Homework 2

Denis Ostroushko

Problem 1

A

Interpretation

The linear probability model with one predictor is given by a general equation $\hat{\pi}(x) = \hat{\beta}_0 + \hat{\beta}_1 * x$. Due to its simple form, the interpretation is straightforward. When x = 0, i.e. a person consumes no alcohol, estimated probability of malformation is the intercept, which is 0.0025.

A coefficient for alcohol consumption is 0.001087, meaning that each additional alcoholic drink per day increases the probability of malformation by 0.001087. There are no other predictors in the model, so we do not have anything to hold constant in this interpretation.

Relative Risk

We can use software output and obtain a fitted model: $P(Y = 1|x) = \hat{\pi}(x) = 0.0025 + 0.001087 * x$

Now we can use this model to estimate the probability of having an event when x = 0 and x = 7.

$\hat{\pi}(0) = 0.0025$

$\hat{\pi}(7) = 0.0025 + 0.001087 * 7 = 0.0101$

Therefore, the relative risk is given by a ratio of fitted probabilities, which is $\frac{\hat{\pi}(7)}{\hat{\pi}(0)} = 4.0436$

Problem 2

We define a probability of a even happening for each observation i to be a random quantity $\pi_i = P(Y = 1)$.

A GLM with a log link means that we model the natural parameter $\eta_i = log(\pi_i)$ in terms of a linear combination of predictors.

Therefore, a GLM equation is given as

$$log(\pi_i) = \hat{\beta}_0 + \hat{\beta}_1 * x_1 + \cdots + \hat{\beta}_p * x_p$$

Consider the case of varying just one variable x_1 by 1 unit, which can either represent the case of switching from one categorical level to the next, or increasing a continuous predictor by 1 unit.

Changing x_1 will change the probability from π_1 to π_2, and the difference of two probabilities on the logarithmic scale is given by

$$log(\pi_2) - log(\pi_1) = \hat{\beta}_0 + \hat{\beta}_1 * (x_1 + 1) + \cdots + \hat{\beta}_p * x_p - \hat{\beta}_0 - \hat{\beta}_1 * x_1 - \cdots - \hat{\beta}_p * x_p =>$$

$$\hat{\beta}_1 = log(\frac{\pi_2}{\pi_1})$$

Therefore, $\frac{\pi_2}{\pi_1} = e^{\hat{\beta}_1}$. Taking the ratio instead of a difference of probabilities results in the relative comparison, therefore we evaluate relative risk here.

We do not use this link function often because of the form that $\hat{\pi}(x)$ takes on. $\hat{\pi}(x) = e^{\hat{\beta}_0 + \hat{\beta}_1 * (x_1 + 1) + \cdots + \hat{\beta}_p * x_p}$ is a function that will always be greater than 0 because of the properties of exponential function, but it is not limited by 1 on the upped end. So, given the data, we can have a scenario where fitted probabilities are greater than 1, which violates axioms of probability.

Problem 3

A

We estimate a general linear logistic regression model using a logit link function. So, taking estimates from the table, we know that the software fitted a model that takes this form:

$$log(\frac{\pi(x)}{1-\pi(x)}) = -3.7771 + 0.1449 * x$$

Using logit function, we can calculate the probability of remission when LI = 8:

$$\pi(LI = 8) = \frac{e^{-3.7771+0.1449*8}}{1 + e^{-3.7771+0.1449*8}} =>$$

$\hat{\pi} = 0.068$

B

In this problem we will fix $\hat{\pi}$ at 0.5 and solve for LI.

$$log(\frac{0.5}{1-0.5)}) = -3.7771 + 0.1449 * x => \frac{log(\frac{0.5}{(1-0.5)}) + 3.7771}{0.1449} = x =>$$

$x = 26.0669 \approx 26$

C

The rate of change in π in the case with one predictor is approximated by $\hat{\beta} * \hat{\pi}(x) * (1 - \hat{\pi}(x))$.

We take $\hat{\beta} = 0.1449$,while $\hat{\pi}(LI = 8) = 0.068$, from part (a). So, the rate of change is 0.1449 * 0.068 * (0.932) = 0.009

Similarly, the rate of change at LI = 26 is 0.1449 * 0.5 * 0.5 = 0.036

D

Using methods from parts (a), (b), (c) we estimate the probability of remission at LI = 14 = $\hat{\pi}(14) = P(Y = 1|LI = 14) = 0.15$.

Probability of remission at LI = 28 is $\hat{\pi}(28) = 0.57$.

Thus, probability increases by 0.42 when LI increases from 14 to 28.

E

Odds ratio for a logistic regression model is given by $e^{\hat{\beta}_1}$ for a predictor x_1. This is the multiplicative change in odds ratio.

In our problem, $\hat{\beta}_1 = 0.1449$, and so the odds ratio is $e^{0.1449} = 1.16$

F

Odds ratio is a function of the model parameter $\hat{\beta}_1$. This parameter is an MLE estimate, so by the invariance property, odds ratio is also an MLE. We know that MLE's are asymptotically normally distributed.

Therefore, we need to do the following steps to a confidence interval for odds ratio.

1. Get a 95% confidence interval for $\hat{\beta}_1$ using 1.96 - 97.5th quantile of the the standard normal distribution and a standard error, which we take from the model output. This is a Wald confidence interval.
2. we exponentiate the lower limit of a 95% confidence interval, an odds ratio, and an upper limit.

Recall that $\hat{\beta}_1 = 0.1449$, and the standard error is 0.0593. Therefore, the 95% confidence interval is (0.029, 0.261).

Taking an exponential of all three quantities gives us quantities that we are looking for. Odds ratio is 1.16 with a (1.03, 1.3) 95% confidence interval.

Note that the odds ratio of 1 implies no effect of a predictor on the estimated relapse probability. Obtained confidence interval does not contain a 1, all values are above 1, therefore we can conclude that increase in LI levels is strongly associated with the chance of relapse. One unit increase in LI multiplies the odds of relapse by 1.16.

Given a different set of observations, fitting model with the same predictor will produce a different $\hat{\beta}_1$. We hope that the true value of β_1 is captured by this confidence interval 95% of the time.

G

In the logistic regression framework, Wald test tells us if the estimate is statistically different from 0

1. Null hypothesis: $H_0 : \hat{\beta} = 0$
2. Alternative hypothesis: $H_a : \hat{\beta} \neq 0$

3. Wald test statistic: $W^* = (\frac{\hat{\beta}-0}{se(\hat{\beta})})^2 = \frac{0.1449}{0.0593} = 5.971$ with 1 degree of freedom
4. Cutoff value is the 95th quantile of $\chi^2_1 = 3.84 = C$
5. $P(C > W^*) = 0.01455$
6. P value is small and the test statistic is greater than the cutoff value for significance at the 95% confidence level. Therefore, we have enough evidence to reject the null hypothesis and conclude that the effect of LI level is not zero. Higher LI levels are positively associated with the chance of relapse.

H

We can conduct a likelihood ratio test for the effect when we compare a model with 1 additional parameter against a model with just the intercept.

1. Null hypothesis: $H_0 : \hat{\beta} = 0$
2. Alternative hypothesis: $H_a : \hat{\beta} \neq 0$
3. Null deviance: 34.372, Residual deviance: 26.073, Test statistic is $X^2 = 34.372 - 26.073 = 8.299$
4. Degree of freedom = 1 due to one parameter subject to test
5. Cutoff for significance is the 95th percentile of a chi-square distribution with 1 degree of freedom: 3.8415
6. $P(\chi^2_1 > X^2) = 0.00397$.
7. We have enough statistical evidence to reject the null hypothesis and conclude that the estimate is different from zero. The drop in deviance is large enough to conclude that the addition of LR levels as a predictor is necessary to improve model fit.

I

I decided to adopt an approach from page 30 of lecture notes for part one. I will first find a standard error and a confidence interval for the linear predictor, $\eta = \mathbf{x}^T\hat{\beta}$. Then we will use a logit function to map linear predictor and a 95% confidence interval to the probability space.

Variance of a linear predictor

Software output shows that the intercept and LI parameter are correlated, therefore, variance of $\hat{\beta}_0 + \hat{\beta}_{LI} * LI$ is given by

$$Var(\hat{\beta}_0 + \hat{\beta}_{LI} * LI) = Var(\hat{\beta}_0) + LI^2 * Var(\hat{\beta}_{LI}) + 2 * LI * Cov(\hat{\beta}_0, \hat{\beta}_{LI})$$

A variance-covariance matrix shows that $Var(\hat{\beta}_{LI})$ = 0.003521, $Var(\hat{\beta}_0)$ = 1.900616, and $Cov(\hat{\beta}_0, \hat{\beta}_{LI})$ = -0.07653.

I will use functions in R to assist with calculations. Function below produces a linear predictor given a set of model parameters:

```
# logit function for mapping to probability space
logit <-
  function(x){
    exp(x)/(1+exp(x))
  }

# estimate linear combination of predictors
lin_predictor <-
  function(beta_0, beta_1, x){
    beta_0 + beta_1 * x
  }
```

We can confirm that when LI = 8, linear predictor $\hat{\eta}$ = -2.6179, and fitted probability is $logit(\hat{\eta})$ = 0.068, which matches our previous results.

Now we can produce variance and standard error for the linear combination of predictors

```
# compute variance
var_lin_predictor <-
  function(beta_0, beta_0_var,
           beta_1, beta_1_var,
           cov,
           x){
    return(
      beta_0_var + x^2 * beta_1_var + 2 * x * cov
    )
  }
```

We estimate that variance of $\hat{\eta}$ = is 0.90148

So, when LI = 8, $\hat{\eta}$ = -2.6179 with a (-4.4788, -0.757) 95% confidence interval.

Confidence interval of a fitted probability

Now, using a logit function, we can map these three estimates to the probability space.

Estimated probability of relapse at LI = 8 is 0.068, bounded by the (0.0112, 0.3193) 95% confidence interval.

Problem 4

We are given a logit equation: $logit(\hat{\pi}) = -10.071 - 0.509 * c + 0458 * x$

Additionally, we have descriptive sample statistics for explanatory variables c and x.

Standardized Coefficients

Denote standardized coefficient as $\hat{\beta}'$, and $\hat{\beta}' = \hat{\beta} * s$ where s is a sample standard deviation of a given predictor. We interpret standardized coefficient as an executed change in log-odds when a given predictor x increases by one standard deviation, after adjusting for other predictors.

This is a simple change from a regular $\hat{\beta}$ which tells us an expected change in log-odds when x changes by one unit, after adjusting for other variables.

In the case of width, a standardized coefficient is $\hat{\beta}'_x = 0.458$ * $2.11 = 0.97$.

Therefore, when width of a crab changes by one standard deviation we expect log -odds of having a satellite to increase by 0.97. For example, this change in width may increase log-odds from 1 to 1.97.

Color has a similar, although an unintuitive explanation. The average color category is 2.44 and the standard deviation is 0.8. Therefore, a one standard deviation from the mean is 2.44 + 0.8 = 3.24, which corresponds to the -0.41 * 0.8 = -0.328 change in log odds. So, as crab color category increases, the log odds of having a satellite crab decrease linearly.

Effect of Crab Color Through Probabilities

Using a logit function, we can obtain probabilities for each combination of predictors. We estimate expected change in probability as a result of color change while holding crab width constant at $\bar{x} = 26.3$

Therefore, probability of having a satellite for a crab of color category 1 and average width is

$P(Y = 1|c = 1, x = 26.3) = \frac{e^{-10.071-0.509*1+0.458*26.3}}{1+e^{-10.071-0.509*1+0.458*26.3}} = 0.8124$

Similarly, for a crab of color category 4 we have

$P(Y = 1|c = 1, x = 26.3) = \frac{e^{-10.071-0.509*1+0.458*26.3}}{1+e^{-10.071-0.509*1+0.458*26.3}} = 0.4846$

So, when the crab color category increases from 1 to 4, probability of having a satellite crab decreases by about 40%.

Problem 5

A

We know that odds ratio $= e^{\hat{\beta}}$, therefore we need an inverse of this function to get $\hat{\beta}$, i.e. $\hat{\beta} = log(Odds\ Ratio)$.

I will use code below to assist me in calculations:

```
# some difference
get_std_err_from_ub <-
  function(or, or_ub){

    return(
      (log(or_ub) - log(or))/1.96
    )

  }

get_std_err_from_lb <-
  function(or, or_lb){

    return(
      (log(or) - log(or_lb))/1.96
    )

  }
```

Some Education vs No Education Predictor

1. Odds Ratio is 4.04, so $\hat{\beta}_{Education} = log(4.04) = 1.3962$
2. To get a standard error, we will take the difference between log(OR) and the lower/upper bound of C.I., and divide that difference by 1.96. We do that for both lower and upper bounds to make sure that we get the same value from each calculation, which will be a standard error for the model estimate
3. Thus, scaled difference between log(OR) and log(Lower Bound of C.I) is $\frac{log(4.04) - log(1.17)}{1.96} = 0.6323$
4. Scaled difference between log(Upper Bound of C.I) and log(OR) is $\frac{log(13.9) - log(4.04)}{1.96} = 0.6304$
5. We have a small rounding error, but the two calculations agree, therefore $\hat{\beta}_{Education} = 1.3962$ with $se(\beta_{Education}) = 0.6304$

Gender

1. Following the same steps, we calculate standard error from the upper bound of C.I. first. Standard error from upper bound: $\frac{log(12.88)-log(1.23)}{1.96} = 1.1396$
2. Standard error from lower bound: $\frac{log(1.38)-log(1.23)}{1.96} = 0.0587$
3. The two results do not agree, which we will address in part (b)

SES High vs Low

1. Standard error from upper bound: $\frac{log(18.28)-log(5.82)}{1.96} = 0.5839$
2. Standard error from lower bound: $\frac{log(5.82)-log(1.87)}{1.96} = 0.5793$
3. Disregarding a small rounding error, we have two estimates that agree with each other, so we have found a standard error. Model estimate is $\hat{\beta}_{SESHigh} = 1.7613$ with $se(\hat{\beta}_{SESHigh}) = 0.5793$

Number of Partners

1. Standard error from upper bound: $\frac{log(11.31)-log(3.22)}{1.96} = 0.641$
2. Standard error from lower bound: $\frac{log(3.22)-log(1.08)}{1.96} = 0.5574$
3. A rounding error is quite obvious here, and I am not sure how this happens. In any case, we have two estimates that *almost* agree with each other, so we have found a standard error. Model estimate is $\hat{\beta}_{SES\ High} = 1.1694$ with $se(\hat{\beta}_{SES\ High}) = 0.5574$

B

As our problem suggests, take 1.38 to be a log-odds ratio, so we need to exponentiate and then take the log again to get the right estimate.

We adjust our formula, and calculate standard error using the upper bound as $\frac{log(12.88)-log(exp(1.38))}{1.96} = \frac{log(12.88)-1.38)}{1.96}$

So, calculating from the upper bound, we have 0.5998

So, calculating from the lower bound, we have 0.5985

Since the two calculation agree, we conclude that the model estimate for Gender is $\hat{\beta}_{Males} = 1.38$ with $se(\hat{\beta}_{Males}) = 0.5998$

Problem 6

We are given a hypothetical data set with count data at different levels of x. When $x = 0$ then we have y_0 success outcomes out of n_0 trials. Therefore, we have $n_0 - y_0$ fails.

The odds of success for the group with $x = 0$ is then given by

$$\frac{\frac{y_0}{n_0}}{\frac{n_0 - y_0}{n_0}} = \frac{y_0}{n_0 - y_0}$$

Using similar arguments, odds of success for group at level $x = 1$ are given by

$$\frac{\frac{y_1}{n_1}}{\frac{n_1 - y_1}{n_1}} = \frac{y_1}{n_1 - y_1}$$

And the Odds Ratio for $x = 1$ to $x = 0$ comparison is given using an odds ratio

$$\frac{\frac{y_1}{n_1 - y_1}}{\frac{y_0}{n_0 - y_0}}$$

Then the Log odds ratio is

$$log(\frac{\frac{y_1}{n_1 - y_1}}{\frac{y_0}{n_0 - y_0}}) = log(\frac{y_1}{n_1 - y_1}) - log(\frac{y_0}{n_0 - y_0})$$

In the framework of a logistic regression model $logit(\hat{\pi}(x)) = \alpha + \beta * x$ means that when $x = 0$, $logit(\hat{\pi}(0)) = \hat{\alpha}$. Similarly, when $x = 1$ we have $logit(\hat{\pi}(1)) = \hat{\alpha} + \hat{\beta}$.

Therefore, we need to show that $\hat{\beta} = log(\frac{\frac{y_1}{n_1 - y_1}}{\frac{y_0}{n_0 - y_0}})$

Intuitively, in the case of this simple model, a log-odds ratio is the linear increase in the log-odds of success when we increase from level $x = 0$ to $x = 1$.

Now we need to set up log-likelihood equations for the logistic regression that we have an show that $\alpha + \beta = log(\frac{y_1}{n_1 - y_1})$

As stated on page 192 of CDA textbook, "When more than one observation occurs at a fixed x_i value, it is sufficient to record the number of observations n_i and the number of successes. We then let y_i refer to this success count rather than to an individual binary response."

So, we state the general likelihood function for such a set up, and then specify it further for our case with two groups and two model parameters.

In general, when we bucket the data as described above, the likelihood function is given by:

$$\Pi_{i=1}^{N}(\pi(x_i)^{y_i} * [1 - \pi(x_i)^{1-y_i}])$$

where:

- y_i is the number of successes, we have levels 0 and 1
- $\pi(x_i)$ is the probability of success at levels 0 and 1

We can now specify likelihood function for our case:

$$L = \Pi_{i=0}^{1}(\pi(x_i)^{y_i} * [1 - \pi(x_i)^{1-y_i}]) = \pi(x_0)^{y_0} * [1 - \pi(x_0)]^{n_0-y_0} * \pi(x_1)^{y_1} * [1 - \pi(x_1)]^{n_1-y_1}$$

Next obvious step is to take the log of this equation and obtain the log-likelihood function:

$$log\ L = y_0 * log(\pi(x_0) + (n_0 - y_0) * log[1 - \pi(x_0)] + y_1 * log(\pi(x_1) + (n_1 - y_1) * log[1 - \pi(x_1)]$$

We know that $logit(\pi) = \alpha + \beta$ so we we can plug in an expression for probability into the log likelihood function:

$$log\ L = y_0 * log(\frac{exp(\alpha)}{1 + exp(\alpha)}) + (n_0 - y_0) * log[\frac{1}{1 + exp(\alpha)}] + y_1 * log(\frac{exp(\alpha + \beta)}{1 + exp(\alpha + \beta)} + (n_1 - y_1) * log[\frac{1}{1 + exp(\alpha + \beta)}]$$

We redistribute the terms and simplify the function to obtain:

$$log\ L = y_0 * \alpha + y_1 * (\alpha + \beta) - n_0 * log(1 + e^{\alpha}) - n_1 * log(1 + e^{\alpha+\beta})$$

Take the first derivative with respect to β, set equal to zero to get an MLE estiamte of β

$$\frac{\partial\ log\ L}{\partial \beta} = y_1 - n_1 * \frac{e^{\alpha+\beta}}{1 + e^{\alpha+\beta}} = 0$$

Which yields

$$y_1 = (n_1 - y_1) * (e^{\alpha+\beta})$$

Take the log and isolate α and β

$$log(y_1) - log(n_1 - y_1) = \alpha + \beta$$

Therefore,

$$log(\frac{y_1}{n_1 - y_1}) = \alpha + \beta$$

Now, we need to show that β is the log odds ratio of the sample, which is a linear increase in log odds when switching from group where $x = 0$ to group where $x = 1$.

If

$$log(\frac{y_1}{n_1 - y_1}) = \alpha + \beta$$

and $\beta = log(\frac{\frac{y_1}{n_1 - y_1}}{\frac{y_0}{n_0 - y_0}})$, then solving for alpha must yield a log odds for a group where $x = 0$.

$$log(\frac{y_1}{n_1 - y_1}) = \alpha + \beta,$$

$$log(\frac{y_1}{n_1 - y_1}) = \alpha + log(\frac{\frac{y_1}{n_1 - y_1}}{\frac{y_0}{n_0 - y_0}}),$$

$$log(\frac{y_1}{n_1 - y_1}) = \alpha + log(\frac{y_1}{n_1 - y_1}) - log(\frac{y_0}{n_0 - y_0}),$$

$$log(\frac{y_0}{n_0 - y_0}) + log(\frac{y_1}{n_1 - y_1}) = \alpha + log(\frac{y_1}{n_1 - y_1})$$

Therefore, $\alpha = log(\frac{y_0}{n_0 - y_0})$ and equation holds.

We showed that β is the log-odds ratio of the sample.