  0.00323
 gender   Male - Female -0.09738   0.011770 -8.27   <0.001  52.8 -0.12045
 smoke    Yes - No        0.08958   0.011846  7.56   <0.001  44.5  0.06636
   97.5 %
  0.01041
  0.00993
  0.07030
 -0.07431
  0.11280

Type: response
```

- Marginal effects for BMI and age only

```
 Term Estimate Std. Error    z Pr(>|z|)     S   2.5 %  97.5 %
  age  0.00986   0.000282 34.9   <0.001 886.6 0.00931 0.01041
  bmi  0.00824   0.000859  9.6   <0.001  70.0 0.00656 0.00993

Type: response
Comparison: dY/dX
```





Interpretation:

- positive effect → increases predicted risk (on average)
- negative effect → decreases predicted risk

11. Predicting risk for new patients.

We now predict arthritis risk for specific individuals.

- Helper function

```
predict_risk <- function(age,
                         bmi,
                         gender = "Female",
                         diabetes = "No",
                         smoke = "No") {

  newdata <- data.frame(
    age = age,
    bmi = bmi,
    gender = factor(gender, levels = levels(arthritis$gender)),
    diabetes = factor(diabetes, levels = levels(arthritis$diabetes)),
    smoke = factor(smoke, levels = levels(arthritis$smoke))
  )

  predict(
    fit,
    newdata = newdata,
    type = "response"
  )
}
```

- Example: low-risk vs high-risk profiles

Low-risk:

- Female, Age 30, BMI 22, non-smoker, no diabetes

High-risk:

- Male, Age 70, BMI 35, smoker, diabetic

```
new_patients <- data.frame(
  age = c(30, 70),
  bmi = c(22, 35),
  gender = factor(c("Female","Male"),
                  levels = levels(arthritis$gender)),
  diabetes = factor(c("No", "Yes"),
                    levels = levels(arthritis$diabetes)),
  smoke = factor(c("No", "Yes"),
                 levels = levels(arthritis$smoke))
```

```
)

predict(fit, newdata = new_patients, type = "response")
```

```
          1           2
0.06001512 0.62751635
```