Additional Exercise 4.1

Data: presence or absence of CHD condition in 100 patients with recorded age. The most natural notation is,

$$\begin{aligned} y_i &= \text{presence (1) or absence (0) of CHD,} \\ x_i &= \text{age,} \end{aligned}$$

for the i'th patient, $i = 1, \ldots, 100$.

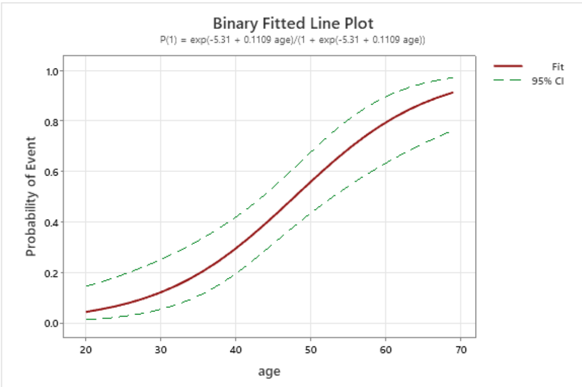
Statistical model
The disease outcomes $y_1, \ldots, y_{100}$ are assumed independent, and with $p_i = \Pr(y_i = 1)$ the linear relation on logit scale is parametrized as,

$$\text{logit}(p_i) = \beta_0 + \beta_1 \cdot x_i.$$

The Minitab output (abbreviated) from fitting this model in the Binary Fitted Line Plot menu is shown below.

HS04_1.CSV

Binary Fitted Line: chd versus age



Response Information

Variable	Value	Count
chd	1	43 (Event)
	0	57
	Total	100

Regression Equation

P(1) = exp(-5.31 + 0.1109 age)/(1 + exp(-5.31 + 0.1109 age))

Coefficients

Term	Coef	SE Coef	VIF
Constant	-5.31	1.13	
age	0.1109	0.0241	1.00

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
age	1.1173	(1.0658, 1.1713)

Analysis of Variance

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	29.31	29.310	29.31	0.000
age	1	29.31	29.310	29.31	0.000
Error	98	107.35	1.095		
Total	99	136.66			

Interpretation of results (Q1):
Estimated parameters with 95% confidence intervals are:

$$\begin{aligned} \text{intercept} \quad &: \quad \hat{\beta}_0 = -5.31, \qquad \text{95\% CI: } -5.31 \pm 1.96 \cdot 1.13 = -5.31 \pm 2.215, \\ \text{slope} \quad &: \quad \hat{\beta}_1 = 0.1109, \qquad \text{95\% CI: } 0.1109 \pm 1.96 \cdot 0.0241 = 0.1109 \pm 0.047. \end{aligned}$$

From the estimated slope, we compute: OR $= \exp(\hat{\beta}_1) = 1.12$. This is for the comparison of two patients one year apart, where the CHD risk (odds) of the older patient is 1.12 times larger. The estimated intercept corresponds to a patient of age 0, which has no meaningful interpretation. We could center age if we wanted to make the intercept more meaningful, but there is no pressing need to do so.

For testing the hypothesis $\mathrm{H}_0 : \beta_1 = 0$ (no association between age and CHD risk), we have two options. The "Analysis of Variance" table shows the likelihood-ratio test as: $G^2 = 29.31$, corresponding to $P < 0.0005$ in the $\chi^2(1)$-distribution. We can also compute the Wald test as: $z = 0.1109/0.0241 = 4.61$, which is obviously very significant as well in $\mathrm{N}(0, 1)$. There is clear evidence of an impact of age on CHD risk.

Assessing the fit of the logistic regression model (Q2):
The listing above does not include any goodness-of-fit tests, but we can get those from the more comprehensive Binary Logistic Regression menu, as shown below.

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	98	107.35	0.243
Pearson	98	101.94	0.372
Hosmer-Lemeshow	8	0.89	0.999

For ungrouped data, the goodness-of-fit statistics based on the deviance and the Pearson chi-square are useless. The Hosmer-Lemeshow test is okay, but not always particularly powerful. One simple alternative method is to add a quadratic term to the linear relation. Without showing the results, it turns out that the improvement in the attained $-2 \ln L$ is minimal: from 107.35 to 107.29. Naturally, this difference ($G^2 = 0.06$) would be totally non-significant as a likelihood-ratio test of the linear model against the quadratic model. The same conclusion is reached when looking at the Wald test ($z = 0.26$) for the quadratic coefficient. There is no indication that a quadratic term would improve the model fit.

Grouping of age, and comparison between linear and categorical effects of age groups (Q3):
The Minitab listings on the next page show the results for the logistic models with `agegrp` (age coded as specified in the question, using the `Data-Recode` menu in Minitab) as continuous and categorical terms, respectively.

HS04_1.CSV

Binary Logistic Regression: chd versus agegrp

Response Information

Variable	Value	Count
chd	1	43 (Event)
	0	57
	Total	100

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -5.20 + 0.1087 agegrp

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-5.20	1.12	-4.65	0.000	
agegrp	0.1087	0.0237	4.58	0.000	1.00

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
agegrp	1.1148	(1.0641, 1.1679)

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	98	108.36	0.223
Pearson	98	101.31	0.389
Hosmer-Lemeshow	5	0.38	0.996

Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	1	20.96	0.000
agegrp	1	20.96	0.000

HS04_1.CSV

Binary Logistic Regression: chd versus agegrp

Response Information

Variable	Value	Count
chd	1	43 (Event)
	0	57
	Total	100

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -2.20 + 0.0 agegrp_25.400 + 0.33 agegrp_32.000 + 1.10 agegrp_36.917 + 1.50 agegrp_42.333 + 2.04 agegrp_47.231 + 2.71 agegrp_51.875 + 3.38 agegrp_56.882 + 3.58 agegrp_63.000

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-2.20	1.05	-2.08	0.037	
agegrp					
32.000	0.33	1.30	0.25	0.802	2.64
36.917	1.10	1.25	0.88	0.378	3.06
42.333	1.50	1.19	1.27	0.205	3.83
47.231	2.04	1.19	1.71	0.086	3.76
51.875	2.71	1.28	2.11	0.035	2.76
56.882	3.38	1.20	2.82	0.005	3.65
63.000	3.58	1.32	2.72	0.007	2.53

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	92	107.96	0.122
Pearson	92	100.00	0.267
Hosmer-Lemeshow	5	0.00	1.000

Analysis of Variance

		Wald Test	
Source	DF	Chi-Square	P-Value
Regression	7	21.86	0.003
agegrp	7	21.86	0.003

The deviance goodness-of-fit test is the likelihood-ratio test comparing a full model with `agegrp` as a factor with the reduced model with `agegrp` as a continuous predictor. The calculation goes as follows,

$$G^2 = 108.36 - 107.96 = 0.40, \quad \text{df} = 7-1 = 6, \quad P = \Pr(\chi^2(6) > 0.40) > 0.99.$$

As seen, the test is absolutely non-significant in the reference chi-square distribution with 6 df. There is no indication of lack of fit. Additionally, one could compute residuals and inspect those, but with such a non-significant goodness-of-fit test it is almost certain that the residuals would show no major model deviations. The test does utilize the grouped predictor (`agegrp`), but it nevertheless confirms our impression from the other analyses above that the logistic regression for the ungrouped age predictor fits really well. Indeed, the loss of fit from the grouping appears fairly small (the $-2\ln L$ values are 108.36 and 107.35).

Presentation of results (Q4):
The interpretation of the estimated coefficient for age in terms of odds-ratio was already mentioned in Q1, and the Minitab output also showed a graph of the predicted probabilities as a function of age. There is hardly more for us to do.