Logistic regression with multiple predictors

Scenario: dengue fever in Vietnam

- *Dengue fever*: a mosquito-borne viral disease, which infects hundreds of millions of people a year. Common in tropical climates
- Researchers in Vietnam are interested in identifying the relationship between specific symptoms and the probability of having dengue
- Lab tests are available, but may take time to return results and/or be expensive

Data

Data on 5720 Vietnamese children, admitted to hospital with possible dengue fever. Variables include:

- `Dengue`: whether the patient actually has dengue fever, based on a lab test (0 = no, 1 = yes)
- `Temperature`: patient's body temperature (in Celsius)
- `Abdominal`: whether the patient has abdominal pain (0 = no, 1 = yes)
- `HCT`: patient's hematocrit (proportion of red blood cells)
- `Age`: patient's age (in years)
- `Sex`: patient's sex
- + several others

Research question

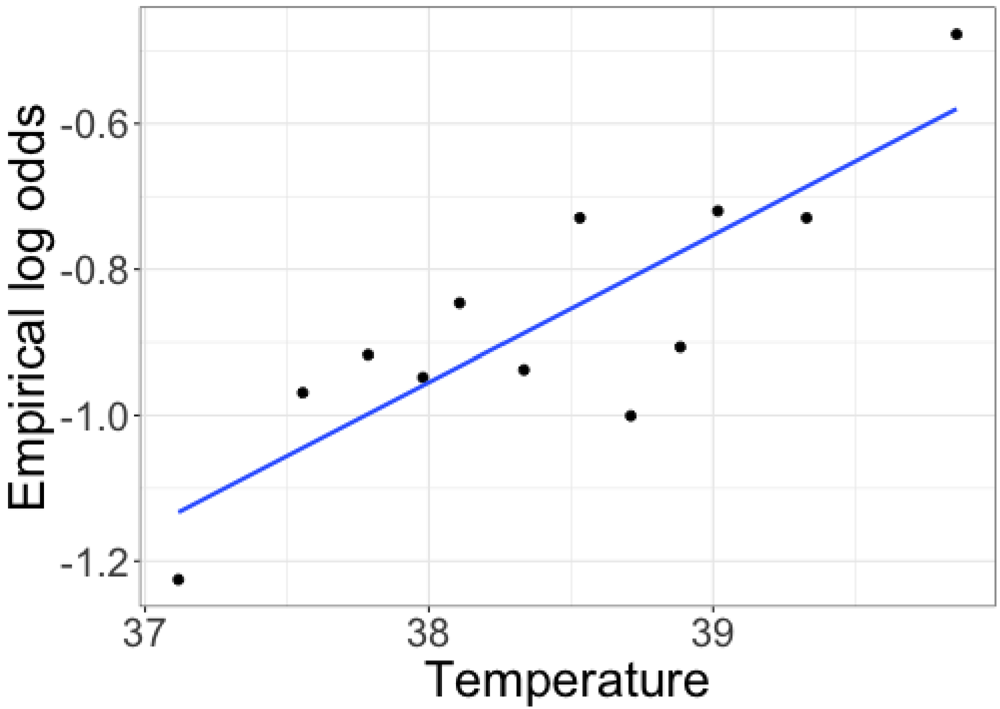
You are approached by the researchers to help analyze their data. Their initial question:

What is the relationship between temperature and the probability of dengue, and does the relationship differ if the patient also presents with abdominal pain?

Exploratory data analysis

> How can we visualize the relationship between temperature and the probability of having dengue?

Exploratory data analysis



> Based on the empirical log odds plot, what would be a reasonable model for the relationship between temperature and the probability a patient has dengue?

Initial model

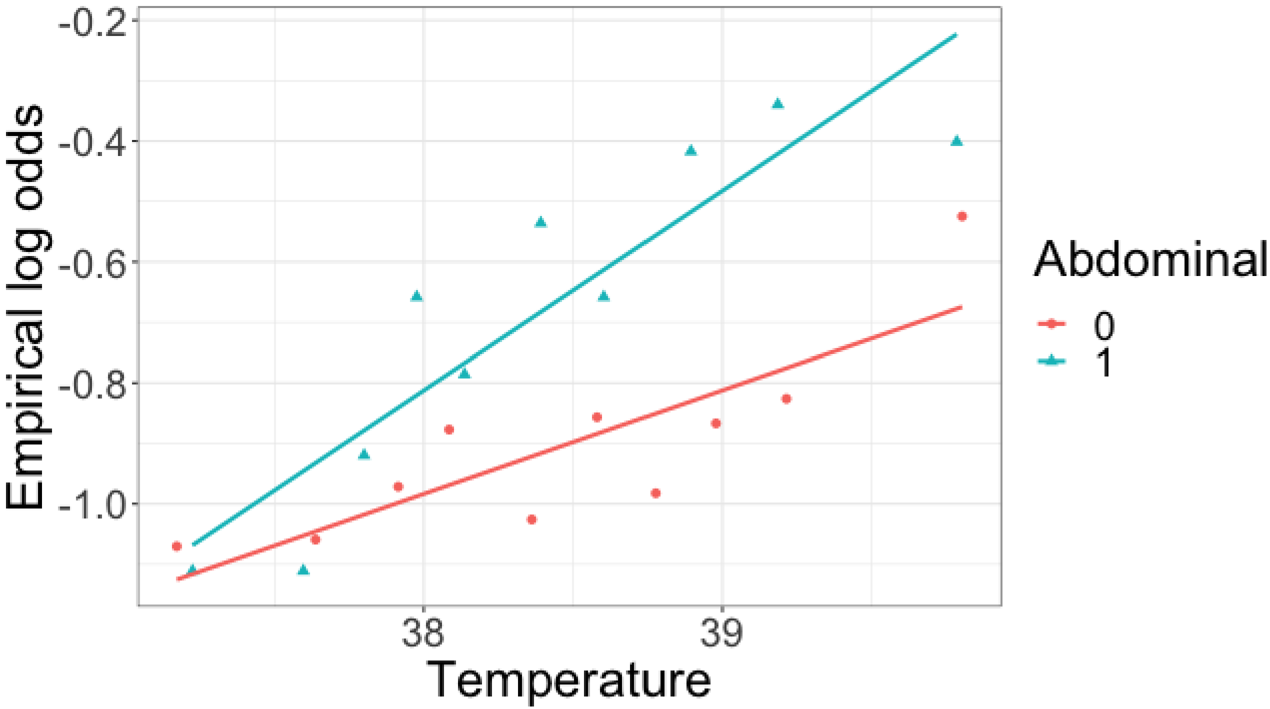
$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\, Temperature_i$$

We're also interested in whether this relationship is different depending on abdominal pain.

> What plot can I use to investigate this question?

Exploratory data analysis



> Based on this plot, what would be a reasonable model that incorporates both temperature and abdominal pain?

Adding abdominal pain

$$Y_i \sim Bernoulli(\pi_i)$$

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1\, Temperature_i + \beta_2\, Abdominal_i \\ &\quad + \beta_3\, Temperature_i \cdot Abdominal_i\end{aligned}$$

- The interaction term allows the relationship between temperature and π_i to change depending on abdominal pain

Class activity

https://sta214-f22.github.io/class_activities/ca_lecture_10.html

Class activity

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -7.745 + 0.178\, Temperature_i - 6.129\, Abdominal_i + 0.166\, Temperature_i \cdot Abdominal_i$$

> What is the estimated probability of dengue for a patient with a temperature of 38C and abdominal pain?

Class activity

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -7.745 + 0.178\, Temperature_i - 6.129\, Abdominal_i + 0.166\, Temperature_i \cdot Abdominal_i$$

> For patients with abdominal pain, what is the estimated change in odds associated with an increase in temperature of 1C?

Class activity

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 \, Temperature_i + \beta_2 \, Abdominal_i \\ + \beta_3 \, Temperature_i \cdot Abdominal_i$$

I want to test whether there is a relationship between abdominal pain and the probability of dengue, after accounting for the relationship between temperature and the probability of dengue.

> What are my null and alternative hypotheses?

Hypothesis testing

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ Temperature_i + \beta_2\ Abdominal_i \\ + \beta_3\ Temperature_i \cdot Abdominal_i$$

Hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_A : \text{at least one of } \beta_2, \beta_3 \neq 0$$

> Which type of test can I use to test these hypotheses (Wald test, likelihood ratio test, or either)?

Likelihood ratio test

$$Y_i \sim Bernoulli(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\, Temperature_i + \beta_2\, Abdominal_i \\ + \beta_3\, Temperature_i \cdot Abdominal_i$$

Hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_A : \text{at least one of } \beta_2, \beta_3 \neq 0$$

> What are my full and reduced models?

Full and reduced models

Full model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\, Temperature_i + \beta_2\, Abdominal_i + \beta_3\, Temperature_i \cdot Abdominal_i$$

Reduced model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\, Temperature_i$$

Test statistic

$$G = \text{deviance of reduced model} \; - \text{deviance of full model}$$

Full model:

```
## Null Deviance:          6956
## Residual Deviance: 6914     AIC: 6922
```

Reduced model:

```
## Null Deviance:          6956
## Residual Deviance: 6927     AIC: 6931
```

$$G =$$

> How do I calculate a p-value for this test statistic?

p-value

$$G = 6927 - 6914 = 13 \quad G \sim \chi^2_k$$

where $k =$ difference in number of parameters between full and reduced models.

> What is k for this test?

p-value

$$G = 6927 - 6914 = 13 \quad G \sim \chi^2_k$$

where $k =$ difference in number of parameters between full and reduced models.

```
pchisq(13, df=2, lower.tail=F)
```

```
## [1] 0.001503439
```

Conclusion

- **Question:** Is there a relationship between abdominal pain and the probability of dengue, after accounting for the relationship between temperature and the probability of dengue?
- **Hypotheses:**

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_A : \text{at least one of } \beta_2, \beta_3 \neq 0$$

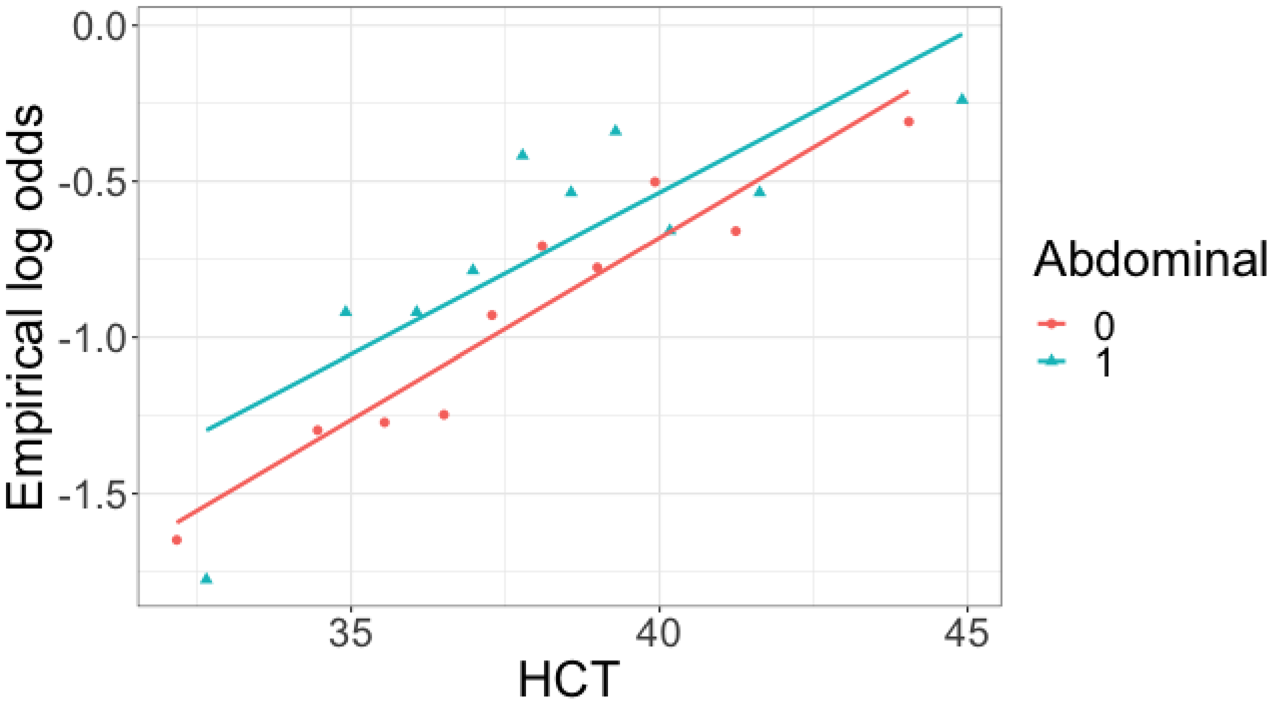
- **p-value:** 0.0015
- **Conclusion:** There is strong evidence that there is a relationship between abdominal pain and the probability of dengue, after accounting for the relationship between temperature and the probability of dengue.

A new question...

You report your results to the hospital, and they ask a follow-up question:

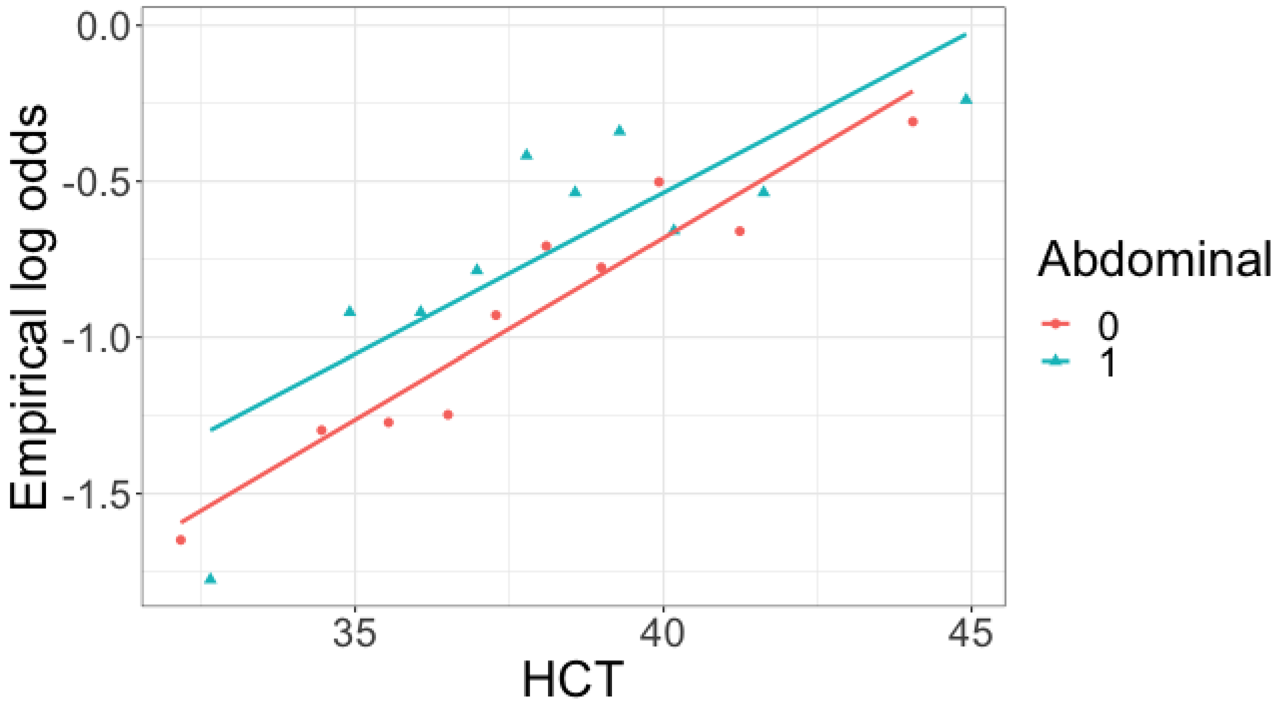
Does the model improve when we add hematocrit (the proportion of red blood cells)?

Exploratory data analysis



> Does it look like we need an interaction between hematocrit (HCT) and abdominal pain?

Exploratory data analysis

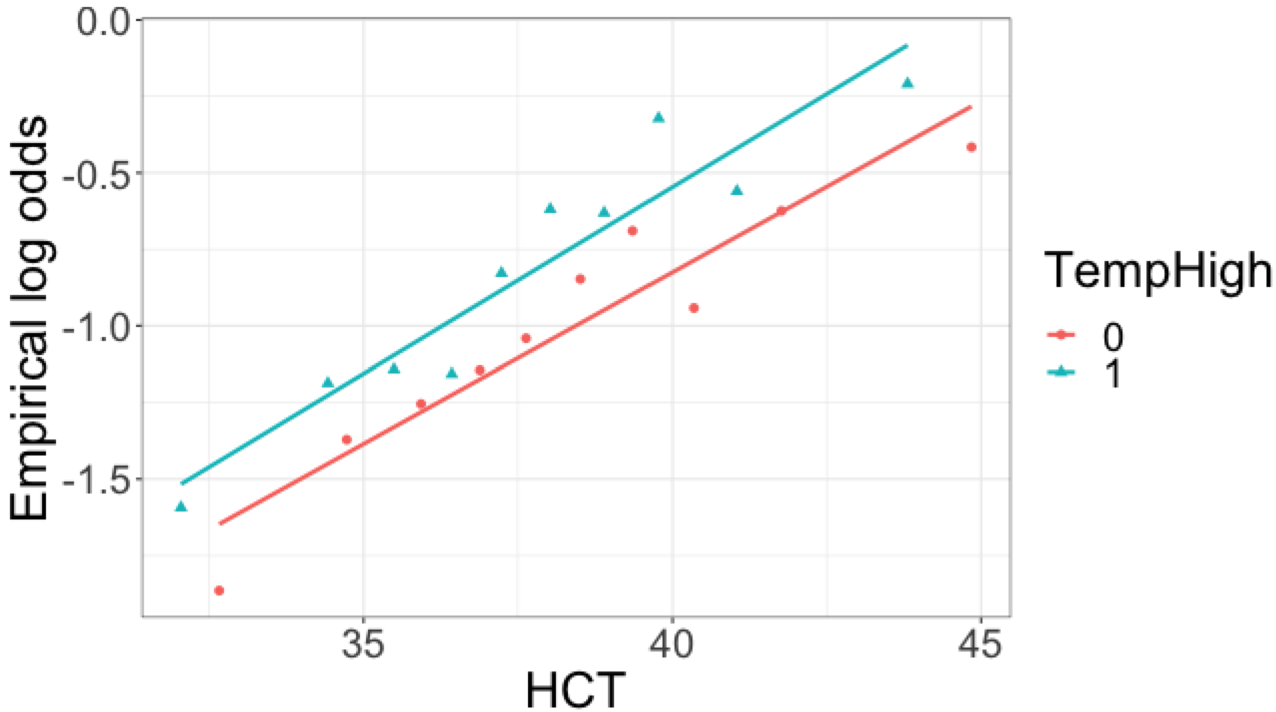


How can I check whether there might be an interaction between temperature and hematocrit?

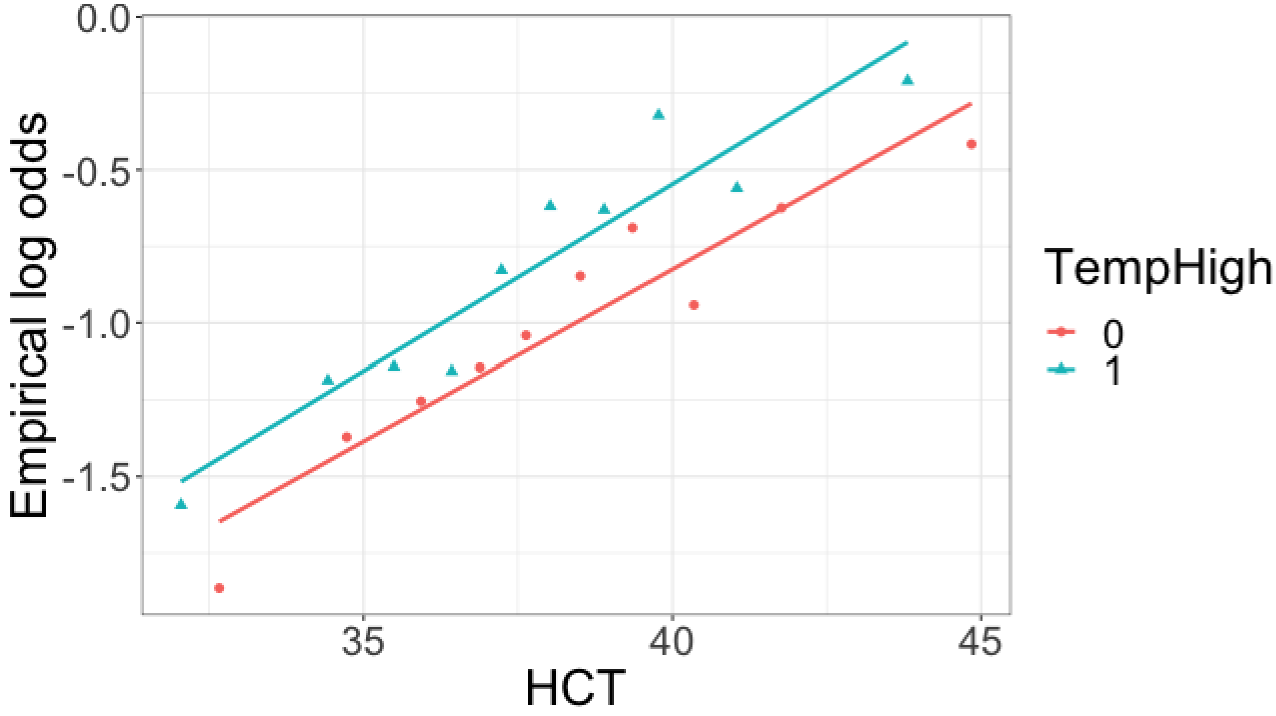
Exploratory data analysis

Define $TempHigh$ by

- $TempHigh_i = 1$ if $Temperature_i > 38$
- $TempHigh_i = 0$ if $Temperature_i <= 38$



Exploratory data analysis



> Does it look like we need an interaction between temperature and hematocrit?

Model

$$Y_i \sim Bernoulli(\pi_i)$$

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 &+ \beta_1\, Temperature_i + \beta_2\, Abdominal_i \\ &+ \beta_3\, Temperature_i \cdot Abdominal_i \\ &+ \beta_4\, HCT_i \end{aligned}$$

- Note that while we binarized temperature for EDA, we use the original variable in the model here