In this lecture, you will learn the following:

1. Introduction to Goodness of Fit for a Model using graphical methods.
2. Assessing the Goodness of Fit for a Logistic Regression Model through:
   a. Graphical method.
   b. Statistical testing method.

Contents

Goodness of Fit for a Model

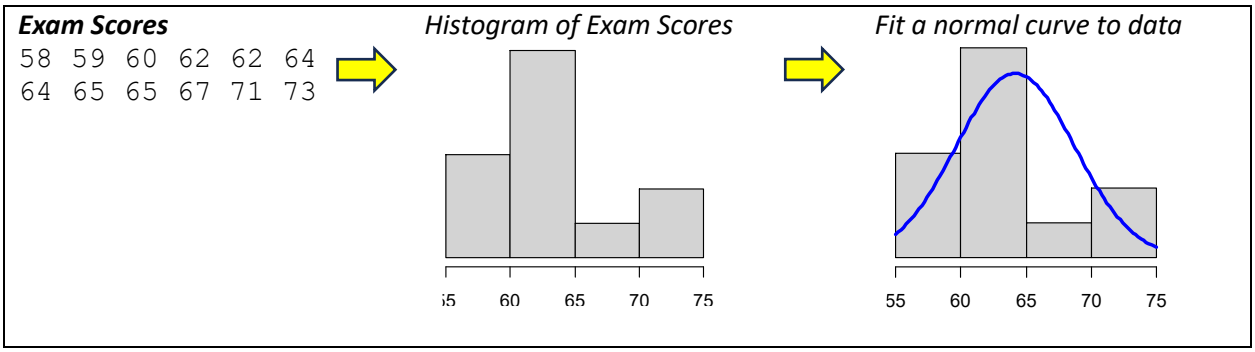
Before fitting a model to data (whether it's a regression model or another type), it's essential to visualize the data using graphs. This helps assess whether the proposed model makes sense.

Example 1: Exam Scores and Normality

Suppose we have a random sample of students' exam scores. We want to determine whether it's appropriate to fit a normal model to these scores. How can we check this?

Steps:

1. Create a histogram of the exam scores.
2. Fit a normal curve to the data and observe how well it matches the distribution.



If the histogram shows a symmetric, bell-shaped curve, it suggests that the data may follow a normal distribution.

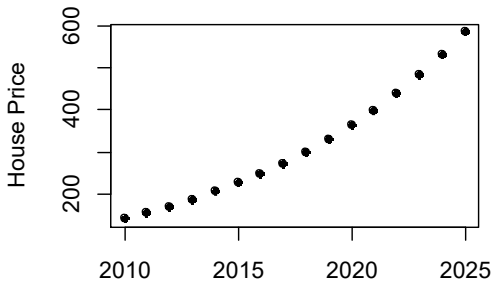
Example 2: Median House Prices Over Time

Let's say you have data on the median house price (in thousands of dollars) in a city from 2010 to 2025. You want to investigate how the median house price changes over time.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Price	140	154	169	186	205	225	248	273	300	330	363	399	439	483	531	585

Steps:

1. Create a scatterplot of the data to examine the relationship between the year and price.
2. Look for patterns: Is the relationship linear or nonlinear? Does it increase or decrease?

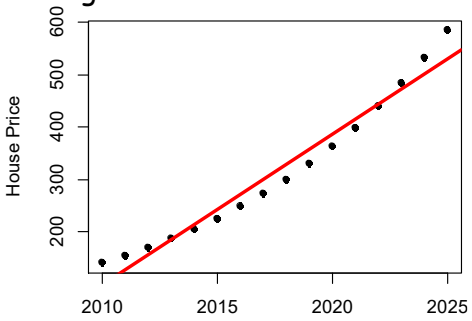


If you fit a regression line to model the median house price as a function of time

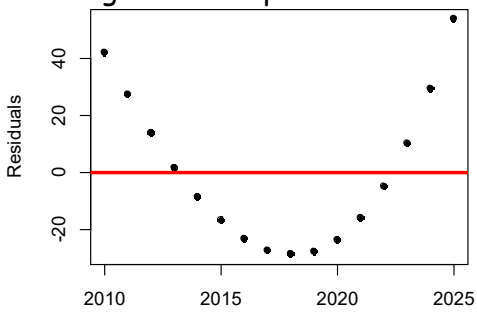
(e.g., **Price = A + B × Year**),

you can evaluate the residuals (the difference between predicted and actual values) to check whether the regression line is a good fit.

Fitting a line to the data



Checking Residuals plot



Since a residual plot that shows an obvious curve, this indicate that a linear model isn't appropriate for the data.

Goodness of Fit for a Logistic Regression Model

Example: Hours Studied and Exam Outcome

Let's consider a random sample of students from a large statistics class. We record the following variables:

- Number of hours studied,
- Exam Grade **(pass P or fail F)**.

The full dataset ***"Hours and Grade"*** is available for download on Brightspace.

Hours	Grade
0	F
0	F
0.5	F
1.5	F
1.5	F
1.5	P
2	F
2.5	F
2.5	F
⋮	⋮
10.5	**P**
11	**P**
11	**P**

We use the **logistic regression model** to relate the **log-odds of** passing to the number of hours studied.

$$\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\boldsymbol{log-odds}} = A + B * x$$

where:

- p is the probability that a student pass the exam,
- A and B are the model coefficients, and
- x represents the number of hours studies

The probability of success, p, is then given by the logistic function:

$$p = \frac{e^{A+Bx}}{1+e^{Bx}}$$

In a logistic regression model, we assume that the **log-odds of success** are a linear function of the hours.

Question: How can we determine if the log-odds of passing increase (or decrease) linearly with the predictor (e.g. hours), and whether the probability of passing can be adequately modeled by the logistic function?

Graphical Methods – Problem

First, we can plot the exam outcome against the number of hours studied.

However, since the exam outcome is a categorical variable (pass/fail), we need to convert it into a numerical variable before creating the plot.

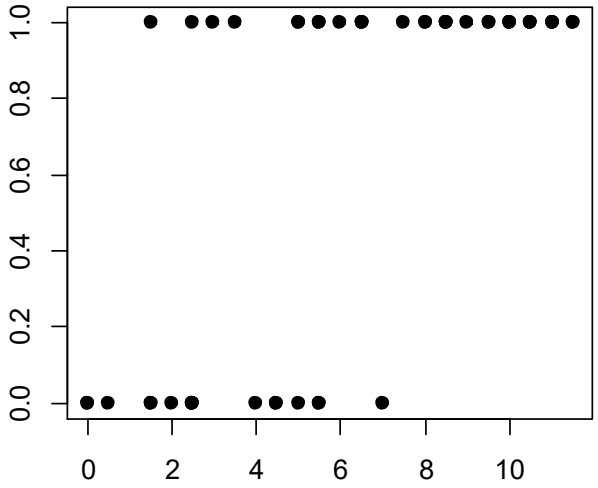
Let's define the numeric exam outcome as follows:

- 1 if the outcome is pass (P),
- 0 if the outcome is fail (F).

Let's do this in R:

```
# Create numeric exam grade using the ifelse(...) function
# If the exam grade is "Pass", assign 1; otherwise, assign 0
numeric.grade <- ifelse(mydata$Grade == "P", 1, 0)

# Create the scatterplot
plot(
  mydata$Hours,          # Hours studied on the x-axis
  numeric.grade,         # Numeric exam outcome on the y-axis
  lwd = 3,               # Set the line width to 3
  pch = 20,              # Use solid circles to represent the points
  cex = 1.2,             # Set the size of the points (1.2x the original size)
  main = "Numeric Exam Outcome vs Hours Studied",  # Title of the plot
  xlab = "Hours",                # Label for the x-axis
  ylab = "Numeric Exam Outcome" # Label for the y-axis
)
```

The following are the results.



As we can see, the plot does not look like a standard scatterplot. Instead, it only displays a distribution of 0s (fail) and 1s (pass) along the x-axis (hours studied).

This graph **doesn't** show any clear relationship in terms of form (e.g. linear or nonlinear) or direction (e.g. positive or negative) that we can interpret.

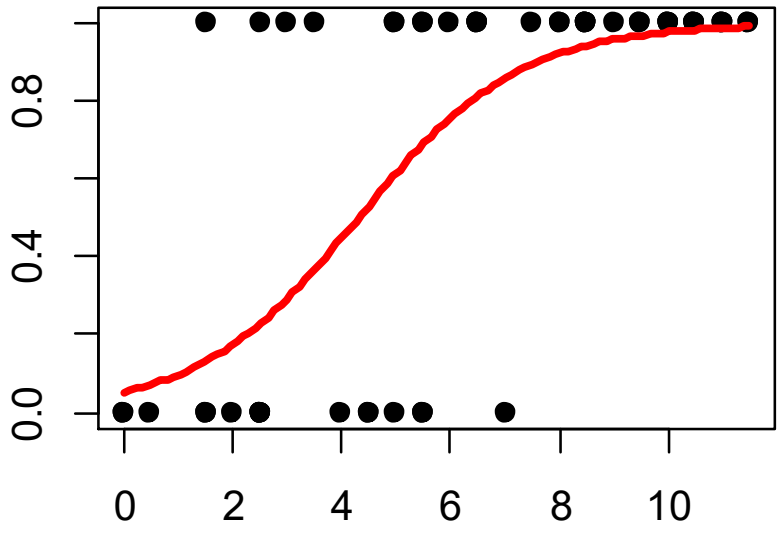
Note: This graph plays an important role in determining whether there will be any estimation issues when fitting a logistic regression model. I will explain it to you later.

Next, we fit the logistic regression model to the data and obtain the estimated equation that predicts the probability of passing. Then, we add the logistic curve to the plot.

Let's do it in R:

```
# Define the predictor variable (x), which is "Hours" in this case
x = mydata$Hours


# Define the response variable (y)
#  - the variable we are predicting its outcomes, which is the "Grade"
y = mydata$Grade

#Convert this into a categorical variable, Grade with two levels: Pass and Fail
y = factor(mydata$Grade)

#Fit the logistic regression model to data
fitted.model = glm( y ~ x, family = binomial)


# Extract the estimated coefficients from the fitted model
intercept <- fitted.model$coef[1]
slope <- fitted.model$coef[2]

# Define the logistic function for plotting
Logistic.func <- function(x) {
  exp(intercept + slope * x) / (1 + exp(intercept + slope * x))
}

# Add the logistic regression curve to the plot
curve(
  logistic.func(x),     # Use the logistic function to generate values
  from = min(x),        # Set the range for the x-axis (from min to max of x)
  to = max(x),          # Set the range for the x-axis (from min to max of x)
  lwd = 3,              # Set the line width of the curve
  col = "red",          # Set the color of the curve to red
  add = TRUE            # Add the curve to the existing plot
)
```

Below are the results.



As we can see, the data points do not closely follow the curve, making it difficult to determine whether the logistic regression model is a good fit.

In my opinion, plotting a categorical response variable (e.g., exam outcome) against a numerical predictor (e.g., hours studied) is **NOT** an effective way to visualize their relationship.

Graphical Methods – Solutions to the Problem

To address this issue, we can take the following steps to better visualize the relationship.

Step 1: Divide the range of the predictor variable (e.g., hours studied) into intervals. For example:

$$0 \leq Hours \leq 3$$

$$3 \leq Hours < 6$$

$$6 \leq Hours < 9$$

$$9 \leq Hours < 12$$

Note: If you are unsure how to choose appropriate intervals, you can use **R** to generate them automatically. Read the attached R file.

The key point is that each interval should contain a sufficient number of observations (e.g., at least 10) to ensure meaningful comparisons.

Step 2 - Within each study-time interval, we count the number of students who **pass** and **fail** the exam, then compute the **proportion of passing**.

Hours	Grade	interval
0	F	$0 \leq Hours < 3$
0	F	
0.5	F	
1.5	F	
1.5	F	
1.5	P	
2	F	
2.5	F	
2.5	F	
2.5	F	
2.5	P	
2.5	F	
3	P	$3 \leq Hours < 6$
3.5	P	
4	F	
4.5	F	
4.5	F	
5	P	
5	F	
5	P	
5.5	F	
5.5	F	
5.5	P	
5.5	P	
6	P	$6 \leq Hours < 9$
6.5	P	
6.5	P	
6.5	P	
7	F	
7.5	P	
8	P	
8	P	
8.5	P	
8.5	P	
8.5	P	
8.5	P	
9	P	$9 \leq Hours < 12$
9.5	P	
9.5	P	
10	P	
10	P	
10	P	
10.5	P	
10.5	P	
10.5	P	
11	P	
11	P	
11	P	
11.5	P	
11.5	P	

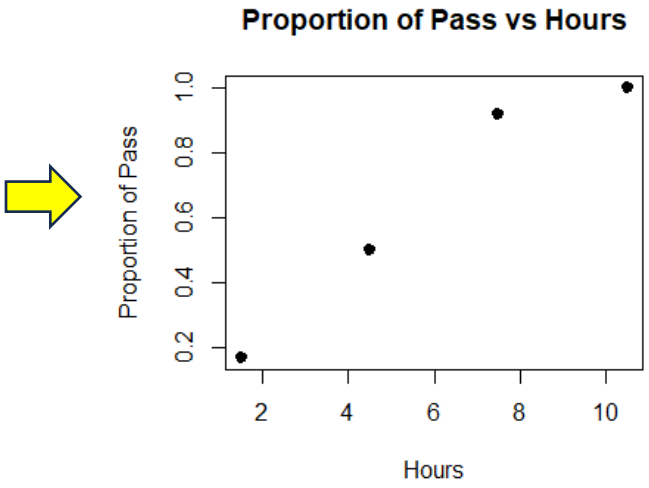
.

Hour Interval	# of Pass	# of Fail	Total	% of Pass
$0 \leq Hours < 3$	2	10	12	16.67%
$3 \leq Hours < 6$	6	6	12	50.00%
$6 \leq Hours < 9$	11	1	12	91.67%
$9 \leq Hours < 12$	14	0	14	100.00%

Step 3 - We plot the **proportion of students who pass** against the **midpoint of each hour interval**.

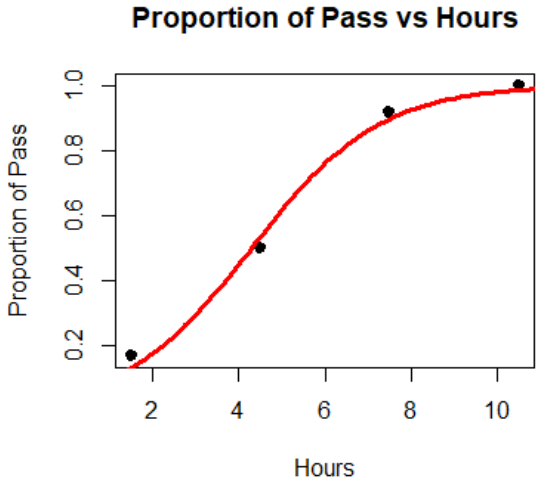
.

Hour Interval	% of Pass	Midt-point
$0 \leq Hours < 3$	16.67%	1.5
$3 \leq Hours < 6$	50.00%	4.5
$6 \leq Hours < 9$	91.67%	7.5
$9 \leq Hours < 12$	100.00%	10.5



From the plot, we observe that the **proportion of students who pass the exam increases as the number of hours studied increases**.

Finally, we overlay the **estimated logistic regression curve** on the plot.
The fitted curve closely follows the observed proportions, indicating that the **logistic regression model provides a good fit to the data**.



It is clearly that the logistic regression model fits the data well.

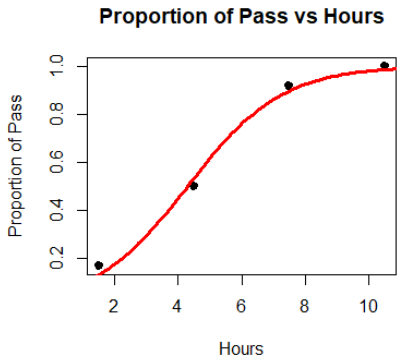
Cautions

This graphical method works best when the dataset is large. A larger sample ensures that each interval contains enough observations to produce more precise estimates of the success proportion .

When the sample size within an interval is small, the estimated proportions often take extreme values (0 or 1), which makes it difficult to visualize the underlying relationship between the predictor and the categorical response.

Statistical Testing Method - Hosmer-Lemeshow test

Using the graphical method to assess the goodness of fit for a logistic regression model can be subjective. For instance, one person might conclude that the fit is good, while another might consider it only moderately good. To address this issue, we use a goodness of fit test known as the **Hosmer-Lemeshow test**, which provides objective results.



Step 1 – State the null and alternative hypotheses:

- **Ho**: The logistic regression model is correct
- **Ha**: The logistic regression model is not correct

Step 2 – Fit the logistic regression model and estimate the probability of success (e.g., passing) for each value of the predictor (e.g., hours) using the regression equation.

Then, calculate the average of the estimated probabilities within each interval.

The output from fitting the logistic regression model to the data is provided below."

```
Call:
glm(formula = y ~ x, family = binomial)

Coefficients:
(Intercept)            x
    -2.8984       0.6734
```

The regression equation used to estimate the probability of passing is

$$Estimated\ Probability = \frac{e^z}{1+e^z}$$

$$where\ z = -2.8984 + 0.6734 * Hours$$

Let's use the regression equation to calculate the probability of passing for each value of hours. The results are presented below

Hour Interval	Hours	Grade	Estimated Probability of Passing (using the regression equation)	Average Probability
$0 \leq Hours < 3$	0	F	0.052235	**Average = 0.157470**
	0	F	0.052235	
	0.5	F	0.071649	
	1.5	F	0.13145	
	1.5	F	0.13145	
	1.5	P	0.13145	
	2	F	0.174873	
	2.5	F	0.228861	
	2.5	F	0.228861	
	2.5	F	0.228861	
	2.5	P	0.228861	
	2.5	F	0.228861	
$3 \leq Hours < 6$	3	P	0.293585	**Average = 0.565537**
	3.5	P	0.367882	
	4	F	0.449029	
	4.5	F	0.532983	
	4.5	F	0.532983	
	5	P	0.61511	
	5	F	0.61511	
	5	P	0.61511	
	5.5	F	0.691164	
	5.5	F	0.691164	
	5.5	P	0.691164	
	5.5	P	0.691164	
$6 \leq Hours < 9$	6	P	0.758099	**Average = 0.881692**
	6.5	P	0.814422	
	6.5	P	0.814422	
	6.5	P	0.814422	
	7	F	0.860052	
	7.5	P	0.895897	
	8	P	0.923379	
	8	P	0.923379	
	8.5	P	0.944059	
	8.5	P	0.944059	
	8.5	P	0.944059	
	8.5	P	0.944059	
$9 \leq Hours < 12$	9	P	0.959403	**Average = 0.981686**
	9.5	P	0.970669	
	9.5	P	0.970669	
	10	P	0.978877	
	10	P	0.978877	
	10	P	0.978877	
	10.5	P	0.984824	
	10.5	P	0.984824	
	10.5	P	0.984824	
	11	P	0.989116	
	11	P	0.989116	
	11	P	0.989116	
	11.5	P	0.992203	
	11.5	P	0.992203	

Step 3 – Calculate the proportion of successes within each interval

Hour Interval	Hours	Grade	Estimated Probability of Passing (using the regression equation)	Average Probability	Passing Proportion
$0 \leq Hours < 3$	0	F	0.052235	**Average = 0.157470**	**# of Pass = 2** **Total = 12** **Proportion of passing** $= \frac{2}{12} = 0.1667$
	0	F	0.052235		
	0.5	F	0.071649		
	1.5	F	0.13145		
	1.5	F	0.13145		
	1.5	P	0.13145		
	2	F	0.174873		
	2.5	F	0.228861		
	2.5	F	0.228861		
	2.5	F	0.228861		
	2.5	P	0.228861		
	2.5	F	0.228861		
$3 \leq Hours < 6$	3	P	0.293585	**Average = 0.565537**	**# of Pass = 6** **Total = 12** **Proportion of passing** $= \frac{6}{12} = 0.5$
	3.5	P	0.367882		
	4	F	0.449029		
	4.5	F	0.532983		
	4.5	F	0.532983		
	5	P	0.61511		
	5	F	0.61511		
	5	P	0.61511		
	5.5	F	0.691164		
	5.5	F	0.691164		
	5.5	P	0.691164		
	5.5	P	0.691164		
$6 \leq Hours < 9$	6	P	0.758099	**Average = 0.881692**	**# of Pass = 11** **Total = 12** **Proportion of passing** $= \frac{11}{12} = 0.9167$
	6.5	P	0.814422		
	6.5	P	0.814422		
	6.5	P	0.814422		
	7	F	0.860052		
	7.5	P	0.895897		
	8	P	0.923379		
	8	P	0.923379		
	8.5	P	0.944059		
	8.5	P	0.944059		
	8.5	P	0.944059		
	8.5	P	0.944059		
$9 \leq Hours < 12$	9	P	0.959403	**Average = 0.981686**	**# of Pass = 14** **Total = 14** **Proportion of passing** $= \frac{14}{14} = 1$
	9.5	P	0.970669		
	9.5	P	0.970669		
	10	P	0.978877		
	10	P	0.978877		
	10	P	0.978877		
	10.5	P	0.984824		
	10.5	P	0.984824		
	10.5	P	0.984824		
	11	P	0.989116		
	11	P	0.989116		
	11	P	0.989116		
	11.5	P	0.992203		
	11.5	P	0.992203		

Step 4 - Compare the observed and estimated probability of success within each interval by calculating the following:

$$z^2 = \left(\frac{p_0 - p_e}{\sqrt{\frac{p_e(1-p_e)}{n}}} \right)^2$$

where
$p_0 =$ observed probability (or proportion) of success
$p_e =$ estimated probability of success based on the logistic model
$n =$ sample size within the interval

In this case:

- The observed probability of success (p_0) is the proportion of students who pass the exam within each interval.
- The estimated probability of success (p_e) is the average probability of a student passing the exam estimated by the logistic model.

Let's calculate z^2 in each interval. Finally, we sum all the z^2 values, which equals 0.61932.

Hour Interval	n (Total)	Proportion of Pass (p_0)	Average Probability of Passing (p_e)	z^2
$0 \leq Hours < 3$	12	0.16667	0.1575	0.00765
$3 \leq Hours < 6$	12	0.50000	0.5655	0.20977
$6 \leq Hours < 9$	12	0.91667	0.8817	0.14072
$9 \leq Hours < 12$	14	1.00000	0.9817	0.26119
			SUM	0.61932

Important Note:

$\sum z^2$ is the Chi-Square Statistic, which follows a Chi-Square distribution with degrees of freedom given by:

$$df = \text{Number of intervals - 2}$$

If **H_0 is true**, meaning the logistic regression model fits the data reasonably well,

- the difference between the observed and estimated probabilities of success should be **small**,
- resulting in a **small Chi-Square statistic**.

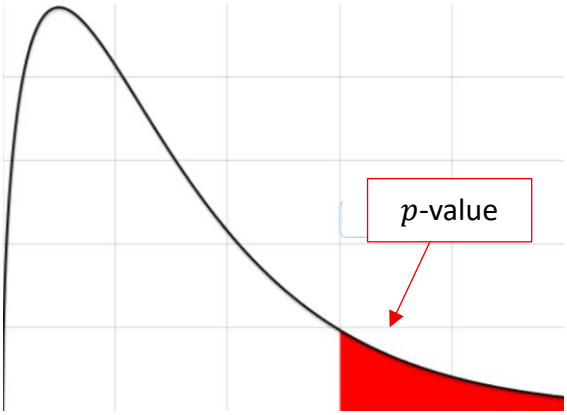
In contrast, a **larger** Chi-Square statistic indicates a **larger disagreement** between the observed and estimated probabilities, providing stronger evidence against H_0.

How can we determine if the Chi-Square statistic gives strong evidence against H_0?

We need to calculate the **p-value**.

Final Step – Calculate the p-value and draw a conclusion:

The p-value is the area under the right tail of the Chi-Square distribution, bounded by the calculated Chi-Square statistic (0.61932).



0.61932

Now, let's run the following R command to calculate the p-value.

```
num.of.intervals  = 4
df = num.of.intervals - 2
chisq.stat = 0.61932
pvalue = 1 - chisq(  chisq.stat, df)
```

```
> num.of.intervals  = 4
> df = num.of.intervals - 2
> chisq.stat = 0.61932
> pvalue = 1 - pchisq(  chisq.stat, df)
>
>pvalue
[1] 0.7336964
```

Let's compare the p-value (0.7337) with the 5% significance level.

Since the p-value is greater than the significance level, there is insufficient evidence to reject H_0, meaning the logistic regression model fits the data reasonably well.

Note: Although we do not reject H_0, this does not imply that the logistic regression model truly represents the underlying relationship between the probability of passing and hours. It simply means that the logistic regression model is a reasonable fit, but it may not be the true model

Cautions

The Hosmer-Lemeshow test is valid when the sample size within each interval is sufficiently large.How large? The expected number of successes and failures should be at least 5.

If this condition is not met, the test results may not be valid and should not be trusted.

Additionally, if the logistic regression model does not appear to provide a reasonable fit to the data based on the graph, then the test results lose their meaning. It is important to interpret the test results alongside with the graphical method.

In this example, the expected number of passes or fails is less than 5 in all intervals, so the test results are **not valid**.

Hour Interval	n (Total)	Average Probability of Passing (p_e)	Expected Number of Passes	Expected Number of fails
$0 \leq Hours < 3$	12	0.1575	12 * 0.1575 = **1.889** $<$ **5**	12 – 1.889 = 10.110
$3 \leq Hours < 6$	12	0.5655	6.786	**5.214** $<$ **5**
$6 \leq Hours < 9$	12	0.8817	10.580	**1.420** $<$ **5**
$9 \leq Hours < 12$	14	0.9817	13.744	**0.256** $<$ **5**