Computer Assignment 2

Gabriel Lindqvist, Jiahui Li

2024-11-24

Exercise2:1

Data on periodontitis (severe infammation of the gums) were collected from a group of adult patients at a large dental clinic.The patients were compared with a group of adult people with normal gums who visited the same clinic.The individuals in both groups were interviewed about the use of dental floss in the last five years,with the following result:

```
##                              Periodontitis
## Regularly use of dental floss Yes  No Sum
##                           Yes  22  75  97
##                           No  148 265 413
##                           Sum 170 340 510
```

```
##                              Periodontitis
## Regularly use of dental floss       Yes        No       Sum
##                           Yes 0.2268041 0.7731959 1.0000000
##                           No  0.3583535 0.6416465 1.0000000
```

Task 1

The problem asks us to analyze the relationship between regular dental floss use and the occurrence of periodontitis (gum disease) using a logistic regression model. The data provided include counts of individuals categorized by whether they regularly use dental floss (Yes/No) and whether they have periodontitis (Yes/No). The logistic regression model is specified as:

$$\text{logit}(p_x) = \beta_0 + \beta_1 x,$$

where

- $x = 1$ if the individual regularly uses dental floss, $x = 0$ otherwise.
- $p_x = P(\text{periodontitis} \mid x$, the probability of having periodontitis given the flossing status.

The task is to answer the following questions:

1. Which parameters in the model can be estimated considering how the data is collected?
2. What estimates do you get and how do you interpret them?

To answer the first question, we start by analyzing the data found in the contingency table above. Firstly, all cell counts in the contingency table seems sufficient. That is, all cell counts are non-zero and are fairly large, indicating that there should be no issues in estimating β_0 and β_1. However, the group of individuals who regularly use dental floss is relatively under-sampled compared to the individuals who do not use dental floss (97 vs 413). This impacts the precision of the estimates for the effect of floss and could limit the statistical power to detect an association between floss use and periodontitis.

In order to answer the second question, we fit a logistic regression and study the estimates.

```
##
## Call:
## glm(formula = per ~ use, family = binomial(link = logit), data = data21,
##     weights = n)
##
## Deviance Residuals:
##       1        2        3        4
## -15.335   17.429   -6.212    8.080
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5825     0.1026  -5.677 1.37e-08 ***
## useyes       -0.6439     0.2633  -2.446   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 649.24  on 3  degrees of freedom
## Residual deviance: 642.80  on 2  degrees of freedom
## AIC: 646.8
##
## Number of Fisher Scoring iterations: 5
```

From the summary we get the following estimates:

- Intercept: $\beta_0 = -0.5825$
- Slope: $\beta_1 = -0.6439$

Substituting these into the model:

$$\text{logit}(p_x) = -0.5825 - 0.6439 \cdot x$$

Next, we want to interpret the parameters.

- **Intercept ($\beta_0 = -0.5825$)**:
  When $x = 0$ (no regular floss use), the log-odds of having periodontitis is -0.5825. Converting this to a probability:
  $$P(\text{Periodontitis} \mid \text{use} = \text{no}) = \frac{\exp(-0.5825)}{1 + \exp(-0.5825)} \approx 0.358$$
  Interpretation: For individuals who do not floss regularly, the probability of having periodontitis is approximately 35.8%.

- **Slope ($\beta_1 = -0.6439$)**:
  When $x = 1$ (regular floss use), the log-odds of periodontists decreases by 0.6439 for individuals who regularly floss compared to those who do not. Converting this to an odds ratio (OR):

$$\text{OR} = \exp(-0.6439) \approx 0.525$$

  Interpretation: Regular flossing reduces the odds of periodontitis approximately 47.5%.

Statistical Significance of the Parameters

- **Intercept (β_0)**: The p-value (< 0.001) shows that the baseline odds of periodontitis for individuals who do not floss regularly is statistically significant.
- **Slope (β_1)**: The p-value (0.0145) indicates that the effect of flossing on the odds of periodontitis is statistically significant at the 5% level.

Summary of Results

- **Regular floss use significantly reduces the likelihood of periodontitis**: The negative slope (β_1) and odds ratio (0.525) suggest that regular flossing is associated with a significant reduction in the odds of periodontitis by approximately 47.5%.
- **Baseline risk for non-floss users**: Individuals who do not floss regularly have a periodontitis probability of 35.8%.

In conclusion, regular flossing has a statistically significant protective effect against periodontitis, as shown by the model ($p = 0.0145$).

Task 2

The task is to explore the same questions found in **Task 1** but now we fit a logistic regression model for the probability of **using dental floss** as a function of periodontitis status. The logistic regression model is specified as:

$$\text{logit}(p_y) = \gamma_0 + \gamma_1 y,$$

where

- $y = 1$ if the individual has periodontitis, and $y = 0$ otherwise.
- $p_y = P(\text{using dental floss} \mid y)$, the probability of using dental floss given the periodontitis status.

Regarding the first question, γ_0 and γ_1 can both be estimated in this study because the contingency provides sufficient information for the model to estimate these parameters (as established in the previous task).

Next, we fit the logistic regression model to get our estimates.

```
##
## Call:
## glm(formula = use ~ per, family = binomial(link = logit), data = data21,
##     weights = n)
##
## Deviance Residuals:
```

```
##       1        2        3        4
## -11.493   -6.405   15.057    9.485
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2622     0.1308  -9.651   <2e-16 ***
## peryes       -0.6439     0.2633  -2.446   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 496.24  on 3  degrees of freedom
## Residual deviance: 489.79  on 2  degrees of freedom
## AIC: 493.79
##
## Number of Fisher Scoring iterations: 5
```

From the summary we get the following estimates:

- Intercept: $\gamma_0 = -1.2622$
- Slope: $\gamma_1 = -0.6439$ (effect of `peryes` relative to `perno`)

Substituting these into the model:

$$\text{logit}(p_x) = -1.2622 - 0.6439 \cdot x.$$

Next, we want to interpret these estimates.

- **Intercept ($\gamma_0 = -1.2622$)**:
  When $y = 0$ (no periodontitis), the log-odds of regular floss use is -1.2622. Converting this to a probability:
  $$P(\text{using floss} \mid y = 0) = \frac{\exp(-1.2622)}{1 + \exp(-1.2622)} \approx 0.22.$$
  Interpretation: Individuals without periodontitis have approximately a 22% probability of using dental floss.

- **Slope ($\gamma_1 = -0.6439$)** When $y = 1$ (periodontitis is present), the log-odds of using dental floss decreases by 0.6439. Converting this to an odds ratio (OR):
  $$\text{OR} = \exp(-0.6439) \approx 0.525$$
  Interpretation: Individuals with periodontitis are 52.5% as likely to use dental floss compared to those without periodontitis. In other words, the presence of periodontitis is associated with a lower likelihood of using dental floss.

Statistical Significance of the Parameters

- **Intercept (γ_0)**:
  - $p < 0.001$: The intercept is significant, indicating that the baseline odds of using dental floss for individuals without periodontitis is significantly below 50%.
- **Slope (γ_1)**:
  - $p = 0.0145$: The slope is significant, meaning periodontitis status has a statistically significant effect on the probability of using dental floss.

Summary of Results **Individuals without periodontitis**:

- The probability of using dental floss is approximately **22%**.

Individuals with periodontitis:

- They are **52.5% as likely** to use dental floss compared to those without periodontitis.

The effect of periodontitis on floss use is statistically significant:

- $p = 0.0145$: Indicates a significant relationship between periodontitis status and floss usage behavior.

Unexpected finding:

- Contrary to expectations, individuals with periodontitis are less likely to use dental floss. This contradicts common assumptions (e.g., individuals with gum disease might be more likely to adopt dental care habits) and warrants further investigation into behavioral patterns or data collection factors.

Task 3

In this task we want to explore the following:

1. Show how the parameter estimates (β_1 and γ_1) from both models (from Questions 1 and 2) can be derived using the actual numbers in the contingency table.
2. Prove that $\beta_1 = \gamma_1$.

Step 1: Recap of the Models

Model 1 (Question 1):

$$\text{logit}(p_x) = \beta_0 + \beta_1 x,$$

where

- $x = 1$ (regular dental floss use), $x = 0$ (no dental floss use).
- $p_x = P(\text{Periodontitis} \mid x)$.

Model 2 (Question 2):

$$\text{logit}(p_y) = \gamma_0 + \gamma_1 y,$$

where

- $y = 1$ (Periodontitis present), $y = 0$ (No periodontitis).
- $p_y = P(\text{Using Floss} \mid y)$.

Step 2: Table Data

The contingency table provides the following counts:

Regularly Use of Dental Floss	Periodontitis (Yes)	Periodontitis (No)	Row Totals
Yes	22	75	97
No	148	265	413
Column Totals	170	340	510

Step 3: Express β_1 and γ_1 Using the Table

For β_1 (Model 1): The logistic regression for Model 1 calculates the odds ratio of periodontitis ($P(\text{Periodontitis} \mid \text{Floss})$) for floss users versus non-users. From the table:

- Odds of periodontitis for floss users:

$$\text{Odds}_{\text{Yes}} = \frac{22}{75}$$

- Odds of periodontitis for non-floss users:

$$\text{Odds}_{\text{No}} = \frac{148}{265}$$

- Odds ratio (OR):

$$\text{OR} = \frac{\text{Odds}_{\text{Yes}}}{\text{Odds}_{\text{No}}} = \frac{\frac{22}{75}}{\frac{148}{265}} = \frac{22 \cdot 265}{75 \cdot 148}$$

- Log odds ratio (β_1):

$$\beta_1 = \log(\text{OR}) = \log\left(\frac{22 \cdot 265}{75 \cdot 148}\right)$$

For γ_1 (Model 2): The logistic regression for Model 2 calculates the odds ratio of floss use ($P(\text{Floss} \mid \text{Periodontitis})$) for individuals with periodontitis versus those without. From the table:

- Odds of floss use for individuals with periodontitis:

$$\text{Odds}_{\text{Yes}} = \frac{22}{148}$$

- Odds of floss use for individuals without periodontitis:

$$\text{Odds}_{\text{No}} = \frac{75}{265}$$

- Odds ratio (OR):

$$\text{OR} = \frac{\text{Odds}_{\text{Yes}}}{\text{Odds}_{\text{No}}} = \frac{\frac{22}{148}}{\frac{75}{265}} = \frac{22 \cdot 265}{148 \cdot 75}$$

- Log odds ratio (γ_1):

$$\gamma_1 = \log(\text{OR}) = \log\left(\frac{22 \cdot 265}{75 \cdot 148}\right)$$

Step 4: Prove $\beta_1 = \gamma_1$

By the definitions above, both β_1 and γ_1 are calculated as:

$$\log\left(\frac{22 \cdot 265}{75 \cdot 148}\right)$$

Thus, we have:

$$\beta_1 = \gamma_1$$

This equality occurs because the odds ratio for Model 1 (Periodontitis as a function of floss use) and Model 2 (Floss use as a function of periodontitis) share the same underlying data structure and proportion relationships in the contingency table. The log odds ratio remains invariant regardless of which variable is treated as the dependent variable.

Step 5: Conclusion

The parameter estimates (β_1 from Model 1 and γ_1 from Model 2) are equal because they are both derived from the same odds ratio in the contingency table. This demonstrates the symmetry of logistic regression in cases like this, where the relationship between two binary variables is analyzed.

Exercise 2:2

In an experiment, 165 mice were exposed to various doses of bensoapyren (BaP) over a ten month period.It was then examined how many of the mice that has developed a lung tumor.The result as a function of logarithmic dose is:

```
##   logdos  x  n
## 1  -7.60  1 18
## 2  -6.22  2 19
## 3  -4.60  4 28
## 4  -3.00  9 32
## 5  -1.39 12 28
## 6   0.92 32 40
```

Task 1

Estimate for each dose separately the risk (probability) to develop a lung tumor. Make a plot of the risk against log(dos). Calculate the log odds for tumor and plot it against log(dose).Seems the logistic regression model appropriate for these data?

The risk (or probability) of developing a lung tumor is calculated as:

$$\text{Risk} = \frac{\text{Tumor (x)}}{\text{Total mice (n)}}$$

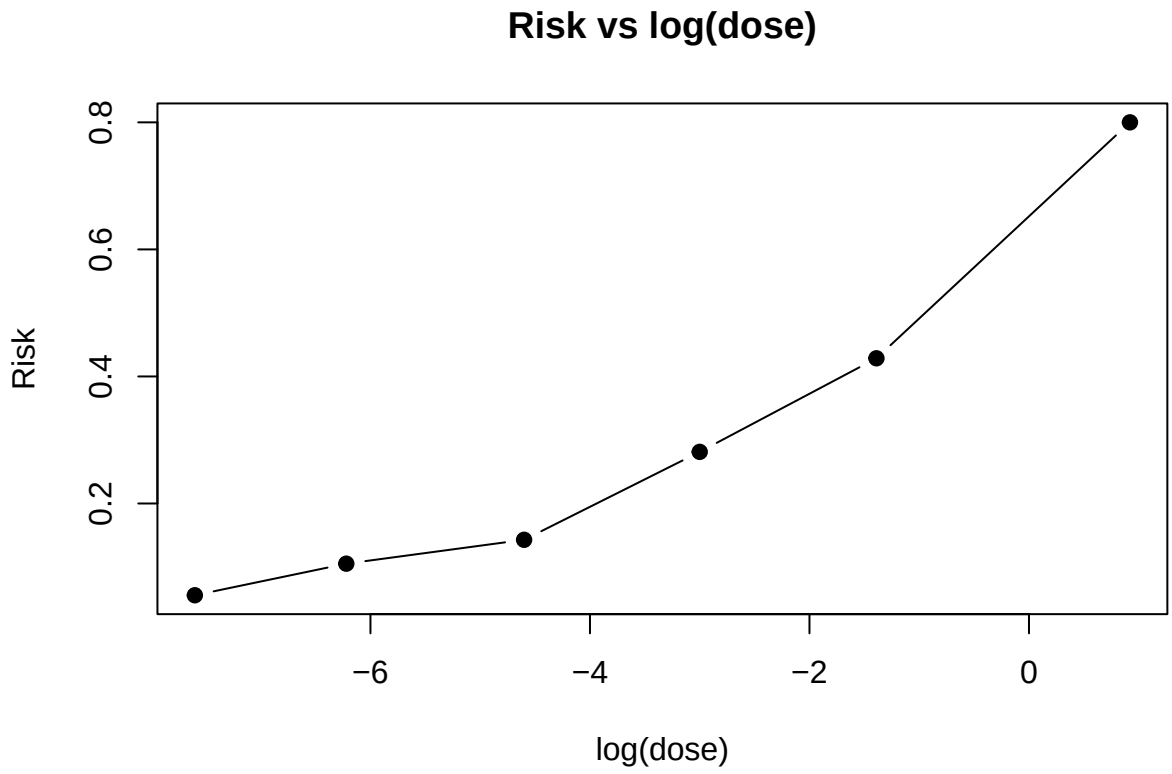
The log-odds (logits) are calculated as:

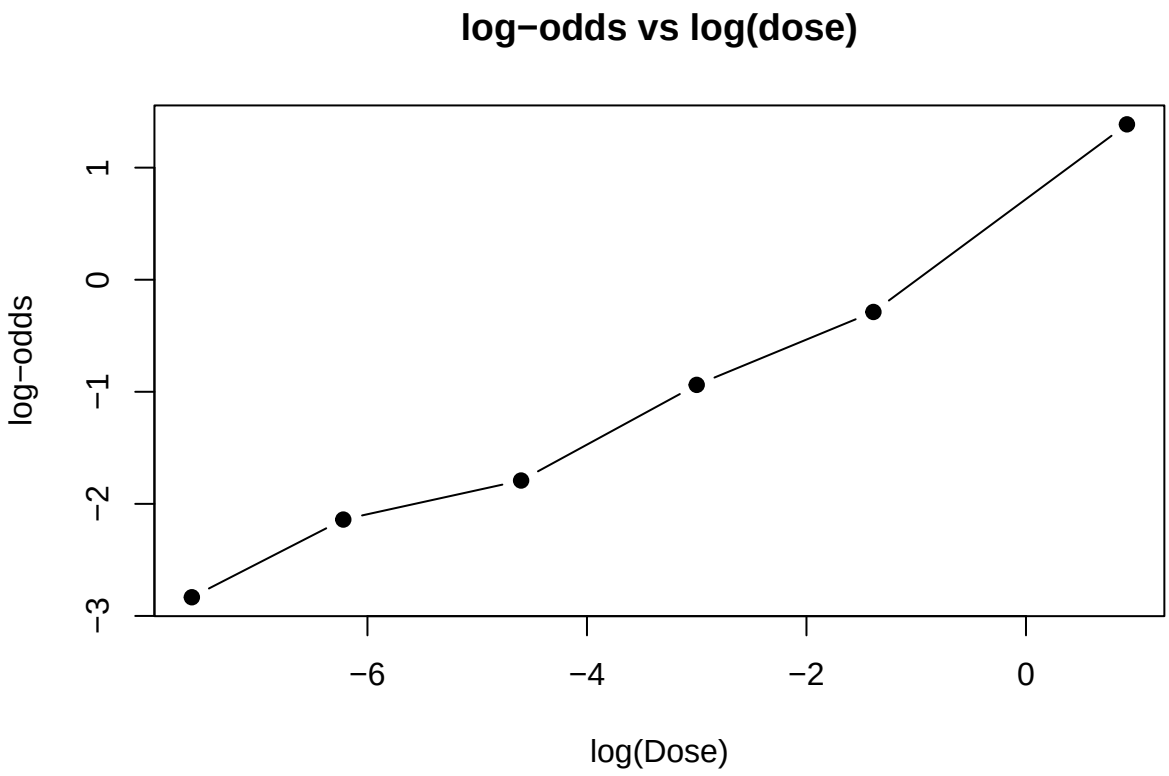
$$\text{Log-Odds} = \log\left(\frac{\text{Risk}}{1 - \text{Risk}}\right)$$

We start by calculating the risk and log-odds and add them as two seperate columns into our dataframe.

```
##   logdos  x  n       risk   log_odds
## 1  -7.60  1 18 0.05555556 -2.8332133
## 2  -6.22  2 19 0.10526316 -2.1400662
## 3  -4.60  4 28 0.14285714 -1.7917595
## 4  -3.00  9 32 0.28125000 -0.9382696
## 5  -1.39 12 28 0.42857143 -0.2876821
## 6   0.92 32 40 0.80000000  1.3862944
```

Next, we plot the calculated risk and log-odds against log(dose) separately





Seems the logistic regression model appropriate for these data?

Data Characteristics:
Logistic regression is suitable for binary data. In this case, the response variable is the probability of tumor occurrence, which meets the conditions for applying a logistic regression model. **Linear Trend of Log-Odds**:
From the plotted relationship between Log-Odds and Log(Dose), a linear trend is observed, it further supports the applicability of the logistic regression model.

Task 2

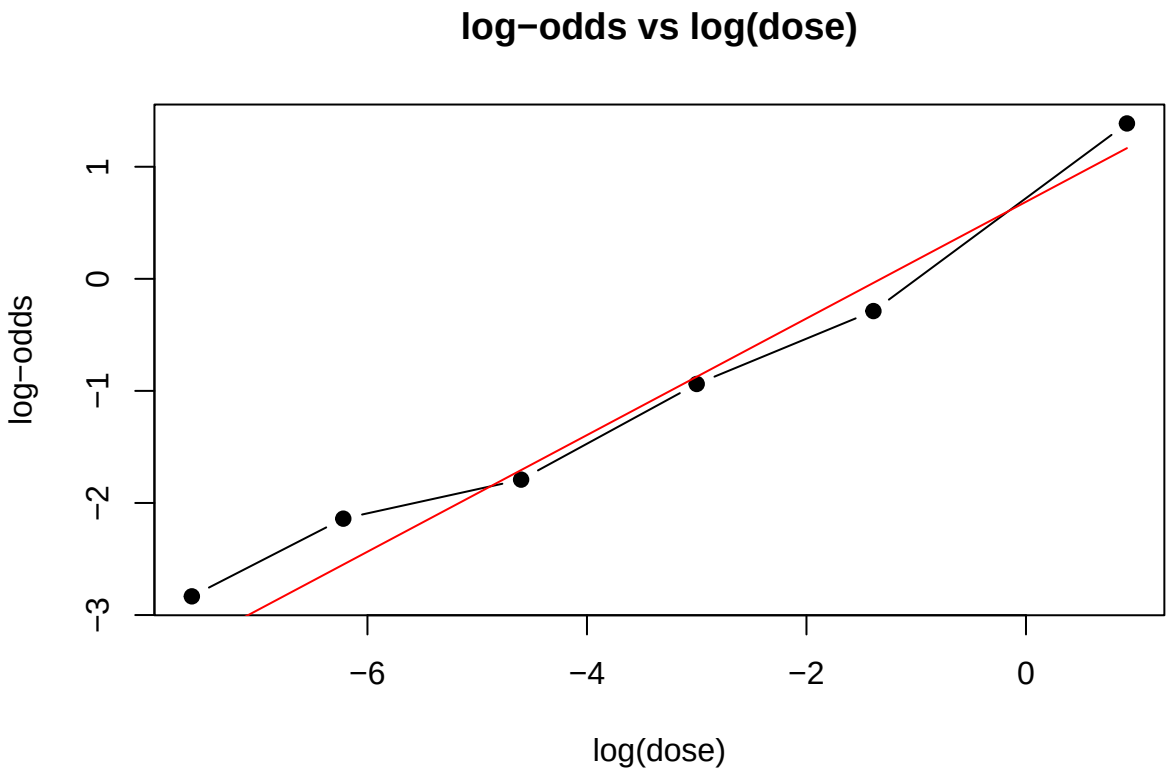
Fit a logistic regression model to this data and interpret the parameter estimates, particularly the slope parameter.

We start by fitting the logistic regression model.

```
##
## Call:
## glm(formula = x/n ~ logdos, family = binomial(link = logit),
##     data = data22, weights = n)
##
## Deviance Residuals:
##       1        2        3        4        5        6
##  0.3965   0.5193  -0.1592  -0.1640  -0.6614   0.5701
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.68702    0.25755   2.668  0.00764 **
## logdos       0.52037    0.08502   6.120 9.35e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 56.5321  on 5  degrees of freedom
## Residual deviance:  1.2416  on 4  degrees of freedom
## AIC: 24.024
##
## Number of Fisher Scoring iterations: 4
```

Next, we visualize by adding fitted values to the previous log-odds vs log(dose) plot.



Graph Observation: log-odds vs log(dose)

- **Linear Trend**: The relationship between log-odds and log(dose) appears approximately linear, and the fitted line aligns closely with the black dots. This indicates that the assumption of linearity between log-odds and log(dose) holds, making the logistic regression model appropriate for the data.

Model Summary Interpretation

Coefficients

- **Intercept ($\beta_0 = 0.687$)**:
  - When log(dose) is 0 (i.e., the dose is 1), the log-odds of developing a tumor is 0.687.
  - Converting to probability:
$$P = \frac{\exp(0.687)}{1 + \exp(0.687)} \approx 0.665$$
    Interpretation: At a dose of 1, the probability of developing a tumor is approximately 66.5%.
- **Slope ($\beta_1 = 0.520$)**:
  - For every one-unit increase in Log(Dose), the log-odds of developing a tumor increases by 0.520.
  - Converting to odds ratio (OR):
$$\text{OR} = \exp(0.520) \approx 1.682$$
    Interpretation: For every one-unit increase in Log(Dose), the odds of developing a tumor increase by approximately 68.2%.

Significance

- The p-value for the intercept is $p = 0.00764$, and for the slope, $p < 0.001$, indicating that both are statistically significant. This shows that Log(Dose) has a significant effect on tumor development.

Model Fit

- **Residual Deviance (1.2416)**:
  - The residual deviance is much smaller than the null deviance (56.5321), indicating that the model fits the data well.
- **Degrees of Freedom**:
  - The residual degrees of freedom (4) are consistent with the complexity of the data.
- **AIC (24.024)**:
  - A low AIC value indicates that the model is both parsimonious and provides a good fit to the data.

Conclusion

1. **Suitability of the Logistic Regression Model**:
   - The relationship between Log-Odds and Log(Dose) is linear.
   - The residual deviance is low, showing a good fit, and all model parameters are statistically significant.
2. **Relationship Between Dose and Tumor Development**:
   - Higher doses (Log(Dose)) significantly increase the probability of tumor development ($\beta_1 > 0$).

Task 3

Find the covariance matrix for the estimates and assess if they are correlated. Find a 95% confidence interval for the parameters. Find a 95 confidence interval for the tumor risk at dose 0.25 (log(dose)=-1.39)).

We start by assessing whether the estimates are correlated

```
##             (Intercept)      logdos
## (Intercept)  0.06633368 0.014358467
## logdos       0.01435847 0.007229089
```

```
## [1] "Pearson's correlation coeff =  0.65569104054274"
```

Covariance and Correlation: - The covariance between `(Intercept)` and `logdos` is `0.0144`, indicating a small linear relationship between the two parameters. This suggests limited correlation between the intercept and the dose effect. However, by computing the correlation coefficient manually, we see from the output that the correlation is roughly 0.66. This contradicts the previous indication of limited correlation, underlying that covariance is not sufficient when assessing linear relationship between variables.

Next, we want to find a 95% confidence interval for the parameters using `confint.default()`.

```
##                 2.5 %    97.5 %
## (Intercept) 0.1822292 1.1918194
## logdos      0.3537212 0.6870093
```

Significance: - From the confidence intervals, the interval for `logdos` does not include 0, confirming that log(dose) has a significant impact on tumor development.

Effect of Dose: - The coefficient for `logdos` lies within $[0.3537, 0.6870]$, indicating that increasing the dose significantly increases the log-odds of tumor occurrence.

Lastly, we want to find a 95% confidence interval for the tumor risk at does 0.25 (equivalent to log(dose) = -1.39) using the formula

$$\text{CI} = \hat{\beta_0} - 1.39\hat{\beta_1} \pm 1.96 \cdot \text{SE},$$

where $\text{SE} = \sqrt{x^t \text{Cov}(\beta_0, \beta_1)x}$, $x^t = (1, -1.39)$ and $\text{Cov}(\beta_0, \beta_1)$ is the covariance matrix. After which, we convert the C.I above to a probability scale using

$$\frac{\exp(\text{CI})}{1 + \exp(\text{CI})}.$$

This is a consequence of the Delta Method.

```
## [1] "Point estimate converted to probability scale: 0.490930147220603"
```

```
## [1] "95% C.I Lower: 0.394087465976418"
```

```
## [1] "95% C.I Upper: 0.588458364428157"
```

From the output above we see that the tumor risk at dose 0.25 is estimated to approximately 0.49% and lies within the 95% confidence interval.

Task 4

Perform a Wald-test of $H_0 : \beta = 0$. Explain and show with an own calculation, based on R-output, how the test statistic is constructed.

The Wald test statistic is calculated as:

$$W = \frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)},$$

where

- $\hat{\beta}_1$: Estimated regression coefficient for log(dose).
- $\text{Var}(\hat{\beta}_1)$: Variance of the estimated coefficient (obtained from the diagonal of the covariance matrix).

```
# Extract the coefficient and standard error for log(dose)
beta1 <- coef(model)["logdos"]  # Extract the coefficient for log(dose)
se_beta1 <- sqrt(vcov(model)["logdos", "logdos"])  # Extract the variance and compute standard error

# Calculate the Wald statistic
wald_stat <- (beta1^2) / (se_beta1^2)

# Display the Wald statistic
cat("Wald Statistic:", wald_stat, "\n")
```

```
## Wald Statistic: 37.457
```

```
# Calculate the p-value (chi-square distribution with 1 degree of freedom)
p_value <- 1 - pchisq(wald_stat, df = 1)

# Display the p-value
cat("p-value:", p_value, "\n")
```

```
## p-value: 9.345077e-10
```

Conclusion: With an extremly small p-value < 0.001, we do reject the null hypothesis that $\beta_1 = 0$. Thus the dose (log-transformed) has a significant positive effect on tumor development probability.

Task 5

Q: The variance of a parameter estimate is generally inversely proportional to the sample size. Is this true also here? Since we have no expression explicitly for the variances, we cannot confirm it straightforward. In order to investigate it,you should multiply all counts in the table by 10 repeatedly (10,100,1000) and report what you observe.

```
##   Scale Factor Variance of Beta1
## 1           10      7.229093e-04
## 2          100      7.229093e-05
## 3         1000      7.229093e-06
```

Observations

- As the scale factor (sample size) increases, the variance of β_1 decreases significantly, confirming the inverse relationship.
- For example:
  - At a scale factor of 10: Variance is relatively large.
  - At a scale factor of 100: Variance is 1/10th of the previous case.
  - At a scale factor of 1000: Variance is further reduced to 1/100th of the original.

Conclusion

This empirical investigation confirms that the variance of a parameter estimate in logistic regression is inversely proportional to the sample size. This is consistent with statistical theory.