STA 303H1S / STA 1002HS: Logistic Regression Practice Problems
SOLUTIONS

1. (a) There were no females over 50. Any comparisons for older people must be based on the assumption that the model still holds and this cannot be verified with these data. Extrapolation is a problem for logistic regression, just as it is for linear regression.

   (b) Males and females might have different tasks and survival could be associated with task.

   (c) i. $3.2 - 0.078 age - 1.6 I_{male}$

   ii. Same equation but with all coefficients negative.

   (d) Estimated probabilities for women are

$$\hat{\pi} = \frac{e^{1.6-0.078age+1.6}}{1+e^{1.6-.078age+1.6}}$$

   and for men are

$$\hat{\pi} = \frac{e^{1.6-0.078age}}{1+e^{1.6-.078age}}$$

   So estimated survival probabilities for males are 0.41 at age 25 and 0.091 at age 50. For females, the estimates are 0.78 at age 25 and 0.33 at age 50.

   (e) Set the estimated log-odds to zero and solve for age. For females, the age of 50% survival is 41.0 years; for males it is 20.5 years.

   (f) Since females are coded as 1 and males as -1, females versus males corresponds to a change of 2 in the explanatory variable `sex`. So the odds ratio for females versus males (of the same age) is found by exponentiating 2 times the estimated coefficient for `sex`.

   (g) The test statistic is $13.326 - 10.127 = 3.199$. Under the null hypothesis that the reduced model adequately fits the data, this is an observation from a chi-square distribution with 1 degree of freedom (since there is 1 less parameter in the reduced model), giving an estimated p-value between 0.05 and 0.1. So there is weak evidence that the reduced model is not adequate; that is, there is weak evidence that 30 is not the age at which the probability of survival for females is 0.5.

   (h) i. The deviance for the model with the interaction term is 47.346 and for the model without the interaction term the deviance is 51.256. The LRT with null hypothesis that the coefficient of the interaction term is zero has test statistic $51.256 - 47.346 = 3.91$. Under the null hypothesis this is an observation from a chi-square distribution with 1 degree of freedom. From tables, we can estimate the p-value as $0.025 < p < 0.05$. So there is moderate evidence that the coefficient of the interaction term is not 0.

   ii. (1) The estimated odds ratio is $e^{-.0325(10)-0.1616(10)} = 0.14$.

   (2) The estimated odds ratio is $e^{6.9267-0.1616age}$. Note that it differs with the age.

   iii. No. If there is no interaction term, $\exp(\beta_1)$ is the odds ratio for a person that is 1 year older than another person, but in the presence of an interaction term, the odds ratio must be adjusted for gender. With an interaction, $\exp(\beta_1)$ is the odds ratio for a 1 year change in age for males only.

(i)

Explanatory variables in model	AIC	Rank by AIC	SC	Rank by SC
age, sex, age*sex	55.346	1	62.573	1
age, sex, age*sex, age^2	55.830	2	64.863	3
age, sex	57.256	3	62.676	2
age, sex, age*sex, age^2, age$^2 *$ sex	57.361	4	68.201	4

Both AIC and SC choose the model with age, sex and their interaction as the model that best fits the data. This is not surprising, considering that all of these have coefficients that are statistically significantly different from 0 (using the likelihood ratio test for the interaction term). Both AIC and SC least favour the most complicated model. The other models are ranked differently, reflecting that SC penalizes more complicated models more severely than AIC. As a rule-of-thumb, a difference of 2 or less in AIC indicates no real difference in the fit of the model (a difference in AIC greater than 10 is considered an important difference), so there is no overwhelming evidence from AIC in favour of any of these models. A parsimonious approach would be to select the simplest model.

2. $\log\left(\frac{Y_i}{1-Y_i}\right)$ is undefined if $Y_i = 1$ or 0.

3. (a) Maximum likelihood estimates are: $\hat{\beta}_0 = -4.7393$, $\hat{\beta}_1 = 0.0677$, and $\hat{\beta}_2 = 0.5986$. The fitted response function is $\text{logit}(\hat{\pi}) = -4.74 + 0.068\,\text{income} + 0.60\,\text{age}$.

(b) $\exp(\hat{\beta}_1) = 1.07$ so an increase of \$1000 in income is associated with a 7% increase in the odds of purchasing a new car in the next year.
$\exp(\hat{\beta}_2) = 1.82$ so an increase of 1 year in the age of the oldest family automobile is associated with an 82% increase in the odds of purchasing a new car in the next year.

(c) $\hat{\pi} = \frac{\exp(-4.74+0.068(50)+0.60(3))}{1+\exp(-4.74+0.068(50)+0.60(3))} = 0.61$

(d) The 0.005 quantile from a standard normal distribution is 2.576. An approximate 99% confidence interval for the coefficient of income is $0.0677 \pm 2.576(0.0281) = (-0.0047,\ 0.14)$. An approximate 99% confidence interval for the odds ratio for families whose incomes differ by \$20 thousand is $(e^{-0.0047(20)},\ e^{0.14(20)}) = (0.91,\ 16.5)$.

(e) $H_0:\ \beta_2 = 0$ versus $H_a:\ \beta_2 \neq 0$.
The test statistic is $(0.5986/0.3901)^2 = 2.35$. Under the null hypothesis, this is approximately an observation from a chi-square distribution with 1 degree of freedom. R gives the p-value as 0.1249. From a chi-square table we can say that $0.1 < p < 0.9$. We can get a more accurate estimate of the p-value by using the fact that $0.5986/0.3901 = 1.53$ is approximately an observation from a standard normal distribution. The p-value is then $2(1 - 0.9370) = 0.126$. So the data are consistent with the coefficient of age being 0.

(f) The models to be compared are $\text{logit}(\pi) = \beta_0 + \beta_1\,\text{income} + \beta_2\,\text{age}$ and $\text{logit}(\pi) = \beta_0 + \beta_1\,\text{income}$. The likelihood ratio test statistic is $39.305 - 36.690 = 2.615$. The estimated p-value from the chi-square table with 1 degree of freedom is $0.1 < p < 0.9$. The conclusion is consistent with the Wald test.

(g) The test statistic is $36.690 - 34.253 = 2.437$. Under the null hypothesis that the coefficients of all the second-order terms are zero, this is approximately an observation from a chi-square distribution with 3 degrees of freedom. From the chi-square table, the p-value is between 0.1 and 0.9 we conclude that no significant contribution to the model is being made by the 3 second-order terms.

4. (a) The R output warns us that "algorithm did not converge" and "fitted probabilities numerically 0 or 1 occurred."

   (b) Separation occurs because a straight line can be drawn in the bottom versus diagonal plot that completely separates the banknotes into genuine and counterfeit. So a linear function of bottom and diagonal perfectly classifies the banknotes.