F-24 DATA-240 Data Mining / Analytic

Homework-2

Prayag Nikul Purani (01741673 7)

Problem 3 Logistic regression concepts

a) Eqn for a logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots \beta_n X_n$$

$$\log\left(\frac{p}{1-p}\right) = -15.3001 + (0.0018 \times \text{factor 1}) + (-0.0061 \times \text{factor 2}) + (0.0057 \times \text{factor 3}) + (0.0066 \times \text{factor 4}) + (0.0071 \times \text{factor 5}) + (0.1113 \times \text{factor 6}) + (-0.0098 \times \text{factor 7}) + (0.0686 \times \text{age})$$

b) (i) $z = \frac{\text{coefficients}}{\text{Std error}}$

intercept $\Rightarrow z = \frac{-15.3001}{0.0015} = \boxed{-10,200.067}$

factor 1 $\Rightarrow z = \frac{0.0018}{0.0103} = \boxed{0.17476}$

factor 2 $\Rightarrow$ $z = \frac{-0.0061}{0.0105} = \boxed{-0.5809}$

factor 3 $\Rightarrow$ $z = \frac{0.0057}{0.0028} = \boxed{2.0357}$

factor 4 $\Rightarrow$ $z = \frac{0.0066}{0.0038} = \boxed{1.7368}$

factor 5 $\Rightarrow$ $z = \frac{0.0071}{0.0199} = \boxed{0.3568}$

factor 6 $\Rightarrow$ $z = \frac{0.1113}{0.0492} = \boxed{2.2622}$

factor 7 $\Rightarrow$ $z = \frac{-0.0098}{0.0037} = \boxed{-2.6486}$

Age $\Rightarrow$ $z = \frac{0.0686}{0.0224} = \boxed{3.0625}$.

(ii) $p = 2 \times (1 - CDF(|z|))$

Intercept

$p = 2 \times (1 - CDF(1020.0667))$

$\boxed{p = 0}$

z is extremely large so p-value is essentially 0

factor 1 $p = 2 \times (1 - CDF(0.1748))$

$= 0.86$

factor 2 $p = 2 \times (1 - CDF(0.5810))$

$= 0.56$

factor 3 $p = 2 \times (1 - CDF(2.0357))$

$= 0.0418$

factor 4 $p = 2 \times (1 - CDF(1.7368))$

$= 0.0824$

factor 5 $p = 2 \times (1 - CDF(0.3568))$

$= 0.7213$

factor 6 $p = 2 \times (1 - CDF(2.2622))$

$= 0.0237$

factor 7 $p = 2 \times (1 - CDF(2.6486))$

$= 0.0081$

Age $p = 2 \times (1 - CDF(3.0625))$

$= 0.0022$

(iii) Lower bound = $\beta - 1.96 \times SE$
upper bound = $\beta + 1.96 \times SE$ } for 95% CI

$CI = [LB, UB]$

Inter apb

$LB = -15.30001 - (1.96 \times 0.0015)$
$= -15.30304$
$UB = -15.30001 + (1.96 \times 0.0015)$
$= -15.29716$

$CI = [-15.30304, -15.29716]$

factor 1

$LB = 0.0018 - (1.96 \times 0.0103) = -0.018388$
$UB = 0.0018 + (1.96 \times 0.0103) = 0.021988$

$CI = [-0.018388, 0.021988]$

factor 2

$LB = -0.0061 - (1.96 \times 0.0105) = -0.0266$
$UB = -0.0061 + (1.96 \times 0.0105) = 0.01448$

$CI = [0.0266, 0.01448]$

factor 3

$LB = 0.0057 - (1.96 \times 0.0028) = 0.000212$
$UB = 0.0057 + (1.96 \times 0.0028) = 0.011188$

$CI = [0.000212, 0.011188]$

factor 4

$LB = 0.0066 - (1.96 \times 0.0038) = -0.000848$
$UB = 0.0066 + (1.96 \times 0.0038) = 0.014048$

$CI = [-0.000848, 0.014048]$

factor 5 $LB = 0.0071 - (1.96 \times 0.0199) = -0.031904$

$UB = 0.0071 + (1.96 \times 0.0199) = 0.046104$

CI [-0.0319, 0.0461]

factor 6 $LB = -0.0098 - (1.96 \times 0.0492) = 0.01486$

$UB = -0.0098 + (1.96 \times 0.0492) = 0.20773$

CI = [0.01486, 0.20773]

factor 7 $LB = -0.0098 - (1.96 \times 0.0037) = -0.01705$

$UB = -0.0098 + (1.96 \times 0.0037) = -0.002548$

CI = [-0.01705, -0.00254]

Age $LB = 0.0686 - (1.96 \times 0.0224) = 0.024698$

$UB = 0.0686 + (1.96 \times 0.0224) = 0.112504$

CI = [0.024698, 0.112504]

CY

Intercept
- p-value = 0.00 (significant)
- CI = [-15.303, -15.297] ⇒ no '0' in it
- ⇒ Significance → Yes

factor 1
- p-value = 0.861 (no)
- CI = [-0.018, 0.0219] ⇒ (include '0')
- ⇒ Significance → NO

factor 2
- p-value = 0.5615 (no significant)
- CI = [-0.0266, 0.01448] → (includes '0')
- Significance → No

factor 3
- p-value = 0.0418 (Significance)
- CI = [0.00012, 0.01188] → (no '0' in it)
- Significance → Yes

factor 4
- p-value = 0.0824 — (not significant)
- CI = [-0.000848, 0.014] — (includes '0')
- → Significance → No

factor 5
- p-value = 0.721 (not significant)
- CI = [0.0148, 0.2077] — (no '0' in it)
- → Significance → Yes

factor 6
- p-value = 0.0237 (Significance)
- CI = [0.0148, 0.2077] — (no '0' in it)
- Significance = Yes

Age factor
- p-value = 0.0022 (significance)
- CI = [0.02469, 0.1125] — (no '0' in it)
- Significance = Yes

factor 7 { p-value = 0.0081 — (significant)
CI = [-0.017, -0.0025] — (no '0' in it)
Significance — Yes.

Analysis of each factor

1) P-value < 0.05
2) CI excluding '0' } for 95% of confidence

1) Age ⟹ b_{coff} = 0.0686 ⟹ +ve so if age ↑ the log-odds of events (death due to the disease) ↑
⟹ 1 unit increase of age, odds increase by 0.0686

Also it is a significant feature as p-value < 2 & CI doesn't have '0' in it.

Factor 7 ⟹ b_{coff} = -0.0098 ⟹ -ve if factor 7 ↑ log-odds decreases
⟹ 1 unit increase of factor 7, the log-odds decreases by 0.0098.

Also it is a significant feature as p-value << 2 & CI don't have '0' in it

e7 (i) a 100 units decreases in factor 7

$$\beta = -0.0098 \times 100 = -0.98$$

$$e^{-0.98} \simeq = 0.3753$$

$$100 \times (1 - 0.3753) = 62.45\%$$

a 100 units decrease in factor 7 is associated with a 62.47% reduction in odds of death.

(ii) an additional year of age

$$\beta = 0.0686 \times 1 = 0.0686$$

$$e^{0.0686} = 1.071$$

$$(1.071 - 1) \times 100 = 7.1\%$$

an year of age associated with 7.1% increase in odds of death.

b) $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$

$= -15.3001 + (0.0018 \times 125) + (-0.0061 \times 105)$
$+ (0.0057 \times 235) + (0.0066 \times 105) +$
$(0.0071 \times 12.5) + (0.1113 \times 42) +$
$(-0.0098 \times 475) + (0.0686 \times 50)$

$= -9.81085$

$$p = \frac{1}{1 + e^{-(-9.81085)}} = \frac{1}{1 + 18298.748}$$

$$= \frac{1}{18299.748}$$

$$p = 0.0005485$$

$\Rightarrow$ So probability of death for 50 year old woman with the given factor values is 0.0548%