

Model Assessment

Logistic regression – Stat 230

Revisiting the Framingham model

- $Y = \text{CHD}$ ($0 = \text{no}$, $1 = \text{yes}$)
- $X = \text{participant's age, sex, total cholesterol, and systolic blood pressure}$

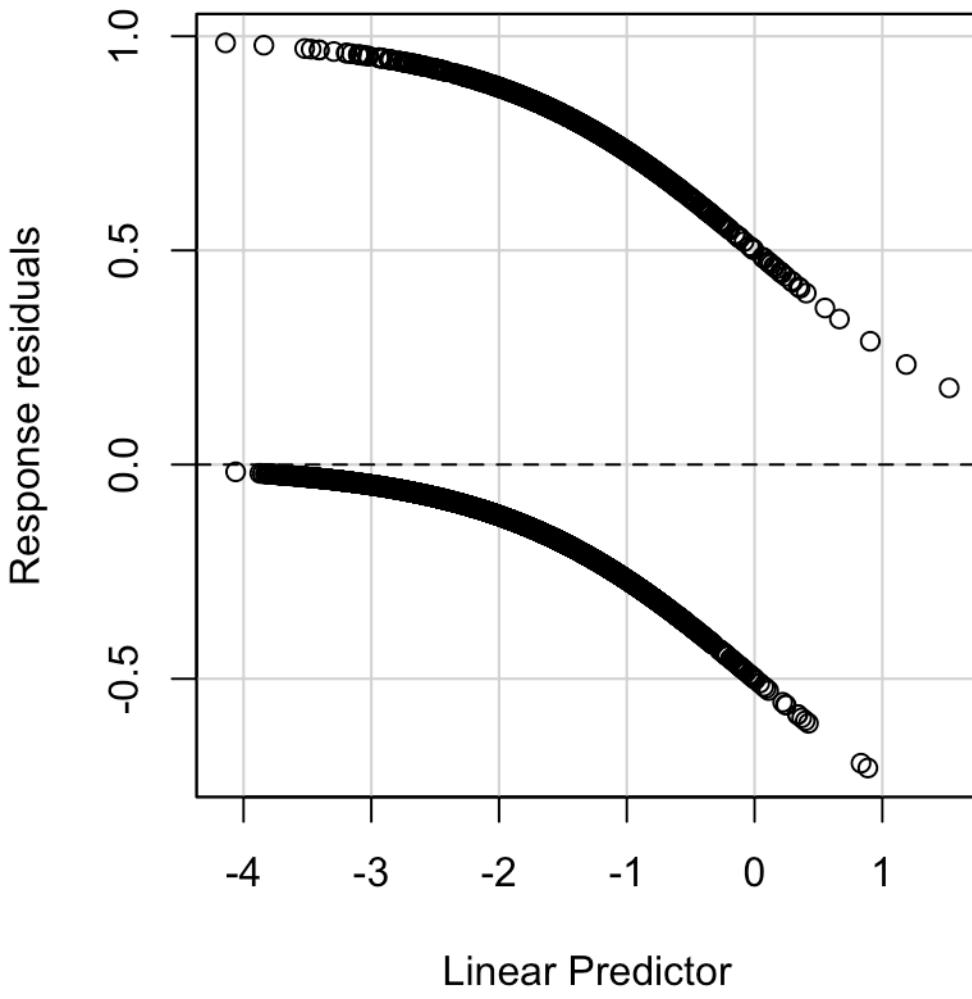
Model assumptions

1. **Linearity:** there is a linear relationship between the log odds (logit) and the predictors
2. **Independence:** no pairing, clustering, etc.
3. **Randomness:** we have a random sample from the population or the data were collected from a randomized experiment

Residual analysis

Response residuals

- $e_i = y_i - \hat{\pi}_i = \begin{cases} 1 - \hat{\pi}_i & \text{if } y_i = 1 \\ -\hat{\pi}_i & \text{if } y_i = 0 \end{cases}$
- Always be between -1 and 1



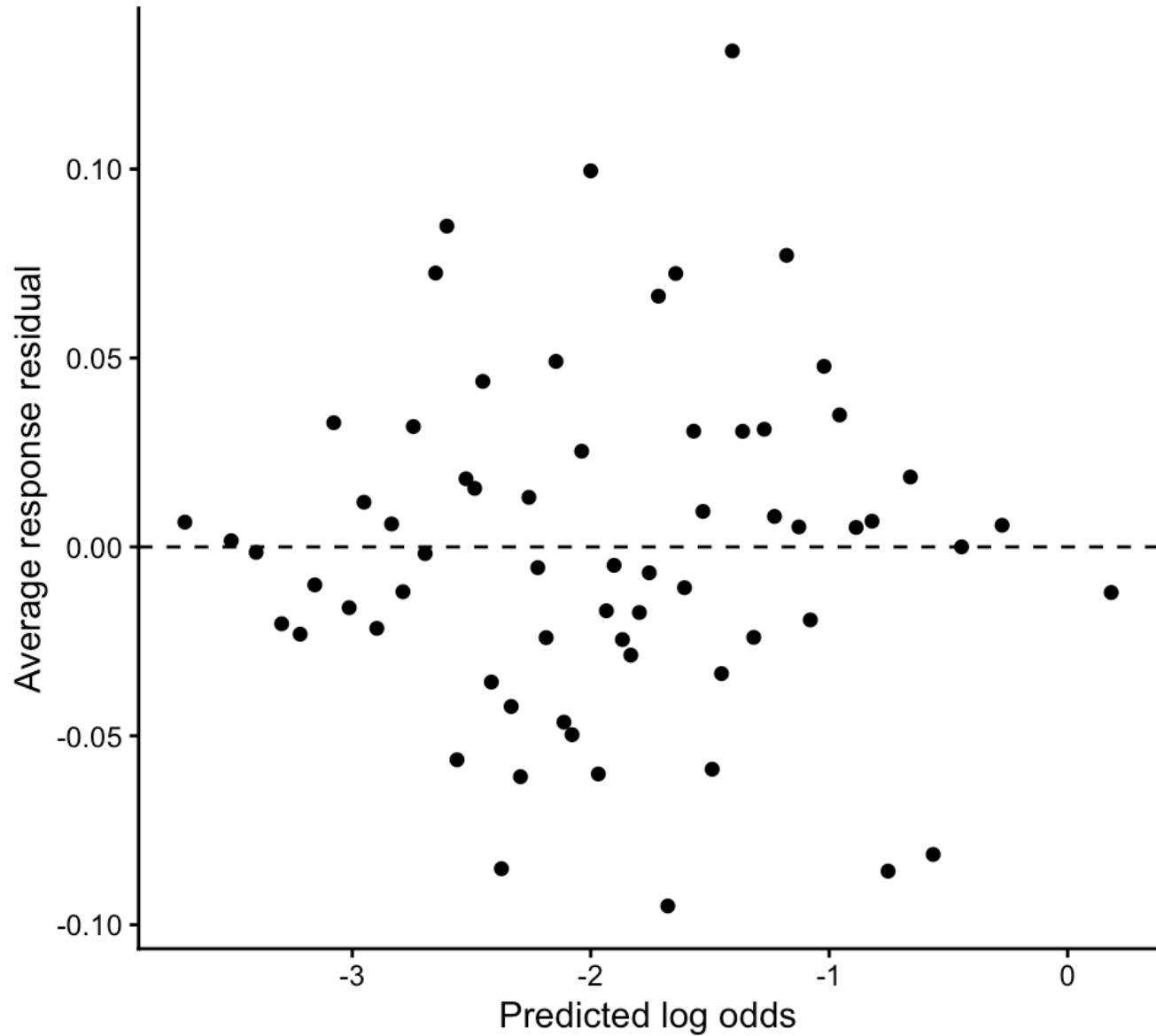
- Not normally distributed
- True distribution under correct model is unknown

Binned residuals

Gelman and Hill suggest binning (grouping) cases by a predictor and computing mean response residuals value within each group

- Calculate response residuals
- Order observations either by the values of the predicted log odds/probabilities, or by quantitative predictor variable
- Use the ordered data to create g bins of approximately equal size.
(Default value: $g = \sqrt{n}$)
- Calculate average residual value in each bin
- Plot average residuals vs. average predicted log odds, probability, or predictor value

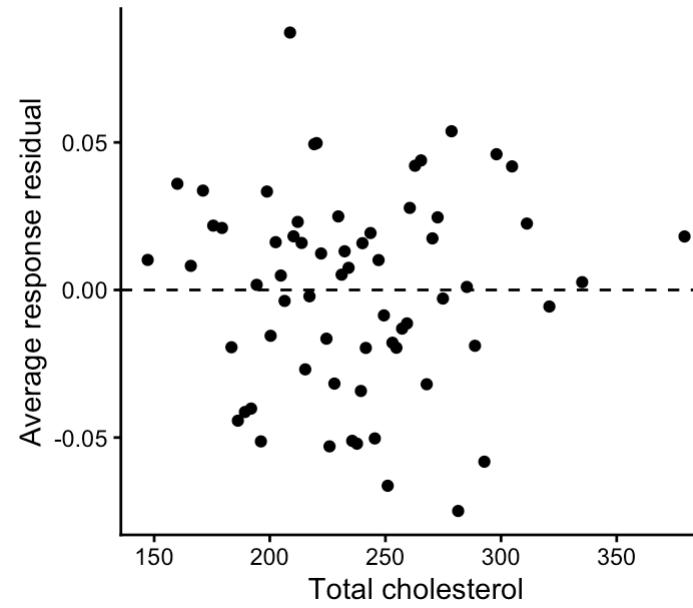
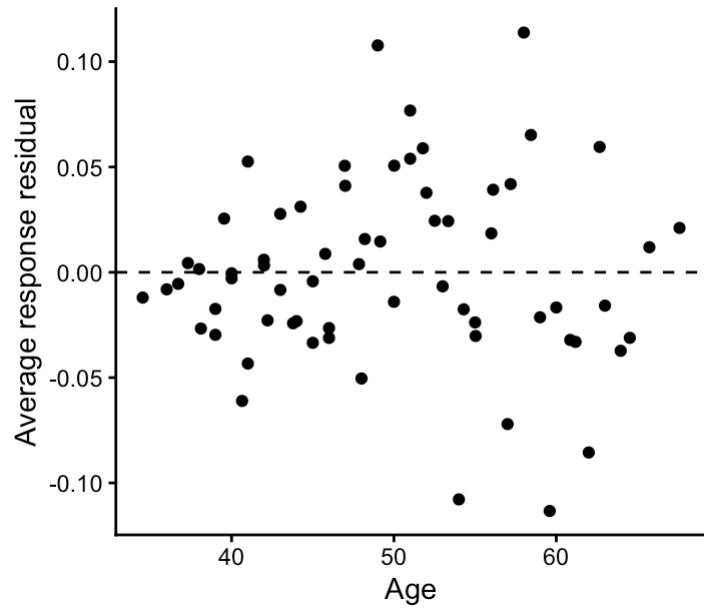
Binned residuals

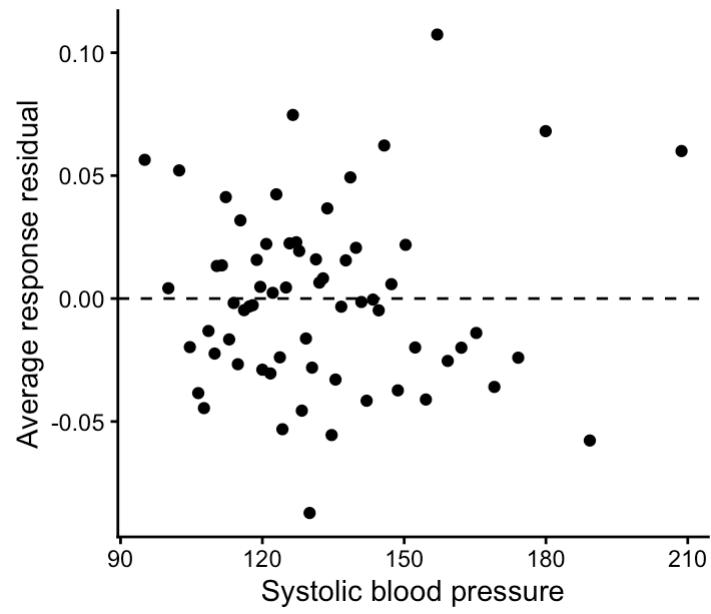


Binned residuals

- Nonlinear trend may be indication that squared term or log transformation of predictor variable required
- If bins have large average residuals (in magnitude)
 - Look at averages of other predictor variables across bins
 - Interaction may be required if large magnitude residuals correspond to certain combinations of predictor variables

Binned residuals





Residuals vs. categorical predictor

- Calculate average residual for each level of the predictor
- All the means should close to 0

male	average residual
0	-8.88e-13
1	-3.51e-13

Pearson residuals

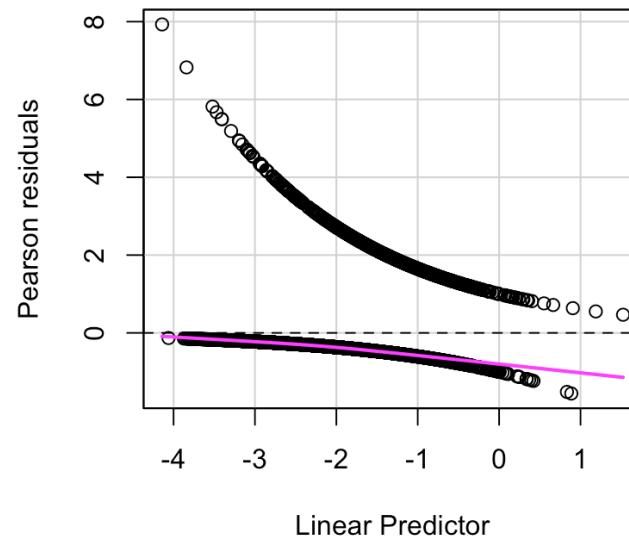
Standardize the distance between the observation (0 or 1) and the expected value (probability)

$$r_{p_i} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

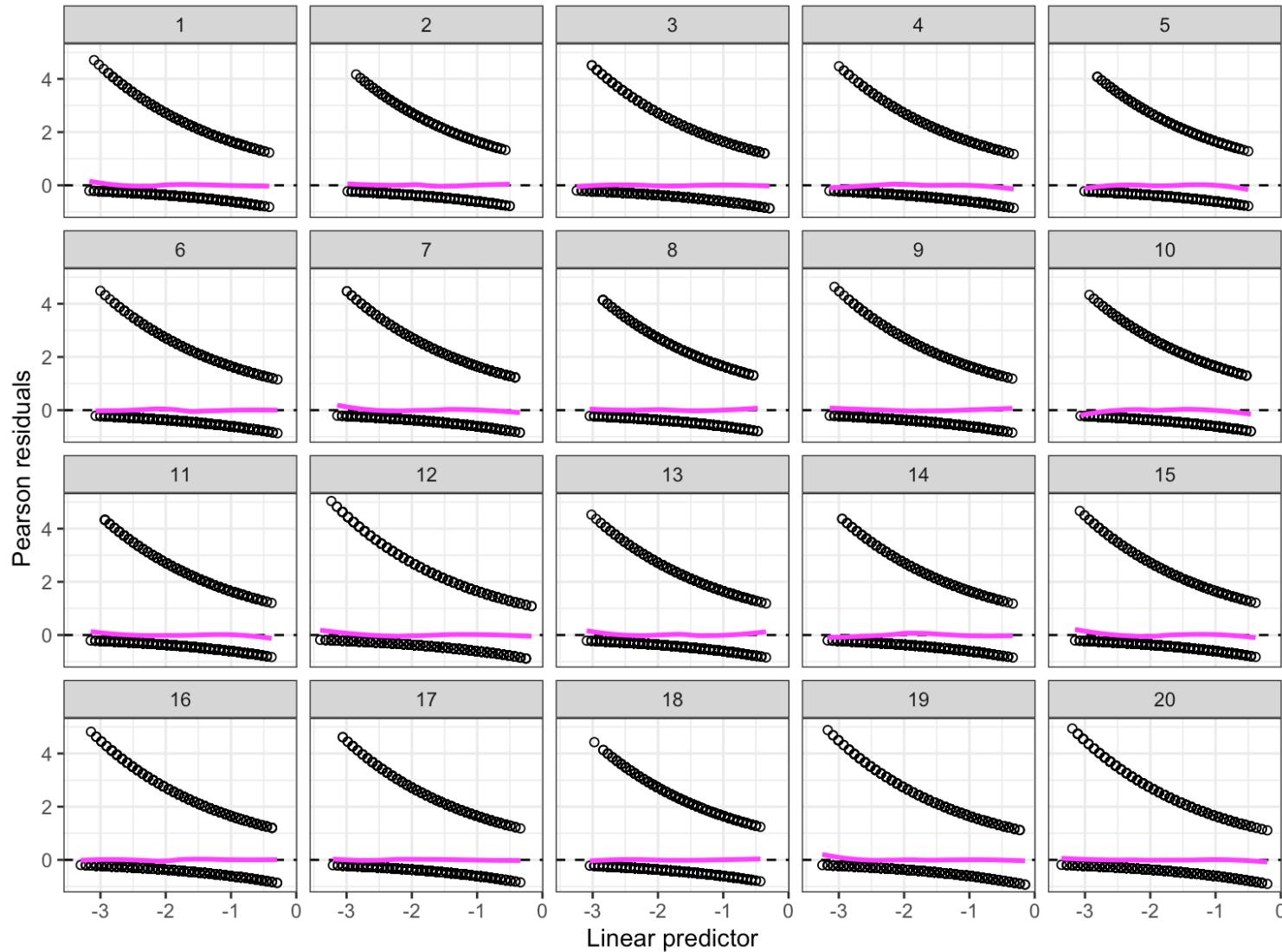
Pearson residuals vs. $\hat{\eta}$ or $\hat{\pi}$

- If the model is correct, then the expected (average) Pearson residual within each bin should be close to 0
- A smoother should be approximately horizontal line with intercept 0

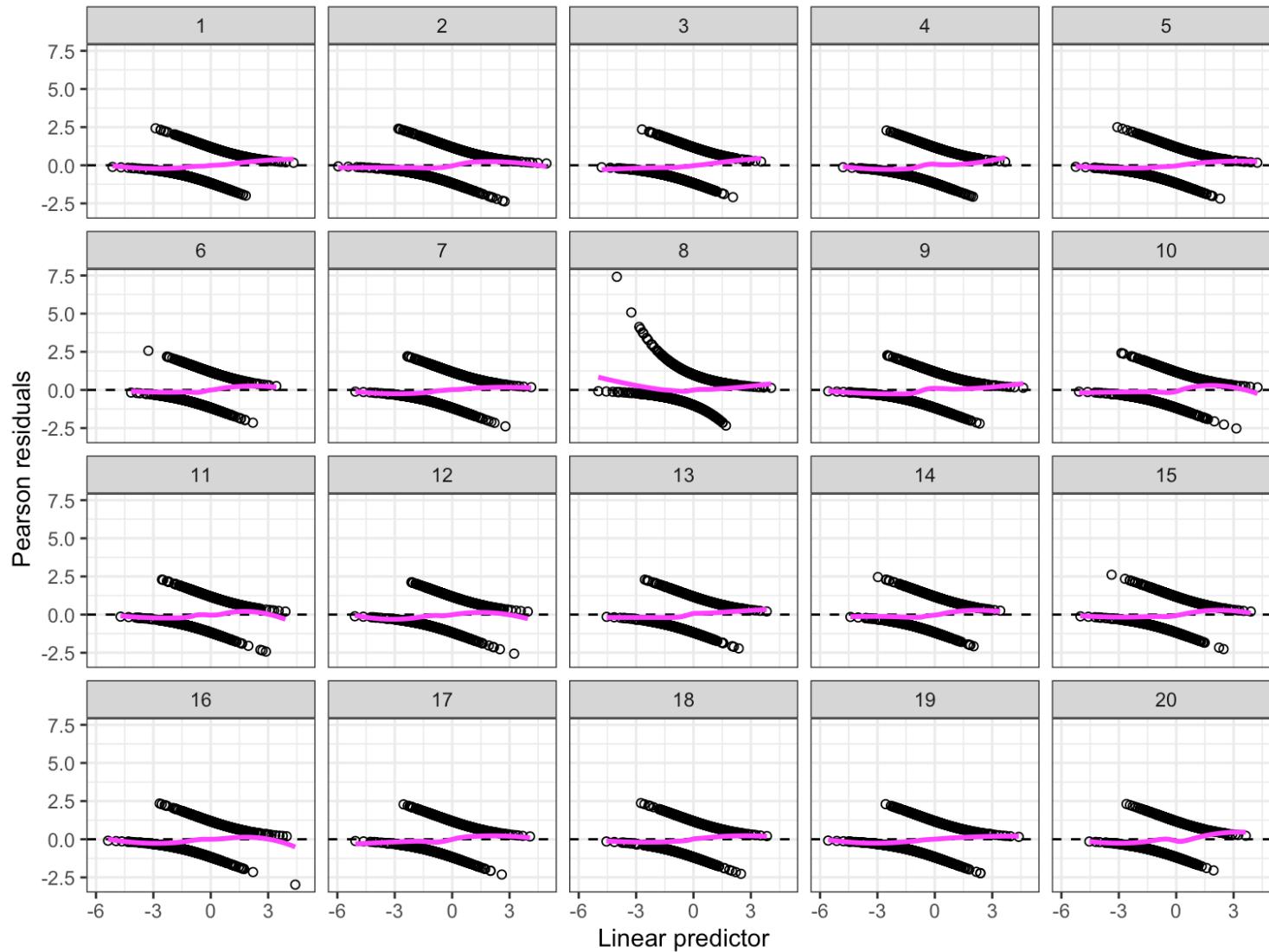
```
1 residualPlot(heart_mod, type = "pearson", smooth = TRUE)
```



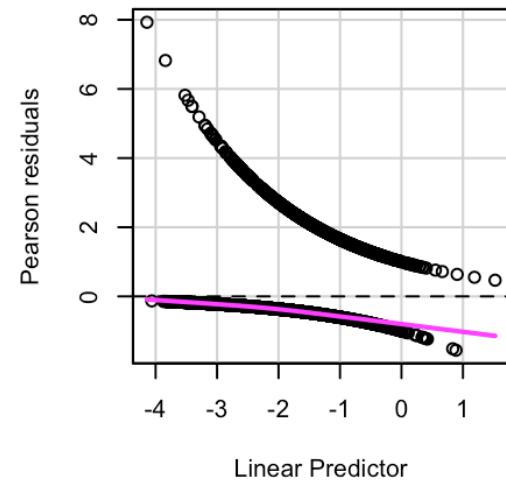
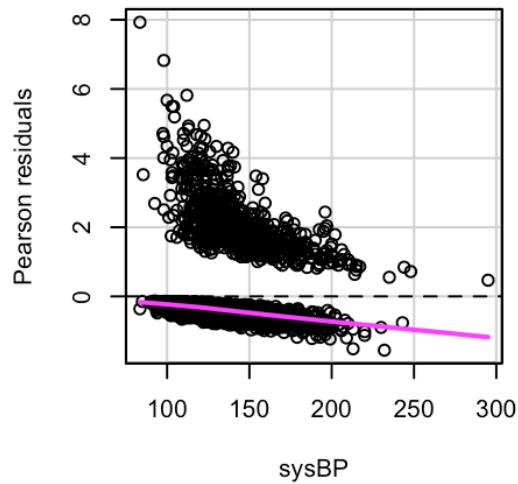
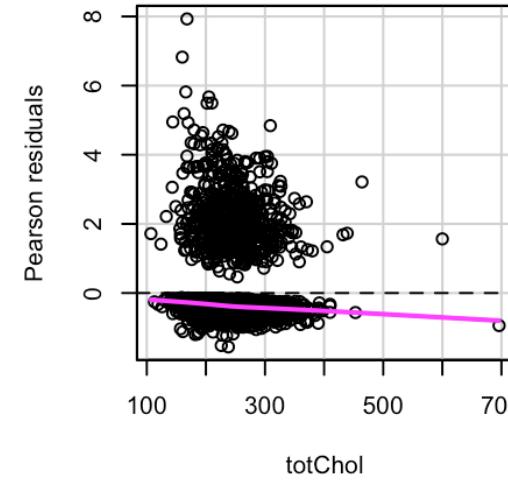
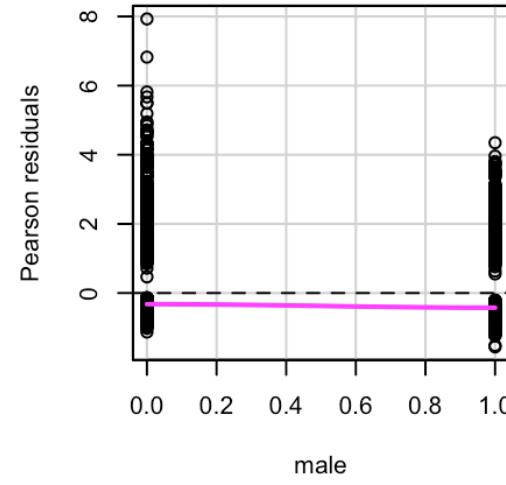
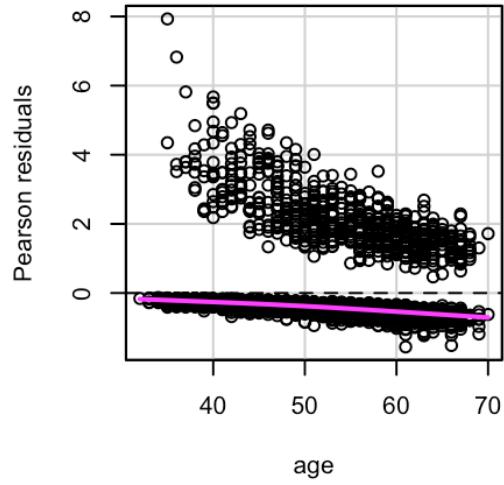
19 of these residual plots were created (simulated) from a model we know is correct, and one is from our actual data



Smoothers can help draw attention to major flaws, but there will always be variation from the horizontal!



Framingham model Pearson residuals



**Checking linearity is
hard!**

Empirical logit

Toy example: Our x values are all unique and range from 2-20. Let's divide the $n = 7$ cases into four groups according to their x value:

$$[0, 5], (5 - 10], (10 - 15], (15 - 20]$$

Empirical logit

For each group, compute the empirical probability and odds of success

Add 0.5 a success and failure to each group to eliminate computational problems when we see 0 S or F in a group.

$$\hat{p} = \frac{\text{\# successes in bin} + 0.5}{\text{\# cases in bin} + 1}$$

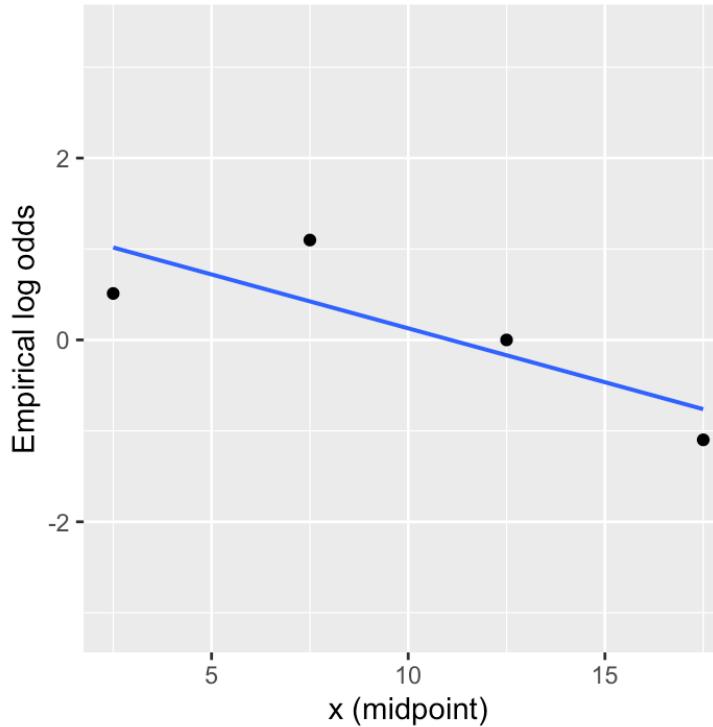
$$\hat{\omega} = \frac{\hat{p}}{1 - \hat{p}} = \frac{\text{\# successes in bin} + 0.5}{\text{\# failures in bin} + 0.5}$$

bin	[0,5]	(5-10]	(10-15]	(15-20]
-----	-------	--------	---------	---------

emp.prob	0.625	0.750	0.500	0.250
emp.odds	1.6666667	3.0000000	1.0000000	0.3333333

Empirical logit plots

Plot the empirical log odds vs. midpoint for each bin.
for linearity, strength, direction, curvature, etc.

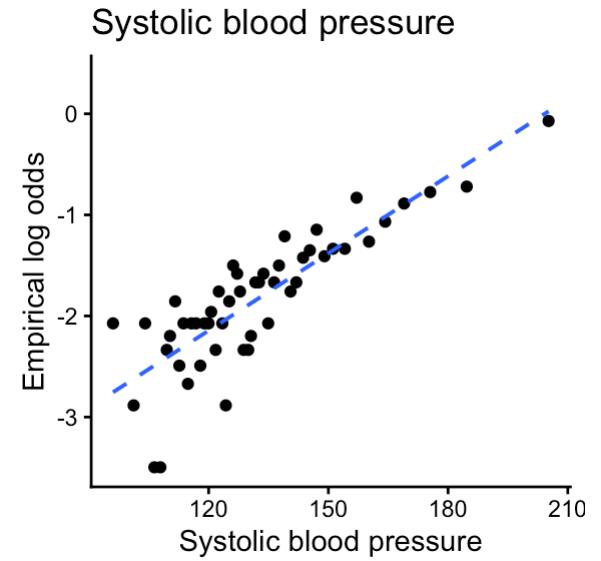
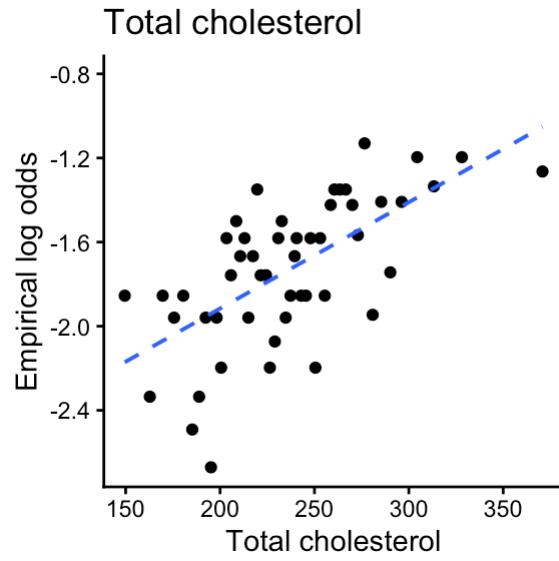
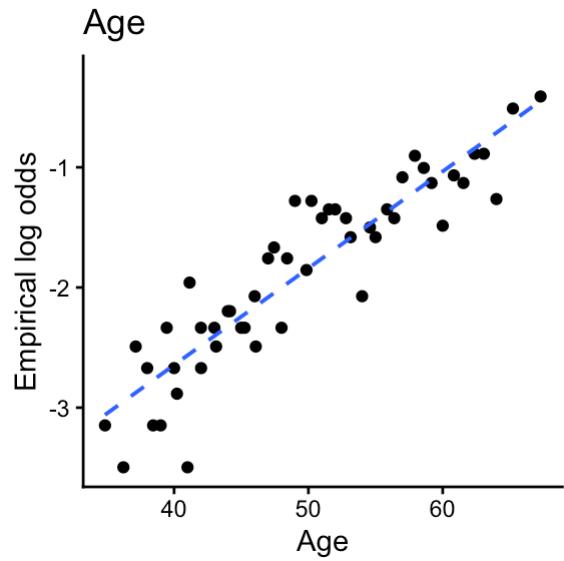


Setting the bin width

Target: relatively constant probability/odds of success within each group *and* enough cases in each group to estimate prob/odds

- Bins that are too wide (e.g. too few bins) means this assumption is likely not met
- Bins that are too narrow (e.g. too many bins) means that we don't have enough cases within each bin to accurately estimate the probability/odds of success.

Back to the heart study



Case influence

We can still calculate and interpret leverage and Cook's distance as before

