Week 12 Hour 1

How NOT to handle binary responses variables.

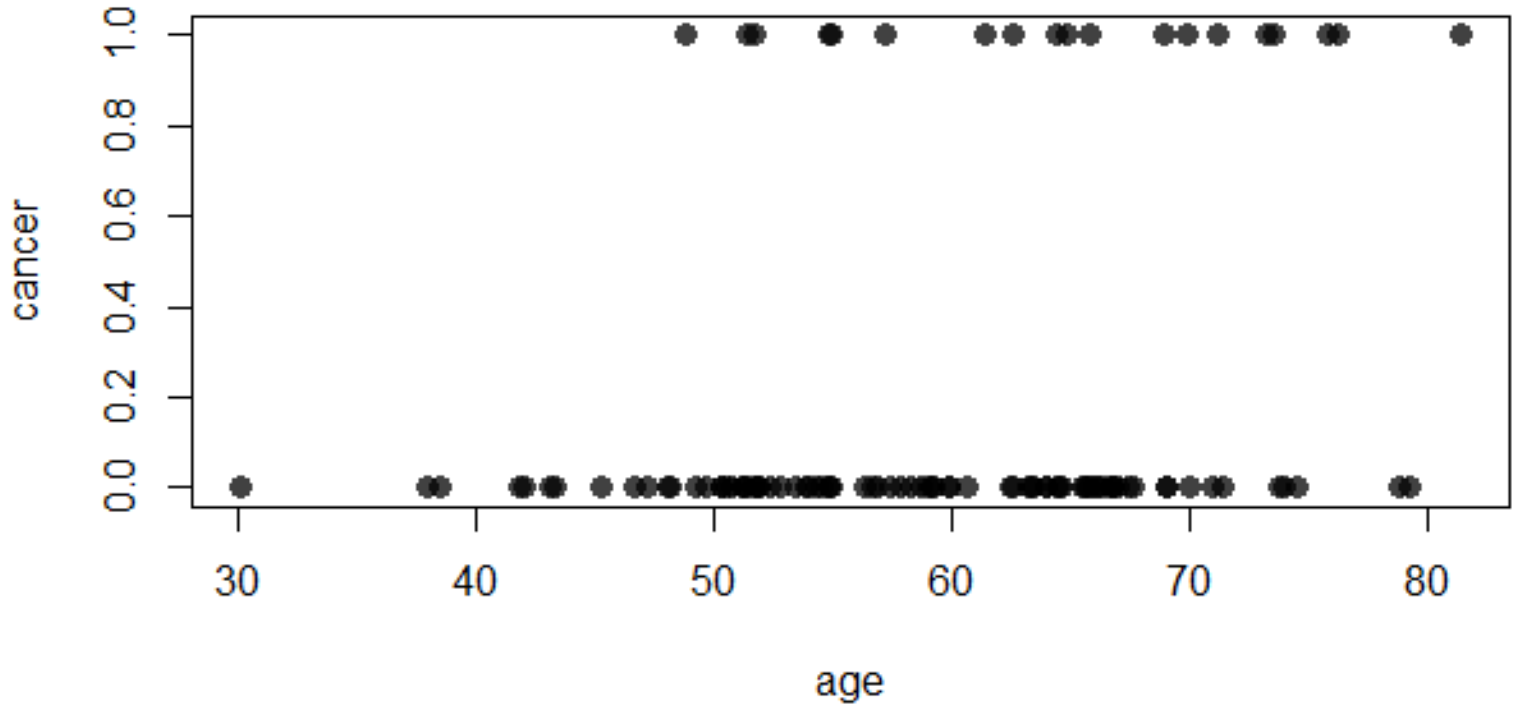
Odds, Log Odds, and the Logit Function

How not to model probability.

Consider the cancer1.csv dataset, which has 100 fictional patients, their ages and whether they have a certain kind of cancer.

	age	cancer
1	56	0
2	47	0
3	74	0
4	69	1
5	55	0
6	57	1

A scatterplot shows that having cancer is relatively rare, and that it appears more often as patients get older.



A linear regression is fit, where the binary variable cancer is the response and age is the explanatory.

Since 1 is ‘yes cancer’, and 0 is ‘no cancer’,
we can interpret a value between 0 and 1 as the probability of cancer.

If the regression model predicts a value such as 0.30, we can interpret that as a 30% of such a person having cancer.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.415413   0.230550  -1.802  0.07465 .
age          0.010213   0.003835   2.663  0.00906 **
```

Pr(Cancer) = -0.415 + 0.010(Age) + error

Pr(Cancer at age 50) = -0.415 + 0.0102*(40)

= -0.415 + 0.510

= 0.085

8.5% chance of cancer. Seems reasonable…

OR DOES IT?!

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.415413   0.230550  -1.802  0.07465 .
age          0.010213   0.003835   2.663  0.00906 **
```

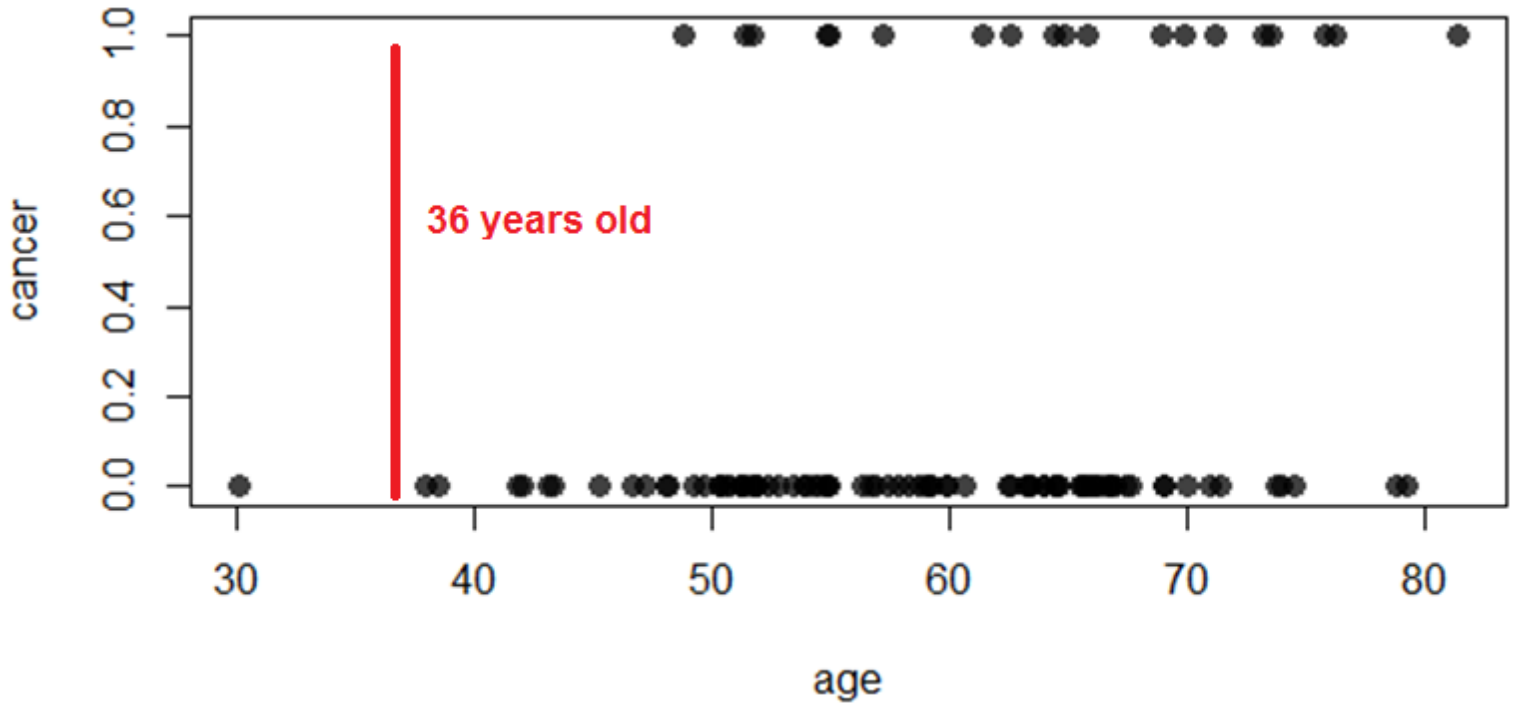
Pr(Cancer at age ***36***) = -0.415 + 0.0102*(***36***)

= -0.415 + 0.3672

= ***-0.0478***

36 year-olds are so health they have a

NEGATIVE 4.78% chance of having cancer. WOO!



Is this an extrapolation like the intercept?

No, 36 years old is within our dataset. It should produce a reasonable answer, but it doesn't.

The prediction intervals show additional problems.

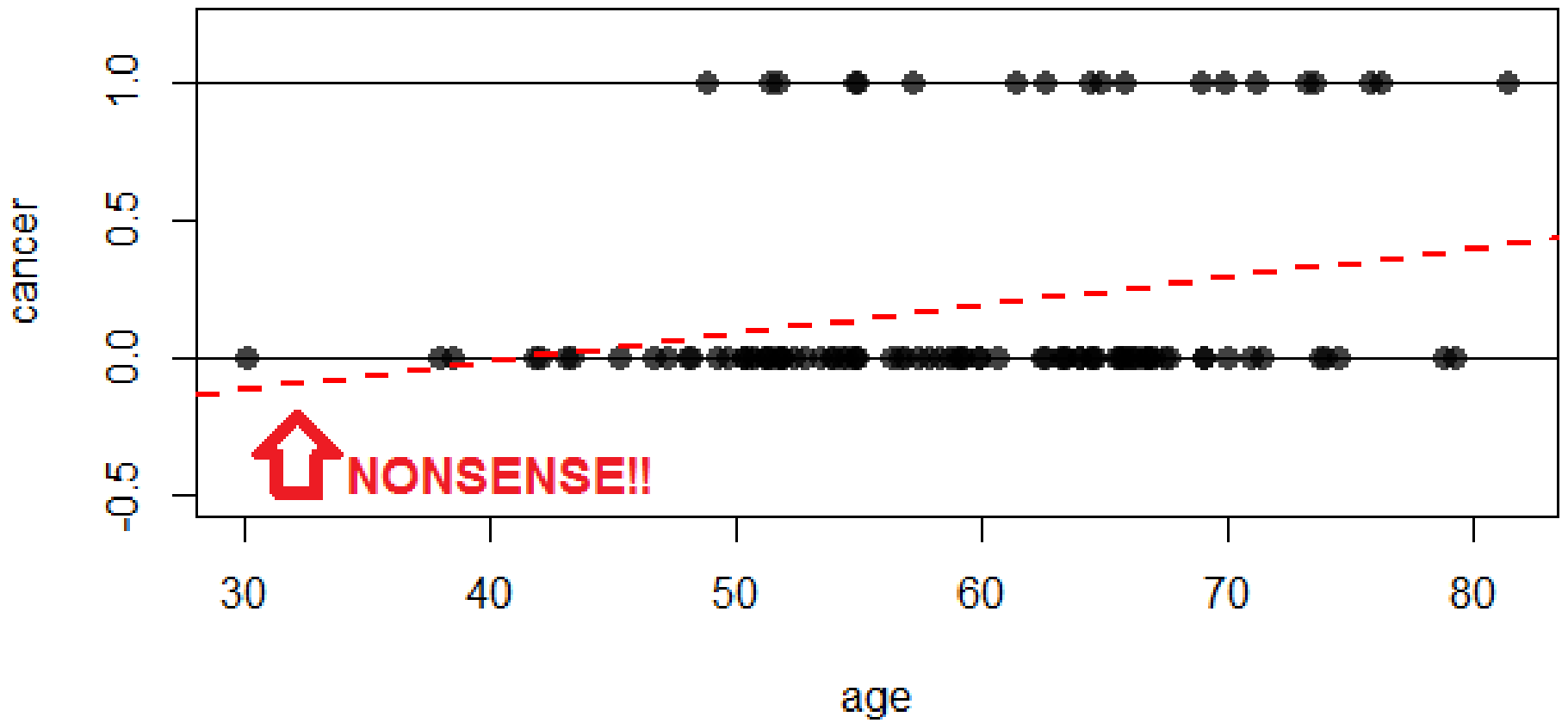
Pr(Cancer at age 66) = -0.415 + 0.0102*(66)

= 0.2586

95% Prediction Interval: -0.5062 to +1.0235

So we're 95% sure that a 66 year-old has between a NEGATIVE 50.62% chance and a 102.35% chance of having this cancer.

...really?



A line just isn't going to work when predicting the response to a binary variable.

So what's really the problem here?

Since probabilities can only take on values from 0 to 1, our model shouldn't be predicting anything outside of that range.

A line is always going to break out of that 0 to 1 range. In fact, it usually will break out somewhere within the range of the explanatory data.

Can we 'clamp' the fitted values to the [0,1] range?

Say you take all the negative values in this example and set them to zero. Are you 100% sure than nobody under 40 can get cancer? There are better options.

Can use a transform?

Yes, but we need a new transform called the logit transform. All the others we've seen (square, log, etc.) will break out of [0,1] as well.

Pay no attention to this method, it is wrong on purpose, as a joke.

Odds

The ‘odds’ of an event is the ratio of the probability of the event happening to it not happening.

Sometimes it’s written at as a ratio

Pr(A) : Pr(Not A), such as 3:2, 7:1, or 1:5.

These are read 3-to-2, 7-to-1, or 1-to-5 respectively.

In statistics, we tend to standardize the odds into a single value by making the second number 1.

So 3:2 becomes 1.50

This means that the event A happens 1.50 times for every time that it doesn't.

1:5 becomes 0.20

7:1 becomes 7

Or, event A is 7 times as likely to happen as it not happening.

We can translate probability into odds.

Example: Probability of A is 0.75

$Pr(A) = 0.75,$

$1-Pr(A) = 0.25$

$Odds = Pr(A) / (1 – Pr(A))$

$= 0.75 / 0.25 = 3$

Work-through example:

Probability of B is 0.125

Pr(B) = 0.125,

1-Pr(B) = 0.875

Work-through example:

Probability of B is 0.125

Pr(B) = 0.125,

1-Pr(B) = 0.875

Odds = Pr(B) / (1 – Pr(B))

Work-through example:

Probability of A is 0.125

Pr(B) = 0.125,

1-Pr(B) = 0.875

Odds = Pr(B) / (1 – Pr(B))

= 0.125 / 0.875 = 1/7 or 0.142

Probability can only range from 0 to 1,

Odds can be 0 up to any positive number.

Pr(A) = 0	translates to	Odds(A) = ***0***
Pr(B) = 0.1	translates to	Odds(B) = ***0.1111***
Pr(C) = 0.5	translates to	Odds(C) = ***1***
Pr(D) = 0.9999	translates to	Odds(D) = ***9999***
Pr(E) = 1	translates to	Odds(E) = ***Infinity***

We can translate odds into probability as well.

Odds of A = 1.5

$$Pr(A) = 1.5 * Pr(\text{not } A)$$

$$Pr(A) = 1.5 * (1 – Pr(A))$$

$$Pr(A) = 1.5 – 1.5* Pr(A)$$

$$2.5* Pr(A) = 1.5$$

$$Pr(A) = 1.5 / 2.5 = 0.6$$

Work-Through Example

Odds of B = 5

$$Pr(B) = 5 * Pr(\text{not } B)$$

$$Pr(B) = 5 * (1 - Pr(B))$$

Work-Through Example

Odds of B = 5

$$Pr(B) = 5 * Pr(\text{not } B)$$

$$Pr(B) = 5 * (1 - Pr(B))$$

$$Pr(B) = 5 - 5 * Pr(B)$$

$$6 * Pr(B) = 5$$

Work-Through Example

Odds of B = 5

$$Pr(B) = 5 * Pr(\text{not } B)$$

$$Pr(B) = 5 * (1 - Pr(B))$$

$$Pr(B) = 5 - 5 * Pr(B)$$

$$6 * Pr(B) = 5$$

$$Pr(B) = 5 / 6 = 0.83333$$

A lot of stats is just making sure your variables make sense.

Log-Odds

Another useful measure is log-odds, which is exactly what it sounds like:

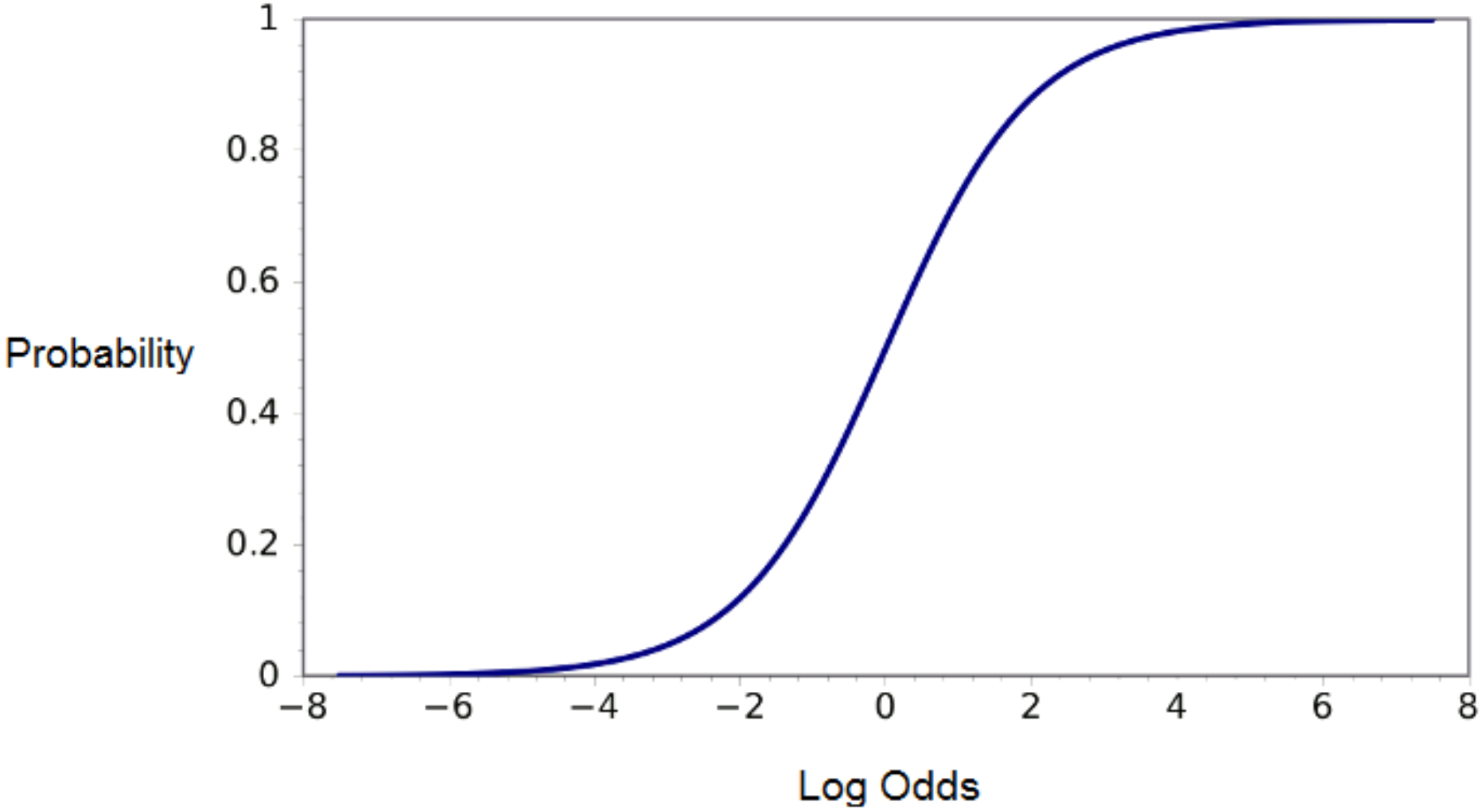
The log-odds is the natural ***logarithm*** of the odds.

…so what?

A few example values

Probability	Odds	Log-Odds
0	0	- Infinity
0.1	0.1111	-2.197
0.25	0.3333	-1.099
0.5	1	0
0.66667	2	0.6931
0.8	5	1.6094
0.9999	9999	9.2102
1	+ Infinity	+ Infinity

The sigmoid curve (name just for interest) shows the whole relationship between log odds and probability.



Probability without any changes can only be described as a value between 0 and 1 inclusive.

Odds, can be described as any value between 0 and positive infinity.

Log-odds, can be described as any real number. That includes every value from negative infinity to positive infinity.

…yeah, but so what?

Why do we care about log-odds and that it can take any number?

Because of regression.

Regression handles many different situations, but it needs the responses to be continuous values.

If we were do a regression to predict how likely something is, we don't want negative probabilities, or ones that are higher than 100%.

Logit Function

The logit function is a function that transforms a probability into a log-odds. It does both steps – turning a probability into an odds, and taking the logarithm, together.

$$logit(x) = \log\left(\frac{x}{1-x}\right)$$

Inverse Logit Function

The inverse logit function does the opposite. It converts a value from a log-odds into a probability.

$$logit^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{e^x}{1+e^x}$$

We will need this function to make sense of some logistic regression results.

Don’t let odd logs trip you up. Take a break for hour 2