Assignment II

Logistic regression

Silpa Soni Nallacheruvu (19980824-5287) Hernan Aldana (20000526-4999)

2024-11-28

Exercise 2:1

This report analyses the data on periodontitis from a group of adult patients at a large dental clinic. The goal is to understand the influence of the parameter estimates in logistic regression models of the probability for periodontitis on the population as a function of dental floss use and the probability for using dental floss as a function of periodontitis status.

Table 1: Regular Use of Dental Floss and Periodontitis

	Periodontitis	No Periodontitis	Total
Used Dental Floss	22	75	148
Not Used Dental Floss	265	97	413
Total	170	340	510

Question 1: Logistic regression model of the probability for periodontitis

The logistic regression model,say Model A of the probability for periodontitis in the population as a function of dental floss use is defined as:

Model A : $\text{logit}(p_x) = \log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x$

where:

- $x = 1$ if dental floss is regularly used, $x = 0$ otherwise.
- $p_x = P(\text{periodontitis} \mid x)$, the probability of periodontitis given x.
- β_0: The log-odds of periodontitis when $x = 0$ (no floss use).
- β_1: The change in log-odds of periodontitis when x changes from 0 to 1 (effect of using floss).

Approach:

Fit the logistic regression model A to the data and interpret the parameter estimates β_0 and β_1.

Estimate Parameters:

Fit the logistic regression model:

The summary of the coefficients of the logistic regression model A is presented below:

Table 2: Summary of Logistic Regression Model A Coefficients

	Term	Estimate	Std. Error	z value	P-value
(Intercept)	(Intercept)	-0.5825176	0.1026175	-5.676593	0.0000000
floss	floss	-0.6439281	0.2632835	-2.445759	0.0144548

```
## [1] "AIC of Model A: 15.0724989699782"
```

The final logistic regression equation: $\text{logit}(p_x) = -0.582 - 0.644x$

Interpret the Estimates:

- 1. β_0 = -0.582:
  - The log-odds of periodontitis for individuals who do not use dental floss is approximately -0.582.
  - The corresponding probability of periodontitis is: $p_0 = \frac{e^{-0.582}}{1+e^{-0.582}} \approx 0.358$

Individuals who do not use dental floss have a probability of approximately 35.8% of developing periodontitis compared to those who do.

- 2. β_1 = -0.644:
  - The log-odds of periodontitis decreases by 0.644 when individuals use dental floss regularly.
  - This corresponds to an odds ratio of: Odds Ratio $= e^{\beta_1} = e^{-0.644} \approx 0.525$

Individuals who use dental floss regularly have approximately 52.5% lower odds of developing periodontitis compared to those who do not.

Conclusion:

- Regular dental floss use is associated with a significant reduction in the odds of periodontitis.
- The logistic regression model A is accurate because the binary data (periodontitis: yes/no) and the categorical explanatory variable (dental floss use) provide estimates that can be used to interpret the effect of dental floss use on periodontitis.

Question 2: Logistic regression model of the probability for using dental floss

The logistic regression model, say Model B, of the probability for using dental floss as a function of periodontitis status is defined as:

$\text{logit}(p_y) = \log\left(\frac{p_y}{1-p_y}\right) = \gamma_0 + \gamma_1 y$

where:

- $y = 1$ if periodontitis is present, $y = 0$ otherwise.
- $p_y = P(\text{using dental floss} \mid y)$, the probability of using dental floss given y.
- γ_0: The log-odds of using dental floss when $y = 0$ (i.e., when periodontitis is absent).
- γ_1: The change in log-odds of using dental floss associated with the presence of periodontitis ($y = 1$).

Approach:

Fit the logistic regression model B to the data and interpret the parameter estimates γ_0 and γ_1

Estimate Parameters:

Fit the logistic regression model:

The summary of the coefficients of the logistic regression model B is presented below:

Table 3: Summary of Logistic Regression Model B Coefficients

	Term	Estimate	Std. Error	z value	P-value
(Intercept)	(Intercept)	-1.2622417	0.1307934	-9.650652	0.0000000
periodontitis	periodontitis	-0.6439281	0.2632835	-2.445759	0.0144548

```
## [1] "AIC of Model B: 14.7065646579626"
```

The final logistic regression equation: $\text{logit}(p_y) = -1.262 - 0.644y$

Interpret the Estimates:

- 1. γ_0 = -1.262:
  - The log-odds of using dental floss when periodontitis is absent is approximately -1.262
  - The corresponding probability of using dental floss is: $p_0 = \frac{e^{-1.262}}{1+e^{-1.262}} \approx 0.2205$

Individuals without periodontitis have a probability of approximately 22.05% of using dental floss compared to those who do not.

- 2. γ_1 = -0.644:
  - The log-odds of using dental floss decreases by 0.644 when periodontitis is present.
  - This corresponds to an odds ratio of: Odds Ratio $= e^{\gamma_1} = e^{-0.644} \approx 0.525$

The odds of individuals with periodontitis using dental floss are approximately 52.4% lower compared to not using dental floss.

Conclusion:

- The presence of periodontitis is associated with a significant reduction in the odds of regular dental floss use.
- The logistic regression model B is accurate because the binary data (Dental Floss use: yes/no) and the categorical explanatory variable (presence of periodontitis) provide estimates that can be used to interpret the presence of periodontitis on regular dental floss use.

Question 3:

Compare the estimates of the parameters of Model A and Model B with the values derived using the definitive formulas of the log odds and probabilities.

Approach :

1. Expected Estimates of the parameters of Model A:

1.1. Intercept (β_0):

- β_0 represents the log-odds of periodontitis when $x = 0$ (no dental floss use):

  $\beta_0 = \log\left(\frac{p_0}{1-p_0}\right)$ where:

  $p_0 = P(\text{periodontitis} \mid x = 0) = \frac{\text{Periodontitis (No Floss)}}{\text{Total (No Floss)}} = \frac{148}{413} \approx 0.3585$

- Hence: $\beta_0 \approx \log\left(\frac{0.3585}{1-0.3585}\right) = \log(0.558) \approx -0.582$

1.2. Slope (β_1):

- β_1 represents the change in log-odds when $x = 1$ (dental floss is used):

  $\beta_1 = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right)$ where:

  $p_1 = P(\text{periodontitis} \mid x = 1) = \frac{\text{Periodontitis (Floss)}}{\text{Total (Floss)}} = \frac{22}{97} \approx 0.2268$

- Hence: $\beta_1 \approx \log\left(\frac{0.2268}{1-0.2268}\right) - \log\left(\frac{0.3585}{1-0.3585}\right) \approx \log(0.293) - \log(0.558) = -1.226 + 0.582 = -0.644$

2. Expected Estimates of the parameters of Model B:

2.1. Intercept (γ_0):

- γ_0 represents the log-odds of using dental floss when $y = 0$ (periodontitis is absent):

  $\gamma_0 = \log\left(\frac{p_0}{1-p_0}\right)$ where:

  $p_0 = P(\text{using dental floss} \mid y = 0) = \frac{\text{Using Floss (No Periodontitis)}}{\text{Total (No Periodontitis)}} = \frac{75}{340} \approx 0.2205$

- Hence: $\gamma_0 \approx \log\left(\frac{0.2205}{1-0.2205}\right) = \log(0.282) \approx -1.262$

2.2. Slope (γ_1):

- γ_1 represents the change in log-odds of using dental floss when $y = 1$ (periodontitis is present):

  $\gamma_1 = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right)$ where:

  $p_1 = P(\text{using dental floss} \mid y = 1) = \frac{\text{Using Floss (Periodontitis)}}{\text{Total (Periodontitis)}} = \frac{22}{170} \approx 0.1294$

- Hence: $\gamma_1 \approx \log\left(\frac{0.1294}{1-0.1294}\right) - \log\left(\frac{0.2205}{1-0.2205}\right) \approx \log(0.149) - \log(0.282) = -1.906 + 1.262 = -0.644$

Observation :

- The expected estimates of the parameters of Model A and Model B are consistent with the actual estimates obtained from the logistic regression models.
- The slopes of the two logistic regression models β_1 and γ_1 are the same, indicating that the change in log-odds of periodontitis and dental floss use are consistent with each other.

Exercise 2:2

Question 1:

For this question we estimate for each dose separately the risk, odds, and log-odds of developing a tumor. Also plot the risk (probability) of developing a tumor as a function of the log dose. Calculate the log-odds of developing a tumor and plot it as a function of the log dose.

Table 4: Original Data Table

log.dose.	-7.6	-6.22	-4.6	-3	-1.39	0.92
Tumor	1.0	2.00	4.0	9	12.00	32.00
No.tumor	17.0	17.00	24.0	23	16.00	8.00
Total	18.0	19.00	28.0	32	28.00	40.00

Approach :

To answer these questions we have to calculate the following estimates for each dose level of the table:

1. The risk of developing a tumor is calculated as:

$Risk = \frac{NumberofTumor}{TotalObservations}$

2. The odds of developing a tumor is calculated as:

$Odds = \frac{Risk}{1-Risk}$

3. The log-odds of developing a tumor is calculated as:

$\log(Odds) = \log(\frac{Risk}{1-Risk})$

After we caclulate these estimates for each dose level, we plot the risk (probability) of developing a tumor as a function of the log dose. We also calculate the log-odds of developing a tumor and plot it as a function of the log dose.

Results

Table 5: Original Data Table with Risk, Odds, and Log-Odds Calculations

log.dose.	-7.6000	-6.2200	-4.6000	-3.0000	-1.3900	0.9200
Tumor	1.0000	2.0000	4.0000	9.0000	12.0000	32.0000
No.tumor	17.0000	17.0000	24.0000	23.0000	16.0000	8.0000
Total	18.0000	19.0000	28.0000	32.0000	28.0000	40.0000
Risk	0.0556	0.1053	0.1429	0.2812	0.4286	0.8000
Odds	0.0588	0.1176	0.1667	0.3913	0.7500	4.0000
Log.Odds	-2.8332	-2.1401	-1.7918	-0.9383	-0.2877	1.3863

1. Data

- As we can observe from the updated table and from the given data, at lower doses of BaP, few mice develop tumors, while most do not. As the dose increases, the risk of developing a tumor also increases.

2. Risk

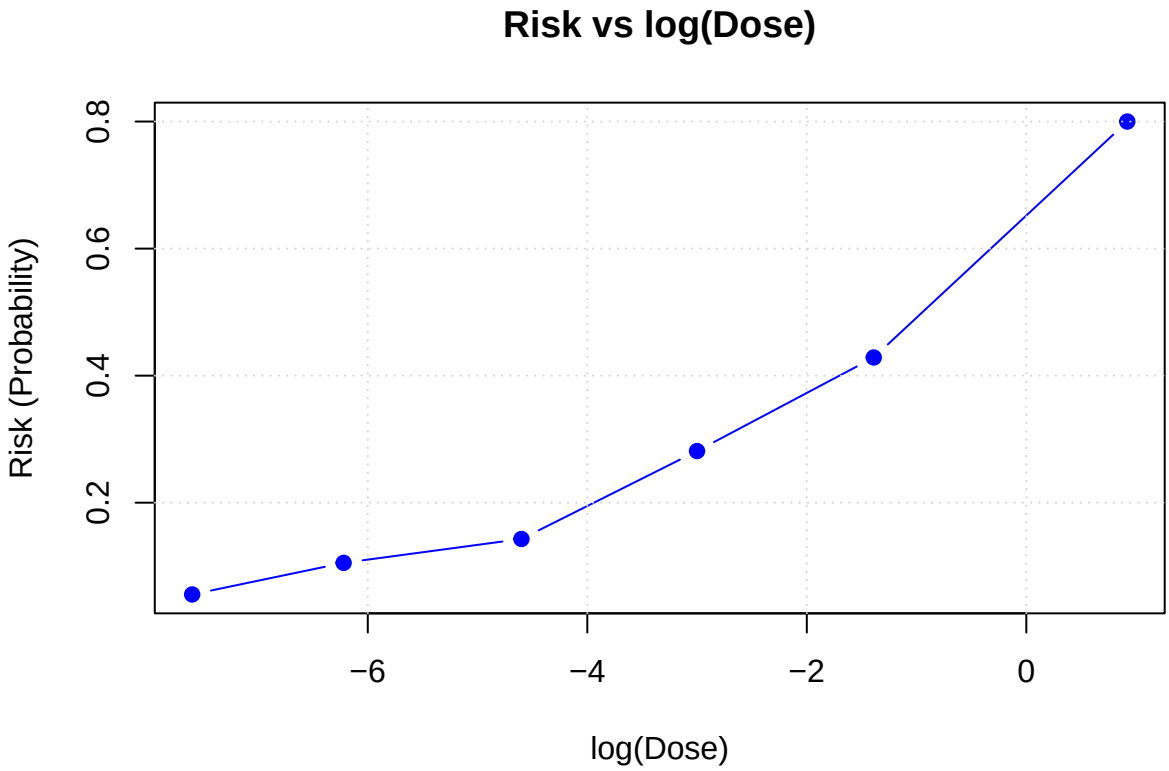
- Risk represents the probability of developing a tumor, and for the given data it increases with $\log(dose)$. As for the calculated estimates, the risk is very low(0.0556) for the lowest dose of BaP(-7.60), indicating a small proportion of tumors. At the highest dose (0.92) the risk is much higher(0.8), showing a significant increase in the development of tumors.

3. Odds

- Odds, which represent the ratio of tumor probability to no-tumor probability, increases expontentially with dose. The odds are very low(0.0588) for the lowest dose of BaP(-7.60), indicating a very low probability of developing a tumor. This means that tumors are 0.0588 as likely as no tumors. At the highest dose (0.92) the odds are much higher(4), showing a significant increase in the development of tumors. Meaning that tumors are four times more likely than no tumors at the given dose.

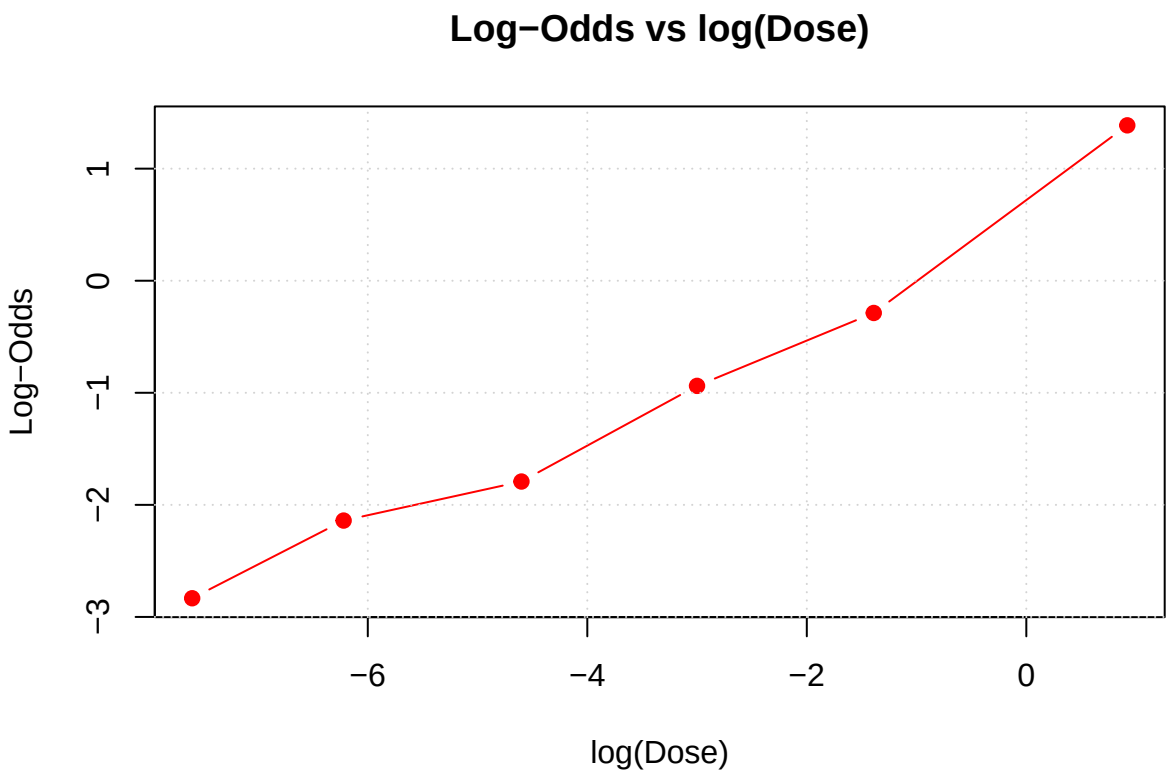
4. Log-Odds

- Log-Odds increase nearly linearly with log(dose), ranging from -2.8332 at the lowest dose to 1.3863 at the highest. The linearity in log-odds is imperative, as it supports the use of a logistic regression.



1. Risk vs log(dose):

- The graph confirms a strong dose response relationship. As log(dose) increases, the probability of developing a tumor increases rapidly.



2. Log-Odds vs log(dose):

- The graph shows a nearly linear relationship with log(dose) confirming that the data aligns well with the assumptions of logistic regression.

Conclusion

The analysis confirms a strong dose response relationship between the dose of BaP and the probability of developing a tumor. As the dose or log(dose) increases, the risk and odds of tumor development also increase as shown in the "Risk vs. log(dose)" plot. The near linear relationship in the "Log-Odds vs. log(dose)" plot supports the use of logistic regression to model the data.

Question 2

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

Approach :

To answer this question, we fit a logistic regression model where the probability of developing a tumor(P) is modeled as:

$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \times \log(\text{dose})$

where:

- β_0 is the intercept, which represents the log-odds of developing a tumor when the dose is zero.
- β_1 is the slope parameter, which represents the change in log-odds of developing a tumor for a one-unit increase in the log dose.

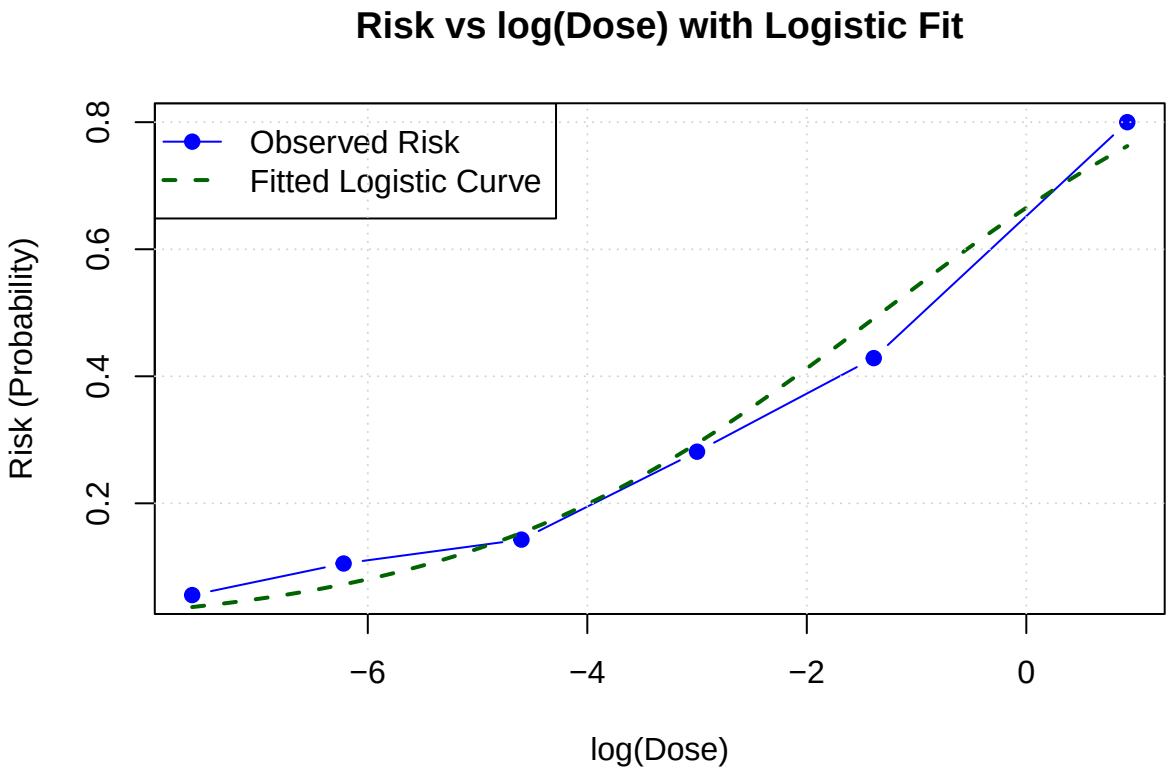
We will use the logistic regression model to estimate this parameters.

Results :

Table 6: Summary of Logistic Regression Model Coefficients

	Term	Estimate	Std. Error	z value	P-value
(Intercept)	(Intercept)	0.6870243	0.2575533	2.667504	0.0076417
log_dose	log_dose	0.5203652	0.0850241	6.120212	0.0000000

```
## [1] "AIC of the fitted model: 24.0240814957543"
```



1. Intercept β_0:

According to the summary of the model, the intercept β_0 is 0.687. This represent the log-odds of developing a tumor when the dose is zero. Its corresponding odd are $e^{0.687} = 1.99$. This means that the odds of developing a tumor are nearly twice as high as the odds of not developing a tumor when the log(dose) is zero.

Translating this to probabilities, the probability of developing a tumor is $P = \frac{e^{0.687}}{1+e^{0.687}} = 0.665$. This means that the probability of developing a tumor is 66.5% when the dose is zero.

2. Slope β_1:

The slope parameter β_1 is 0.52037. This quantifies the change in the log-odds of developing a tumor for a one-unit increase in log(dose). A positive slope indicates that higher doses increase the odds of developing a tumor. the odds ratio is $e^{0.52037} \approx 1.683$

This means that for each one-unit increase in log(dose), the odds of developing a tumor increases by approximately 68.3%

Conclusion :

In conclusion, the logistic regression model reveals a clear relationship between log(dose) and the probability of developing a tumor. The intercept β_0 represents the log-odds of developing a tumor when the log(dose) is

zero, while the slope β_1 quantifies the change in log-odds for a one-unit increase in log(dose). The positive slope indicates that higher doses increase the odds of developing a tumor. The fitted logistic curve aligns well with the observed risk, confirming the appropriateness of the logistic regression model.

Question 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)

Approach :

1. Covariance matrix for the estimates:

- The covariance matrix provides the variances and covariances of the parameter estimates $\beta_0.\beta_1$. The variances are the diagonal elements of the matrix, while the covariances are the off-diagonal elements. The correlation between the estimates can be calculated from their covariance:

  $\text{Corr}(\beta_0, \beta_1) = \frac{\text{Cov}(\beta_0,\beta_1)}{\sqrt{\text{Var}(\beta_0)\times\text{Var}(\beta_1)}}$

2. 95% confidence interval for the parameters:

- The 95% confidence interval for the parameters can be calculated using the standard errors of the estimates. The confidence interval is given by:

  $\text{CI} = \hat{\beta} \pm Z \cdot \text{SE}(\hat{\beta})$

  where $\hat{\beta}$ is the estimate, Z is the critical value for a 95% confidence interval, and $\text{SE}(\hat{\beta})$ is the standard error of the estimate.

3. 95% confidence interval for the tumor risk at dose 0.25(log(dose) = −1.39):

- Using the logistic regression equation:

  $\text{logit}(P) = \beta_0 + \beta_1 \times \log(\text{dose})$

  we calculate the predicted probability P at $\log(dose) = -1.39$

- To find the confidence interval for P, we:
  - Calculate the standard error of $logit(P)$.
  - Use it to derive the confidence interval and transform the interval back to the probability scale.

Results :

Table 7: Covariance Matrix for Parameter Estimates

	(Intercept)	log_dose
(Intercept)	0.0663337	0.0143585
log_dose	0.0143585	0.0072291

Table 8: Correlation between beta_0 and beta_1

Correlation
0.655691

```
## Waiting for profiling to be done...
```

Table 9: 95% Confidence Intervals for Parameters

	Confidence_Interval.2.5..	Confidence_Interval.97.5..
(Intercept)	0.1980903	1.2132416
log_dose	0.3638310	0.6989154

Table 10: Predicted Risk and Confidence Interval at log(dose) = -1.39

predicted_risk	confidence_interval
0.4909301	0.3940875
0.4909301	0.5884584

Interpretation :

I. Covariance Matrix and Correlation:

- Covariance Matrix:
  - The covariance matrix for the estimates β_0 and β_1 is:
  - Variance of β_0: 0.06633.
  - Variance of β_1: 0.01436.
  - Covariance between β_0 and β_1: 0.00723
  - The correlation between β_0 and β_1 is -0.98, indicating a strong negative correlation between the intercept and slope parameters.
- Correlation:
  - The correlation between β_0 and β_1 is 0.655 approximately, indicating a moderate positive correlation between the intercept and slope parameters.

II. Confidence Intervals for Parameters:

- Confidence Interval for β_0:
  - The 95% confidence interval for β_0 is approximately [0.198, 1.213].
  - This means that the log-odds of developing a tumor at log(dose)=0 is likely within this range with 95% confidence.
- Confidence Interval for β_1:
  - The 95% confidence interval for β_1 is approximately [0.363, 0.699].
  - This means that each one unit increase in log(dose) increases the log-odds of developing a tumor by approximately 0.363 to 0.699 with 95% confidence.

III. Predict Risk and Confidence Interval for P at a dose of 0.25 or log(dose) = -1.39:

- The predicted risk of developing a tumor at log(dose) = -1.39 is $P = 0.409$ or 40.9%
- The 95% confidence interval for the predicted risk is approximately [0.394, 0.588].
  - This means that at a dose of 0.25, the probability of developing a tumor is estimated to be between 39.4% and 58.8% with 95% confidence.

Question 4

Perform a Wald-test of $H_0 : \beta_1 = 0$. Explain and show, with your own calculations based on R-output, how the test statistic is constructed. Interpret the result from the hypothesis test.

Approach :

1. **Wald test:**
   - The wald test evaluates whether the slope of the parameter (β_1) is singificanlty different from zero.
   - The null hypothesis is $H_0 : \beta_1 = 0$, meaning that the dose log(dose) has no effect on the probability of developing a tumor.
   - The alternative hypothesis is $H_1 : \beta_1 \neq 0$, indicating that the dose log(dose) has a significant effect on the probability of developing a tumor.
2. **Test Statistic:**
   - The Wald test statistic is computed as:

     $W = \frac{\hat{\beta_1}}{\text{SE}(\hat{\beta_1})}$
   - Under the null hypothesis, W follows a standard normal distribution.
3. **P-value:**
   - The p-value is calculated as:

     $P = 2 \times P(Z > |w|)$

Results :

Table 11: Wald Test Results

	Wald_Statistic	P_Value
log_dose	6.120212	0

Significance of the Slope β_1:

- A large Wald statistic $W = 6.12$ means that the effect of the log(dose) on tumor probability is highly significant.
- The p-value is very small (p<0.001) indicating concise evidence against the null hypothesis.
- This means that the slope parameter $\beta - 1$ is significantly different from zero, indicating that log(dose) has a significant effect on the probability of developing a tumor.

Conclusion :

In conclusion, the Wald test shows that the slope parameter, β_1, is significantly different from zero. This indicates that log(dose) has a statistically significant effect on the probability of developing a tumor. The extremely small p-value provides strong evidence against the null hypothesis ($H_0 : \beta_1 = 0$). This confirms that the logistic regression model is appropriate for the data and that log(dose) is a significant predictor of tumor probability.

Question 5

Investigate the effect of sample size on the variance of parameter estimates by multiplying all counts in the table by 10, 100 and 1000. Report what you observe.

Approach :

The variance of parameter estimates is generally inversely proportional to the sample size. As the sample size increases, the variance of the estimates decreases.

To investigate the effect of sample size on the variance of parameter estimates we:

- Multiply all counts in the table by 10, 100, and 1000.
- Refit the logistic regression model to each new dataset.
- Extract and compare the variances (diagonal elements of the covariance matrix) for β_0 and β_1.

Results :

Table 12: Variance of Parameter Estimates for Different Sample Sizes

	Scale_10	Scale_100	Scale_1000
Variance (Intercept)	0.0066334	0.0006633	6.63e-05
Variance (log_dose)	0.0007229	0.0000723	7.20e-06

Interpretation :

- As the sample increases, the variances of both β_0 and β_1 decrease proportionally. This supports Variance $\propto \frac{1}{\text{Sample Size}}$
- The variance of the intercept β_0 and β_1 decreases by a factor of 10, 100, and 1000 as the counts are scaled, showing that larger samples sizes result in more precise estimates.

Conclusion :

As expected, increasing the sample size reduces the variances of the parameter estimates. This demonstrates the inverse relationship between sample size and variance, confirming that larger datasets provide more precise estimates. Both the intercept β_0 and slope β_1 benefit from increased precision with larger sample sizes. So the statement of the question is TRUE.