CMPINF 2100
Introduction to Data Centric Computing

Week 09

Introduction to Regression

Linear Model Assumptions

At the start of the semester we introduced the goals of **predictive modeling** or ***supervised learning***

- We discussed the two main areas: regression and classification
- We discussed at high level the goals and steps involved.
- Please review the “supervised learning workflow” from week 01.

This lecture introduces regression with linear models!

- Consider a simple case of a response, y, and a single input variable, x.
- We observe the input-output pair, $\{x_n, y_n\}$, N times.
- We wish to **train** or ***fit*** a linear model, relating the response to the input.

What are the assumptions of the linear model?

To “see” the assumptions, let’s start out with the math

- A **linear model**, assuming a ***linear relationship*** between the response and input for the n-th observation is typically written as:

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$

$$\epsilon_n \sim \text{normal}(0, \sigma)$$

To “see” the assumptions, let’s start out with the math

- A **<u>linear model</u>**, assuming a ***linear relationship*** between the response and input for the n-th observation is typically written as:

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$

The model has 2 COEFFICIENTS.

The INTERCEPT, β_0, and the SLOPE, β_1.

$$\epsilon_n \sim \text{normal}(0, \sigma)$$

To “see” the assumptions, let’s start out with the math

- A **<u>linear model</u>**, assuming a ***linear relationship*** between the response and input for the n-th observation is typically written as:

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n$$

The errors, ϵ_n, are NORMALLY (Gaussian or Bell Curve) distributed with mean zero and standard deviation σ.

$$\epsilon_n \sim \text{normal}(0, \sigma)$$

However, I prefer a different notation…

- The **LIKELIHOOD** associated with the n-th observation:

$$y_n \mid x_n, \beta_0, \beta_1, \sigma \sim \text{normal}(y_n \mid \beta_0 + \beta_1 x_n, \sigma)$$

However, I prefer a different notation…

- The **LIKELIHOOD** associated with the n-th observation:

$$y_n \mid x_n, \beta_0, \beta_1, \sigma \sim \text{normal}(y_n \mid \beta_0 + \beta_1 x_n, \sigma)$$

The vertical bar, |, means "**GIVEN**".

The LEFT HAND side above reads as:

"y **GIVEN** x, beta 0, beta 1, and sigma"

GIVEN means we KNOW the values to the RIGHT side of the vertical bar.

However, I prefer a different notation…

- The **LIKELIHOOD** associated with the n-th observation:

$$y_n \mid x_n, \beta_0, \beta_1, \sigma \sim \text{normal}(y_n \mid \beta_0 + \beta_1 x_n, \sigma)$$

The ~ character is the tilde character (left of 1 on most keyboards).

The ~ character here means “as distributed as”.

The notation in MATH reads as: “the values on the LEFT are DISTRIBUTED AS the distribution on the right”.

The variable y is therefore NORMALLY distributed GIVEN x, beta 0, beta 1, and sigma

However, I prefer a different notation…

- The **LIKELIHOOD** associated with the n-th observation:

$$y_n \mid x_n, \beta_0, \beta_1, \sigma \sim \text{normal}(y_n \mid \beta_0 + \beta_1 x_n, \sigma)$$

- The observed response is a random variable!
  - The observations are **stochastic**.

However, I prefer a different notation…

- The **LIKELIHOOD** associated with the n-th observation:

$$y_n \mid x_n, \beta_0, \beta_1, \sigma \sim \text{normal}(y_n \mid \beta_0 + \beta_1 x_n, \sigma)$$

- The observed response is a random variable!
  - The observations are **stochastic**.
- The response is Normally distributed. What's the average value?

However, I prefer a different notation…

- The **LIKELIHOOD** associated with the n-th observation:

$$y_n \mid x_n, \beta_0, \beta_1, \sigma \sim \text{normal}(y_n \mid \beta_0 + \beta_1 x_n, \sigma)$$

- The observed response is a random variable!
  - The observations are **stochastic**.
- The response is Normally distributed. What's the average value?

$$\mu_n = \beta_0 + \beta_1 x_n$$

The mean of the response changes with the input! Setting β_0, β_1, and x_n allows us to calculate μ_n. The mean is **deterministic** given the **coefficients** and **input**.

Our simple linear model can therefore be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

$$\mu_n = \beta_0 + \beta_1 x_n$$

Linear regression is therefore modeling the MEAN or AVERAGE behavior of the response!

Our simple linear model can therefore be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

$$\mu_n = \beta_0 + \beta_1 x_n$$

What is the variability of the response around μ_n?

Our simple linear model can therefore be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

$$\mu_n = \beta_0 + \beta_1 x_n$$

What is the variability of the response around μ_n?

σ

The response's standard deviation around the MEAN TREND is σ.

Our simple linear model can therefore be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

$$\mu_n = \beta_0 + \beta_1 x_n$$

What is the variability of the response around μ_n?

σ

The response's standard deviation around the MEAN TREND is σ.

Does the variability around the mean change with x?

Our simple linear model can therefore be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

$$\mu_n = \beta_0 + \beta_1 x_n$$

What is the variability of the response around μ_n?

σ

The response's standard deviation around the MEAN TREND is σ.

Does the variability around the mean change with x?

NO!!! Constant variance!

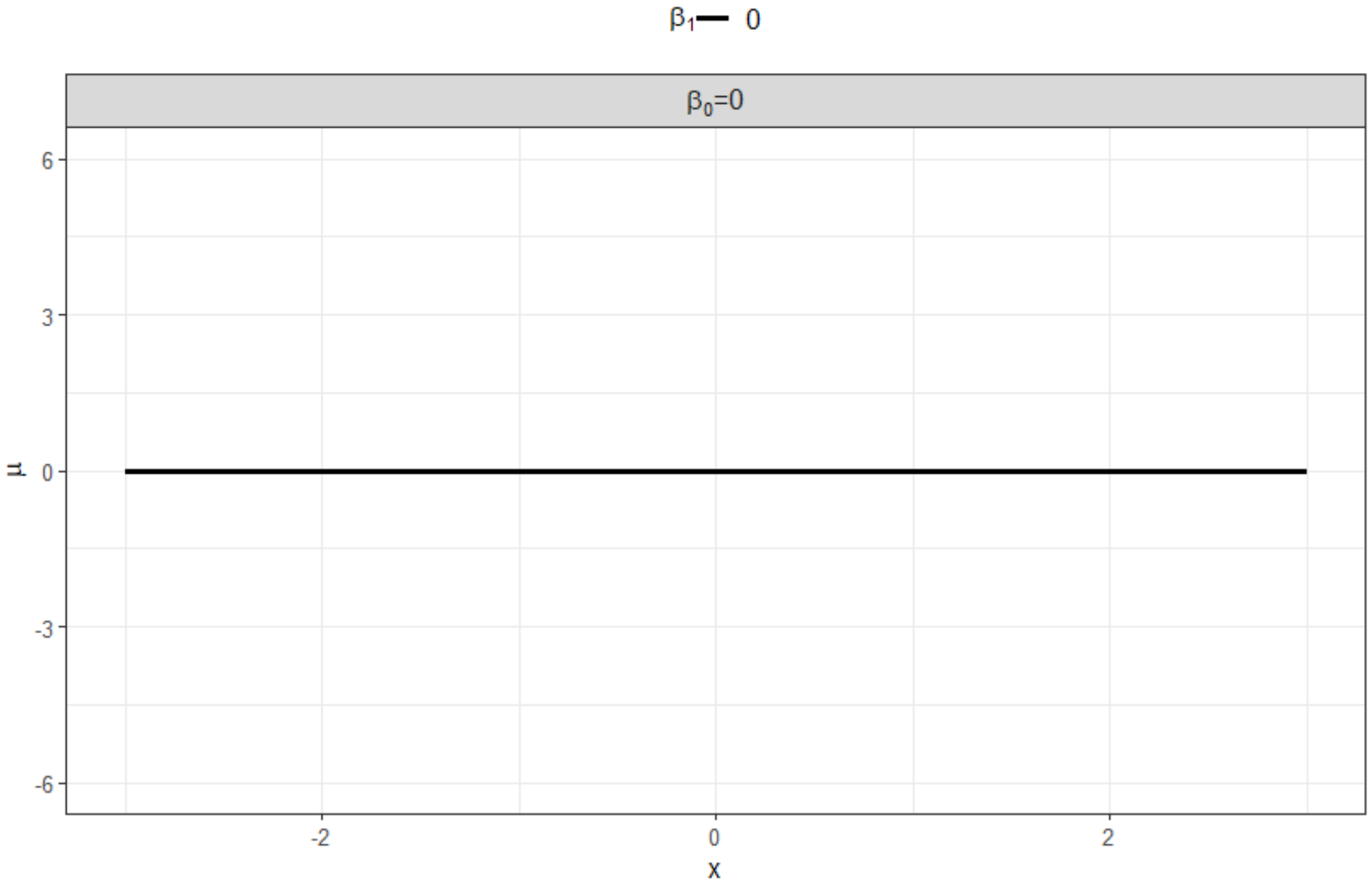
Remember there are N observations

- We have focused on an individual observation's **LIKELIHOOD**.
- We assumed that each observation is **CONDITIONALLY INDEPENDENT** given the input and parameters.
- This means one observation does NOT depend on or is NOT related to any other observation AS LONG AS we KNOW the input and the parameters!!!!!
- The random observations are UNCORRELATED if we know the input parameters.

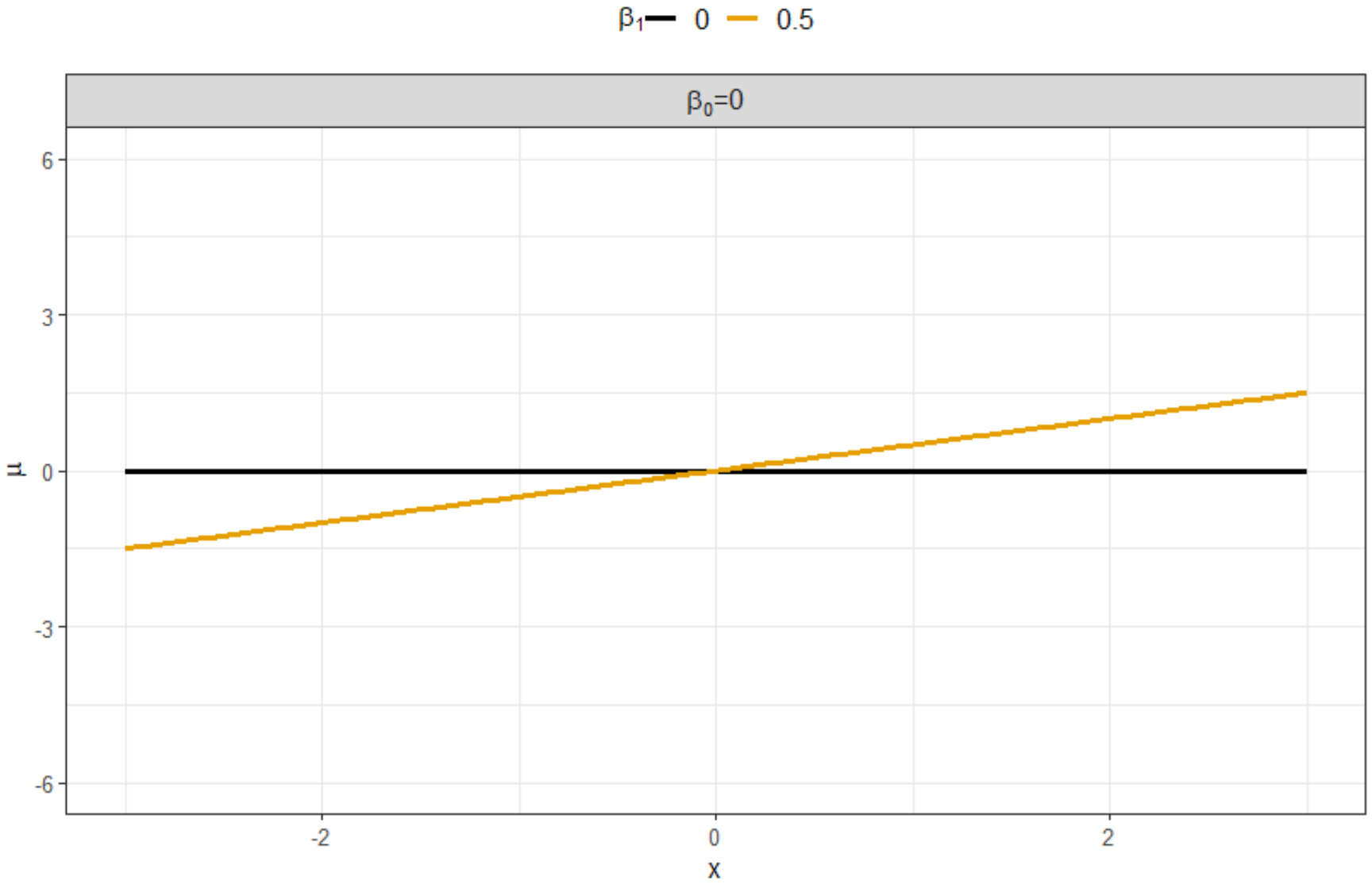
Trends: start with the mean, μ, vs the input x

- β_0 is the intercept and β_1 is the slope.
- Consider an input, x, between -2 and +2.
- What happens if $\beta_0 = 0$ and $\beta_1 = 0$?

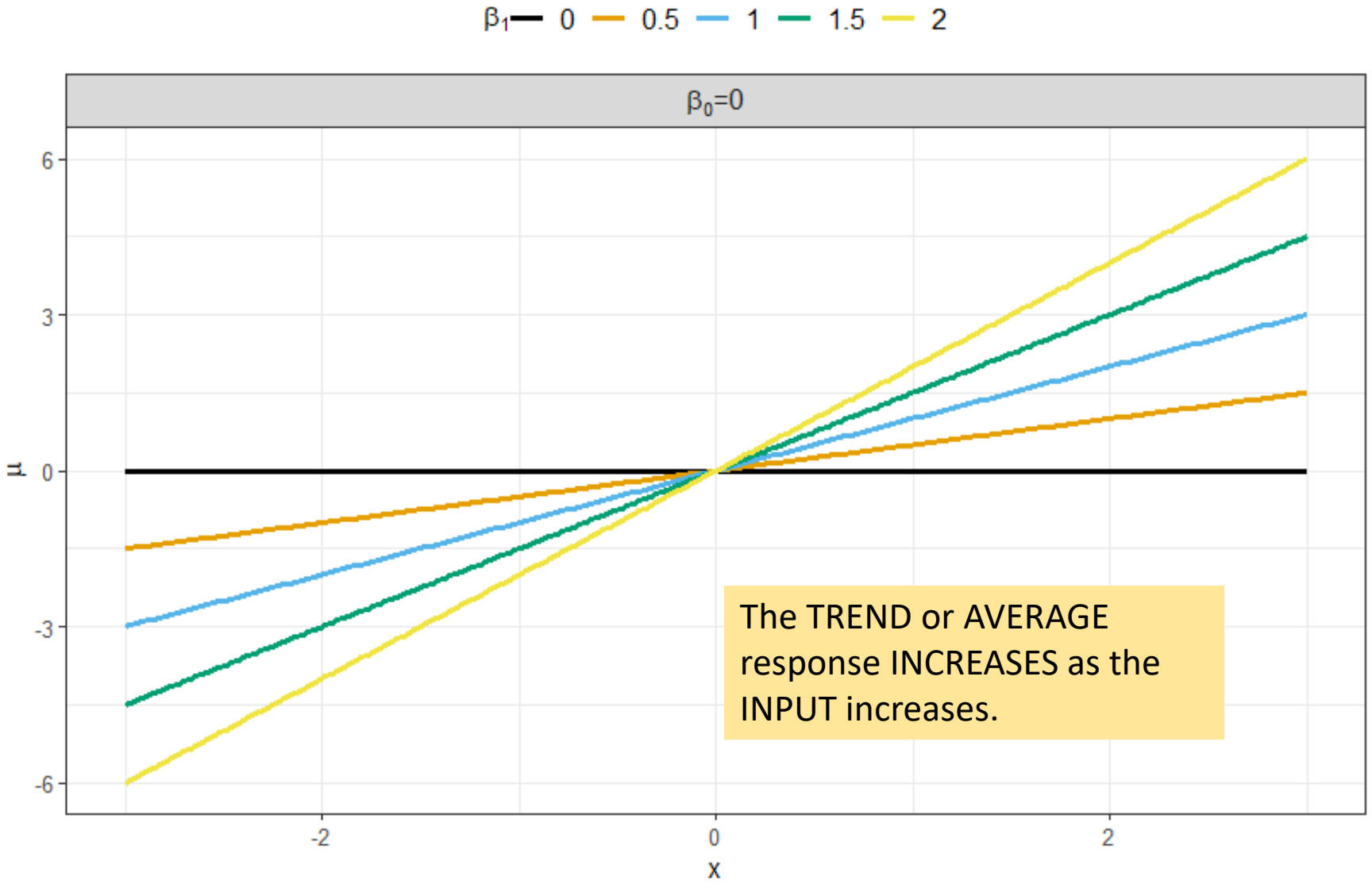
There is no trend between the mean and the input if the SLOPE is ZERO!!!



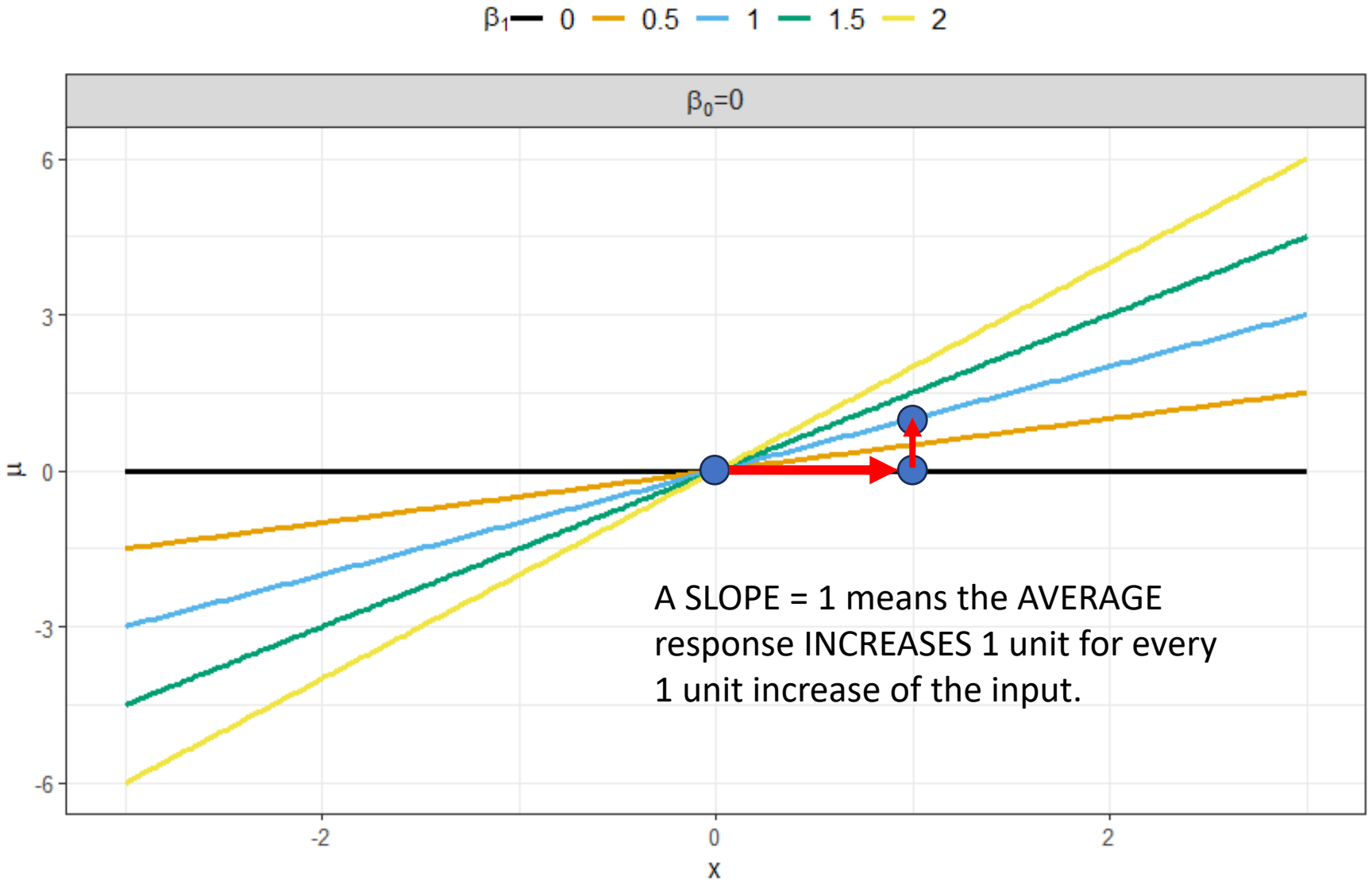
As the slope increases…



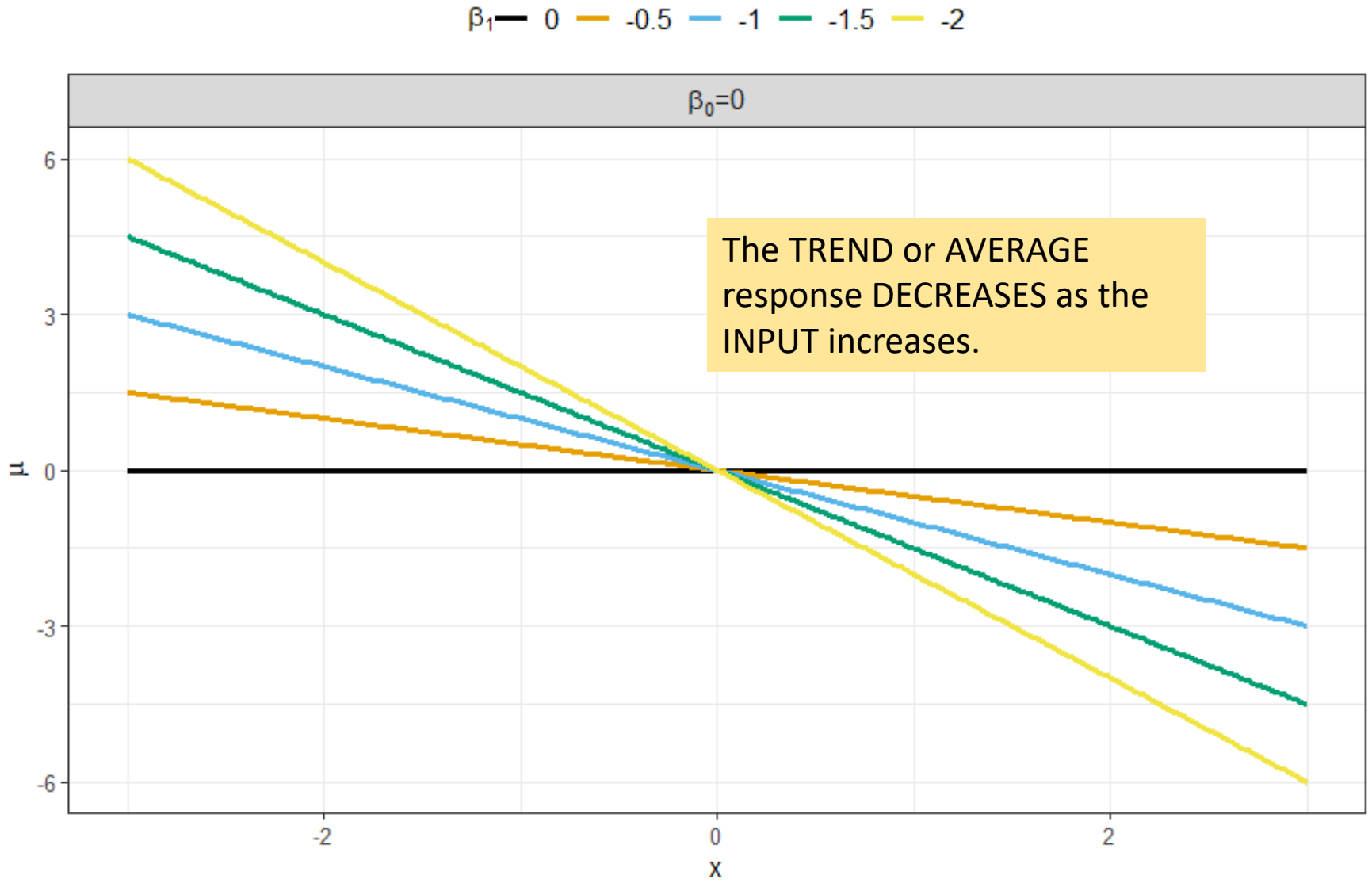
As the slope increases…



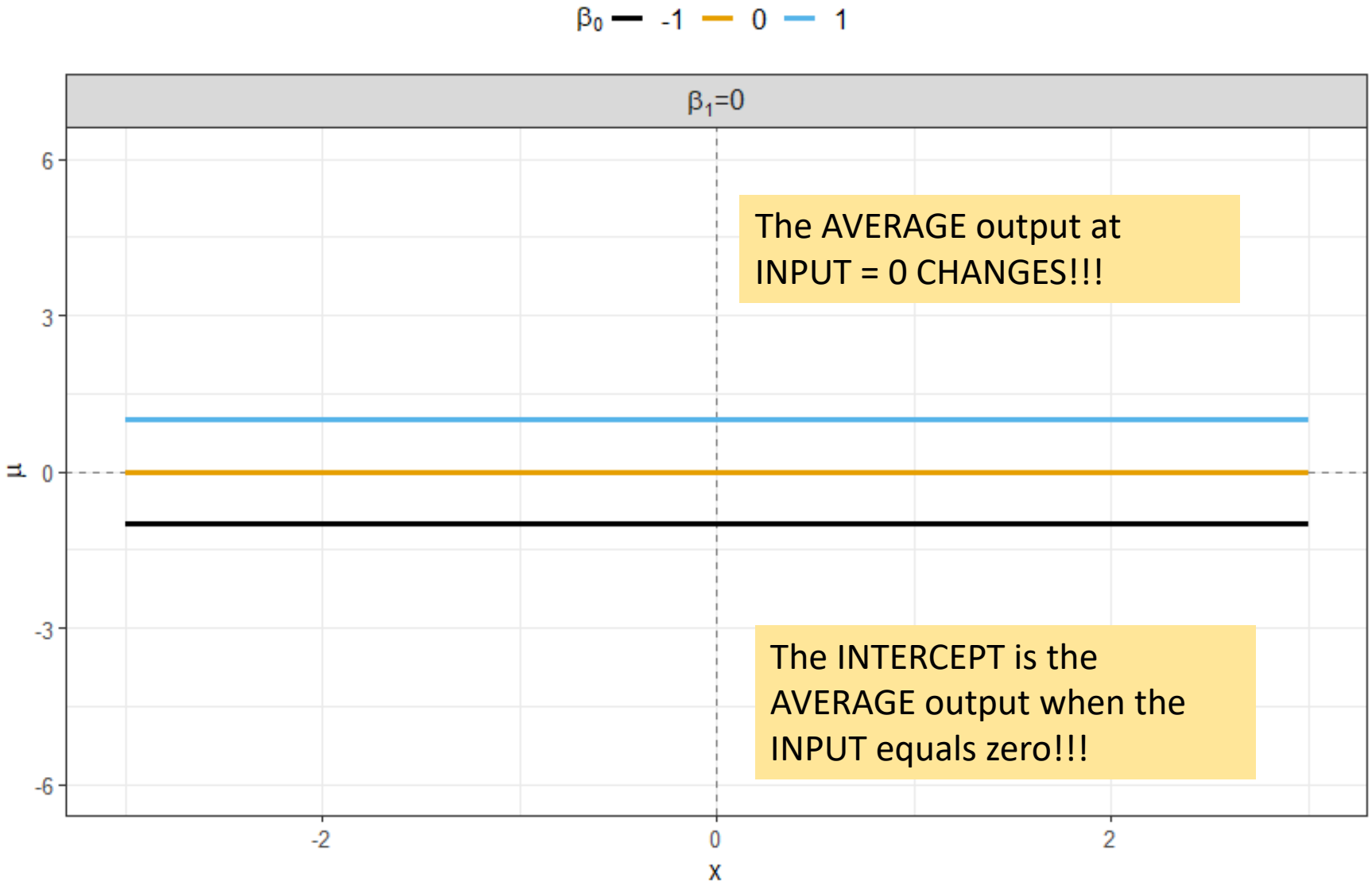
For example, focus on SLOPE = 1 (blue line)



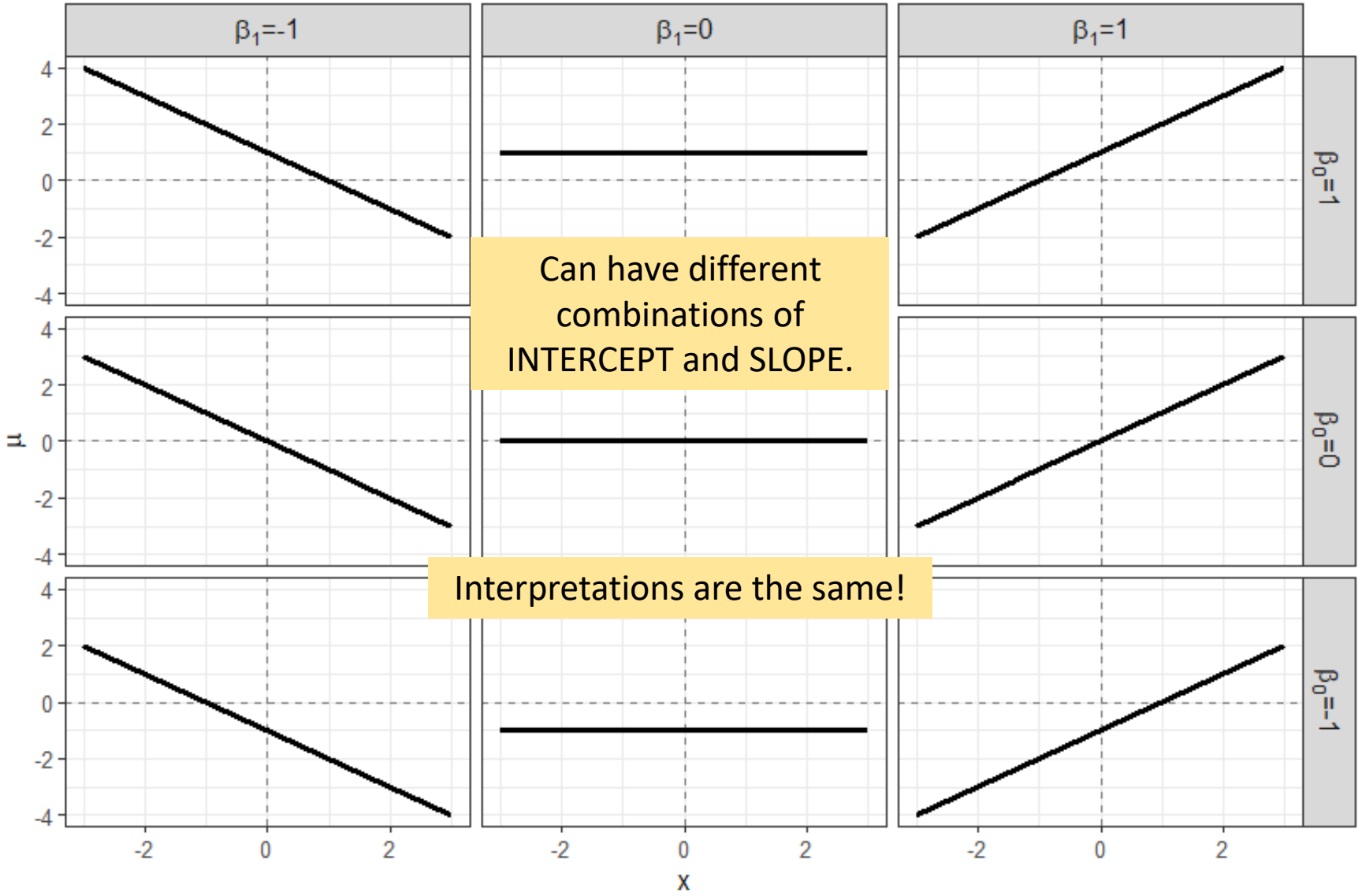
Negative slopes



Changing the intercept…



Changing the intercept and slope together

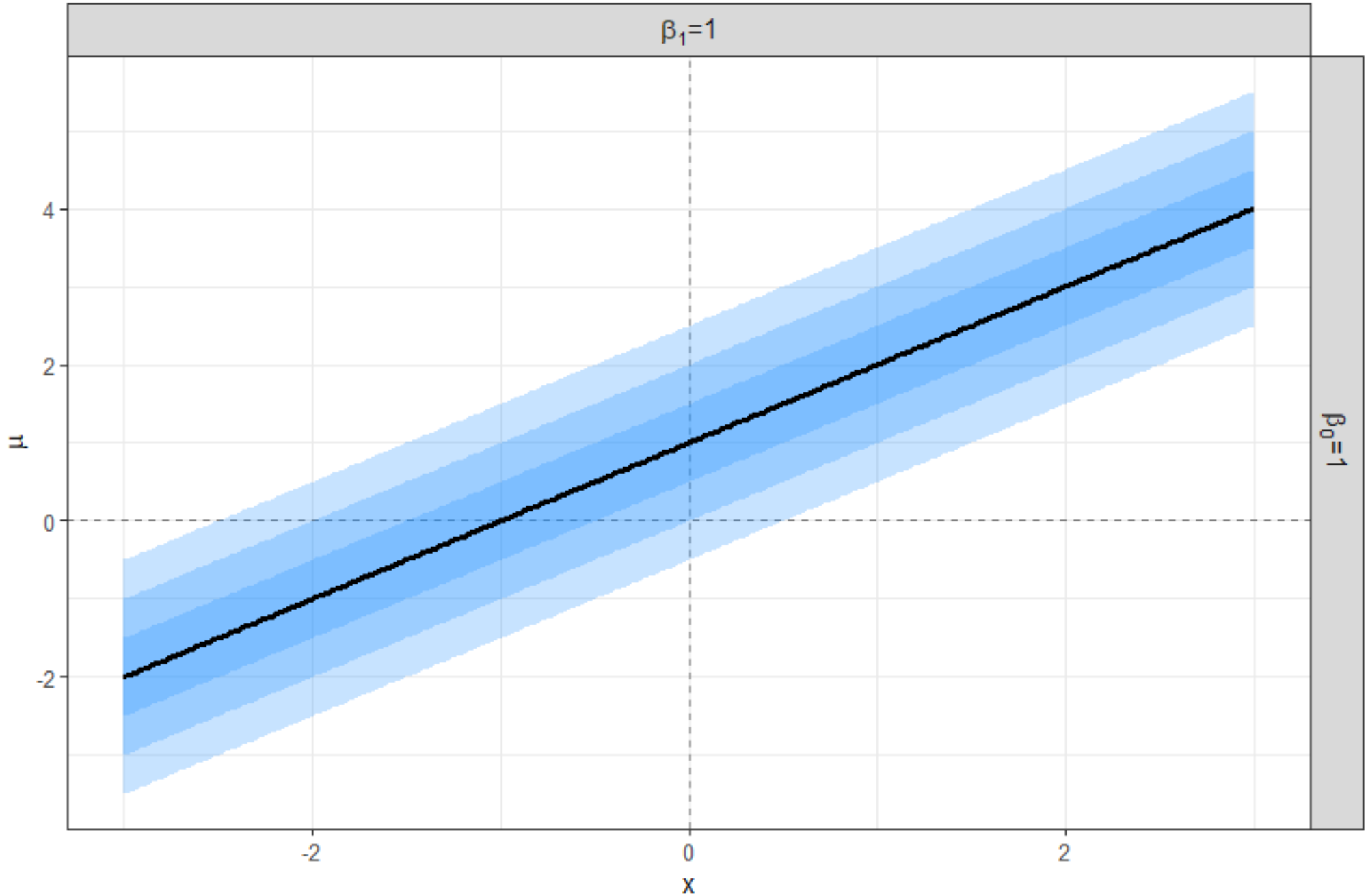


Can have different combinations of INTERCEPT and SLOPE.

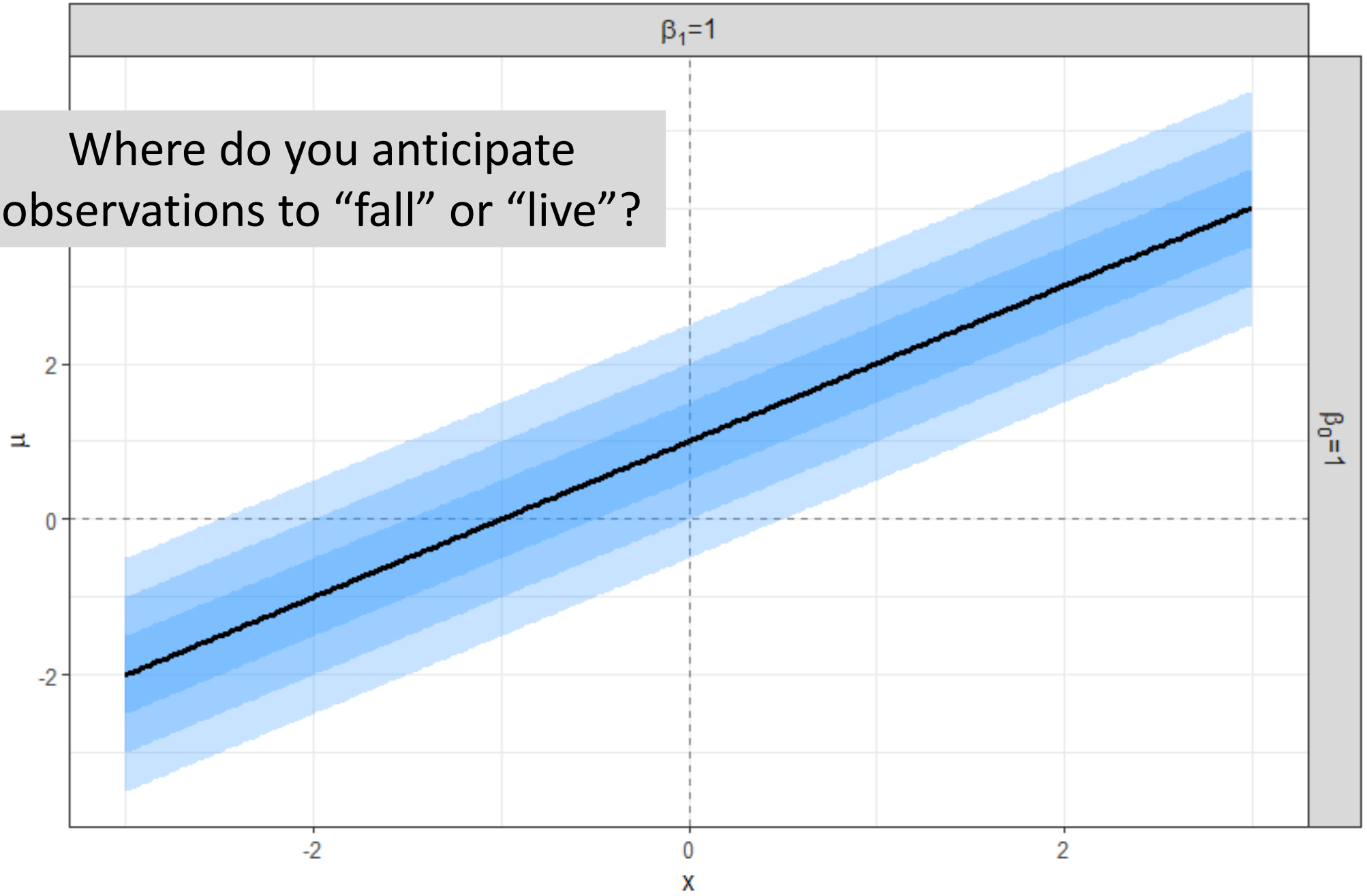
Interpretations are the same!

What about the random component?

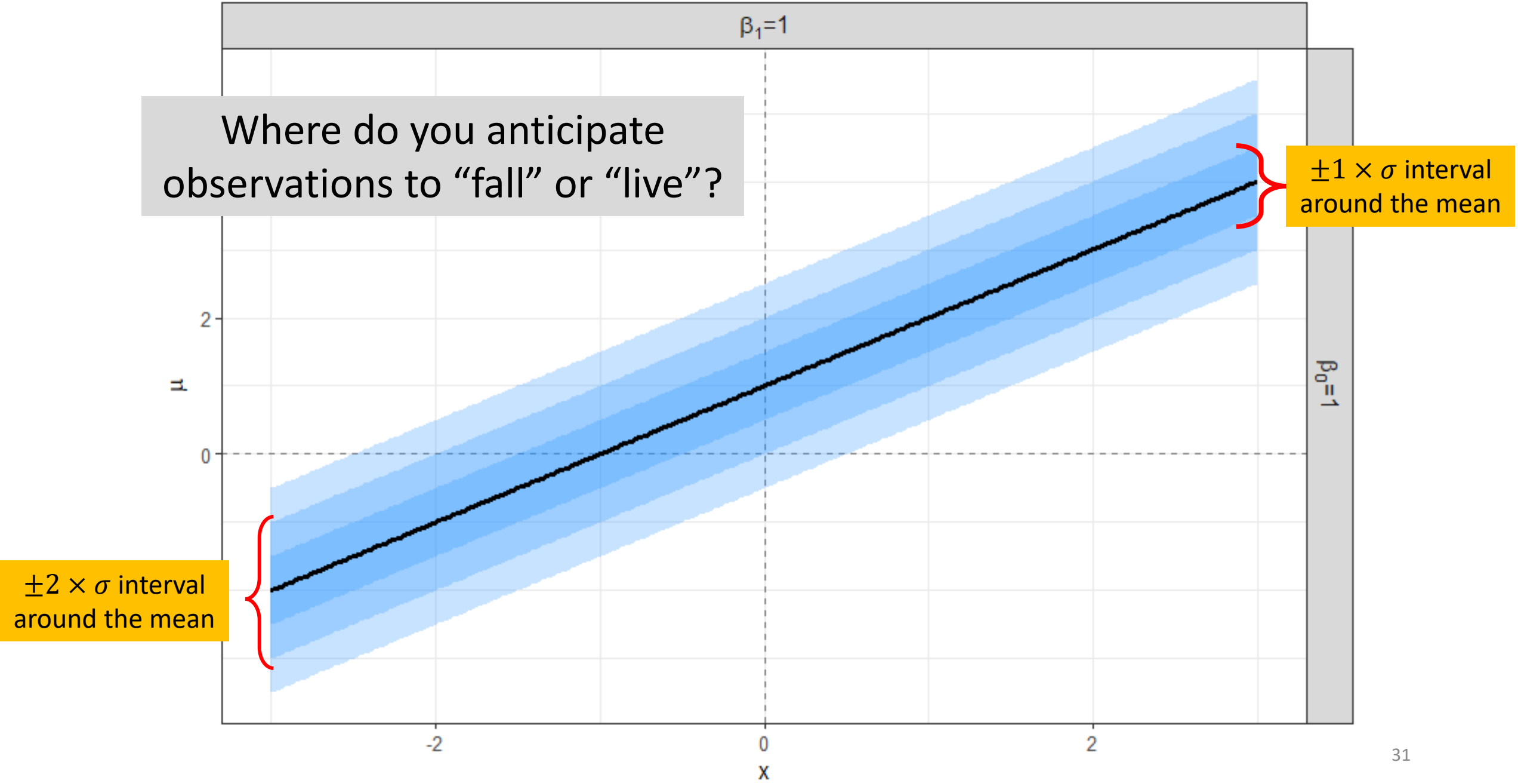
- Observations are normally distributed around the mean with standard deviation σ.
- We will assume that $\sigma = 0.5$ for visualization purposes.
- The variability or **uncertainty** will be represented by overlapping ribbons to capture the $\pm 1 \times \sigma$, $\pm 2 \times \sigma$, and $\pm 3 \times \sigma$ intervals.


β_1=1
β_0=1
μ
x
4
2
0
-2
-2
0
2

Where do you anticipate observations to “fall” or “live”?



Where do you anticipate observations to “fall” or “live”?



$\pm 1 \times \sigma$ interval around the mean

$\pm 2 \times \sigma$ interval around the mean

Let’s generate random draws!

- Specify β_0 and β_1.
  - Specifically we will set both equal to 1.
- Specify σ.
  - Specifically we sill set equal to 0.5.

Let's generate random draws!

- Specify β_0 and β_1.
  - Specifically we will set both equal to 1.
- Specify σ.
  - Specifically we sill set equal to 0.5.
- Specify $\mathbf{x} = [x_1, \ldots, x_N]^T$.
  - We will use 9 evenly spaced points between -2 and 2.

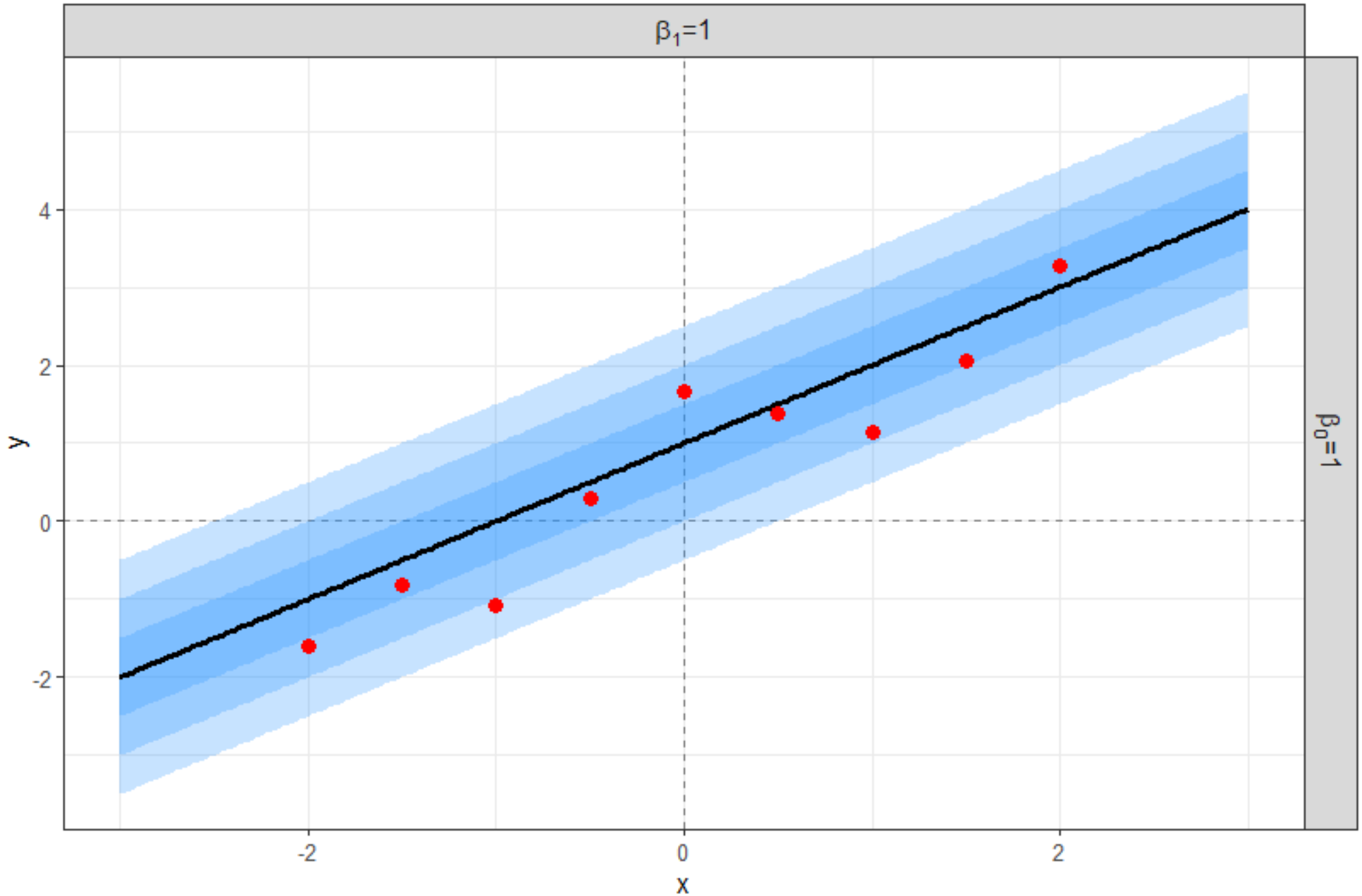
Let's generate random draws!

- Specify β_0 and β_1.
  - Specifically we will set both equal to 1.
- Specify σ.
  - Specifically we sill set equal to 0.5.
- Specify $\mathbf{x} = [x_1, \ldots, x_N]^T$.
  - We will use 9 evenly spaced points between -2 and 2.
- Calculate the mean, μ_n, associated with each x_n GIVEN $\boldsymbol{\beta}$.

$$\mu_n = \beta_0 + \beta_1 x_n$$

- Generate a normal random number:

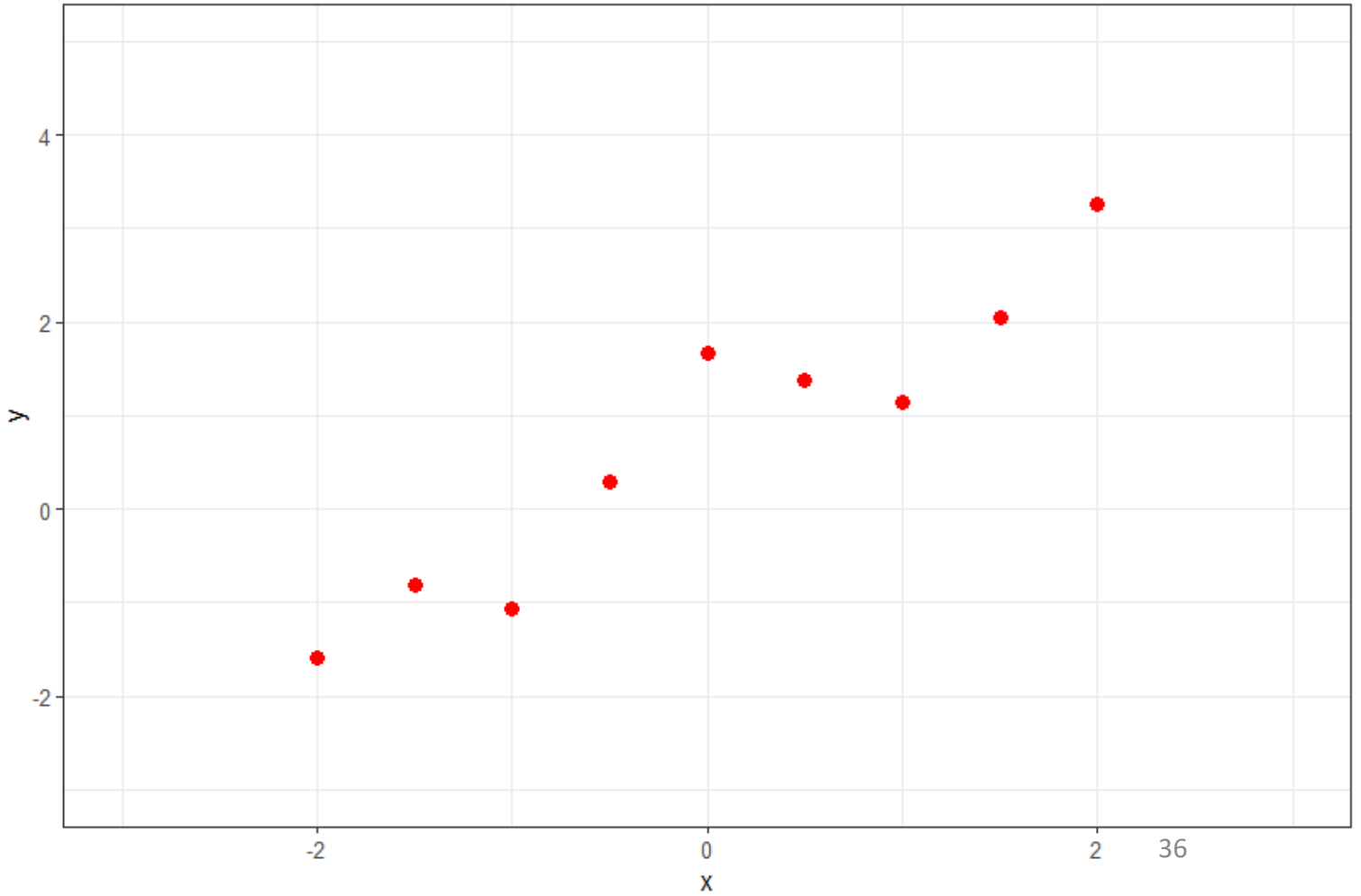
$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$


β_1=1
β_0=1
y
x
4
2
0
-2
-2
0
2

We know the observations are generated from a linear relationship between the response and the input…

However, if we only had random observations…

Would we be tempted to think the relationship is non-linear…

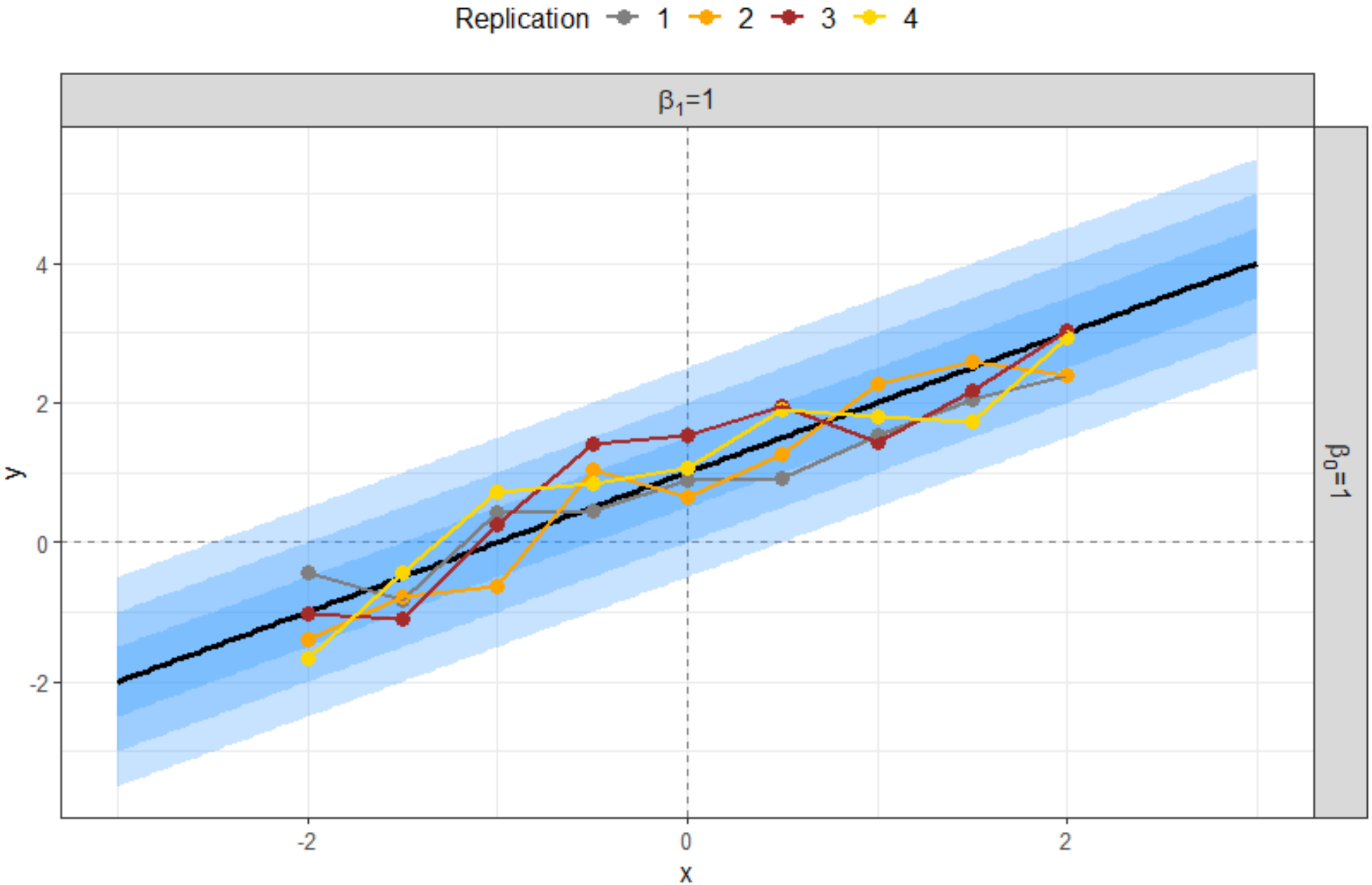


What happens if we increase the sample size at each input location?

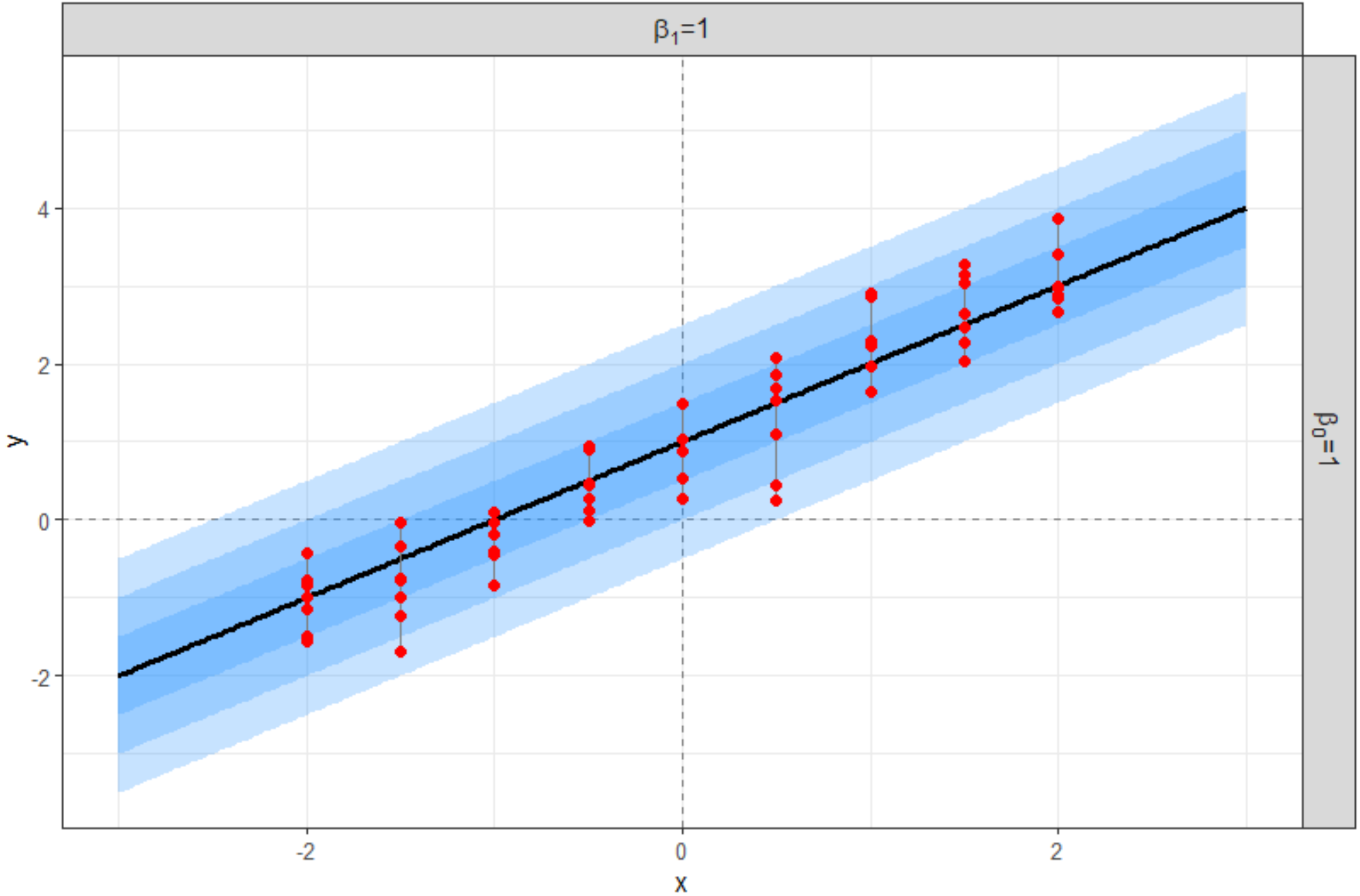
- Rather than generating one random value per input location, generate 4 repeats or replications per x value.

- The next slide connects the observations associated with each replication just to help the visualization.

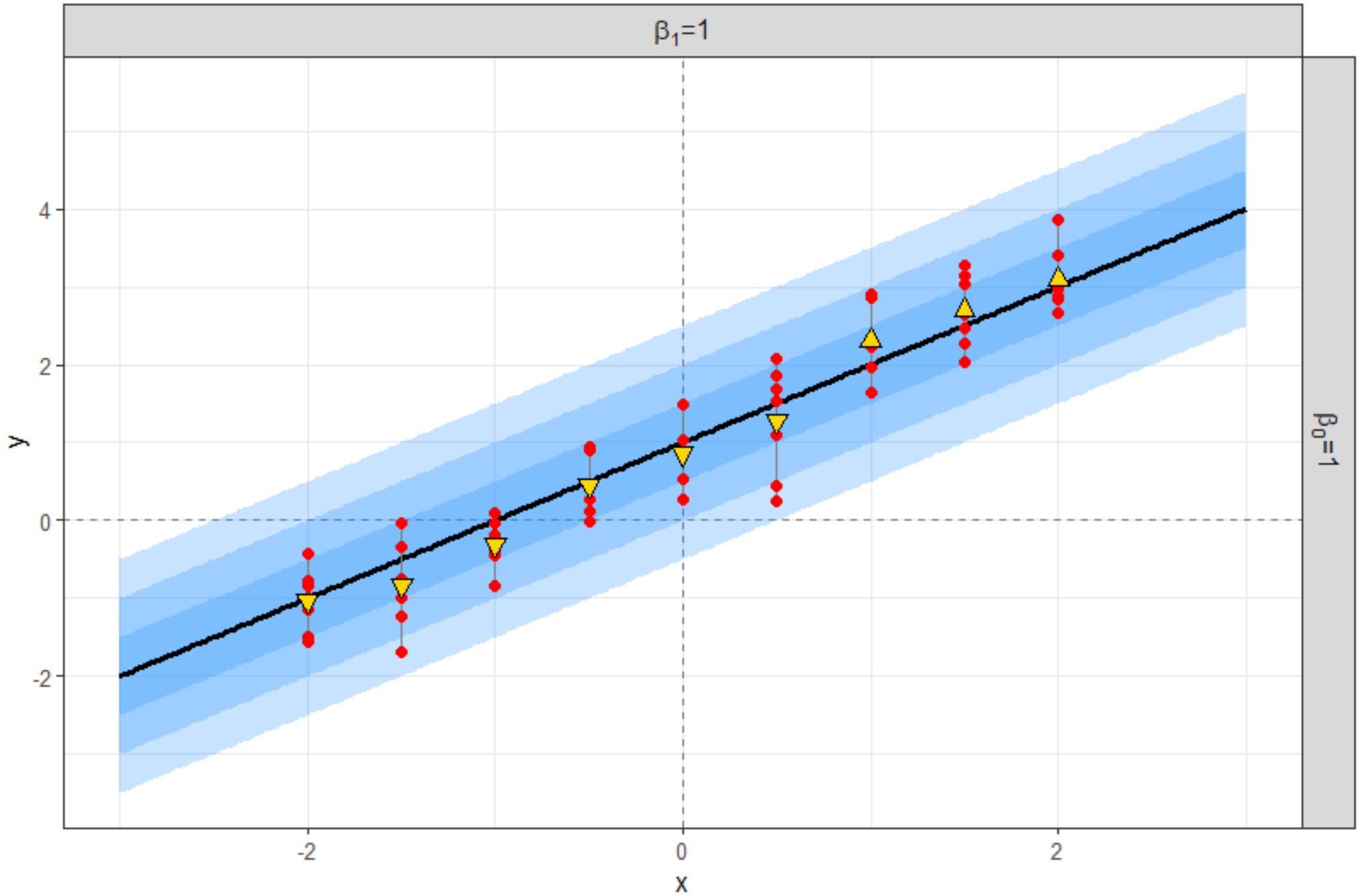
- Remember, we generated random draws based on a **deterministic linear relationship** between the mean response and input!


Replication 1 2 3 4
β_1=1
β_0=1
y
x
4
2
0
-2
-2
0
2

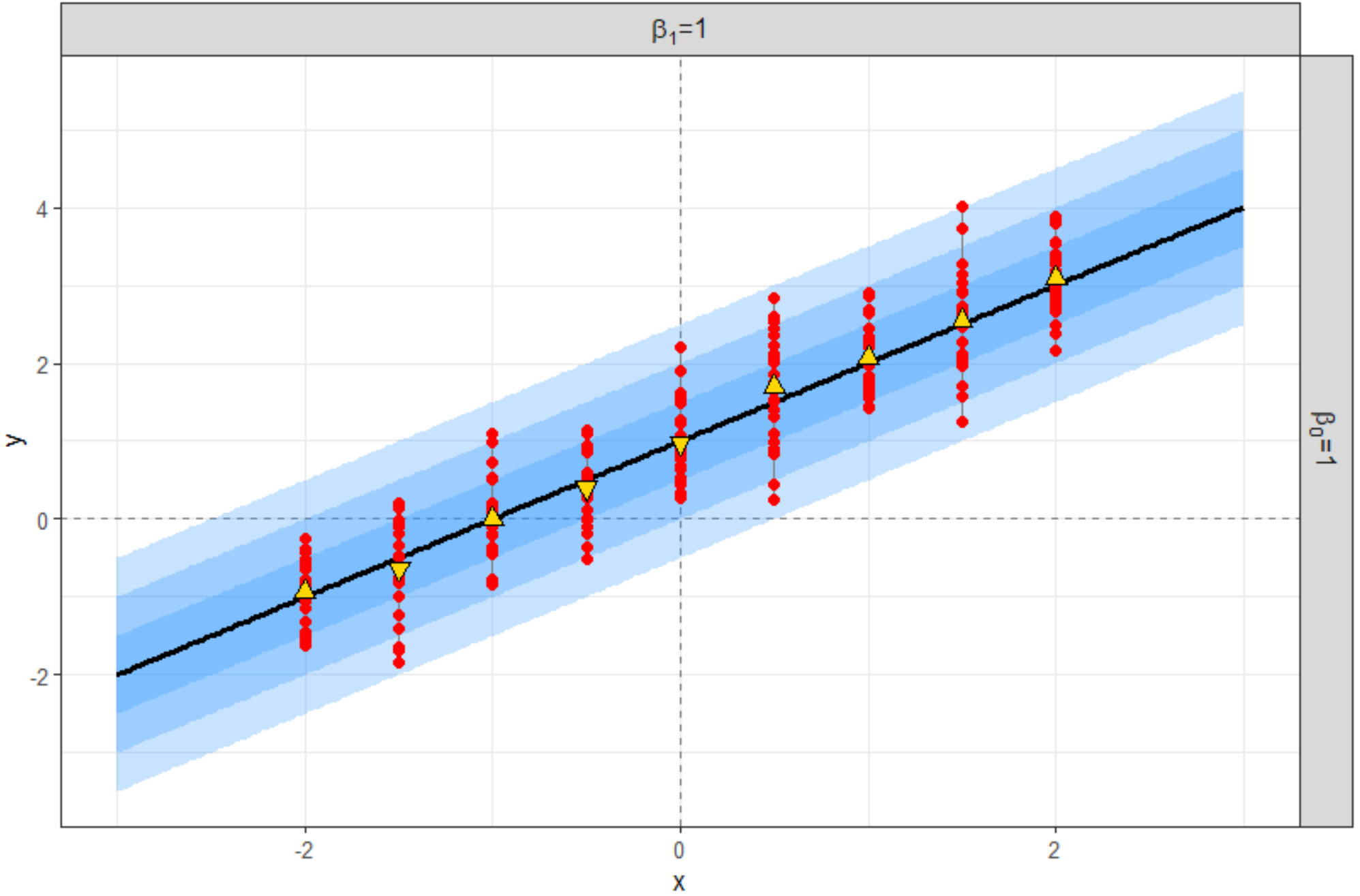
Increase the sample size from 4 to 7 replications per input location



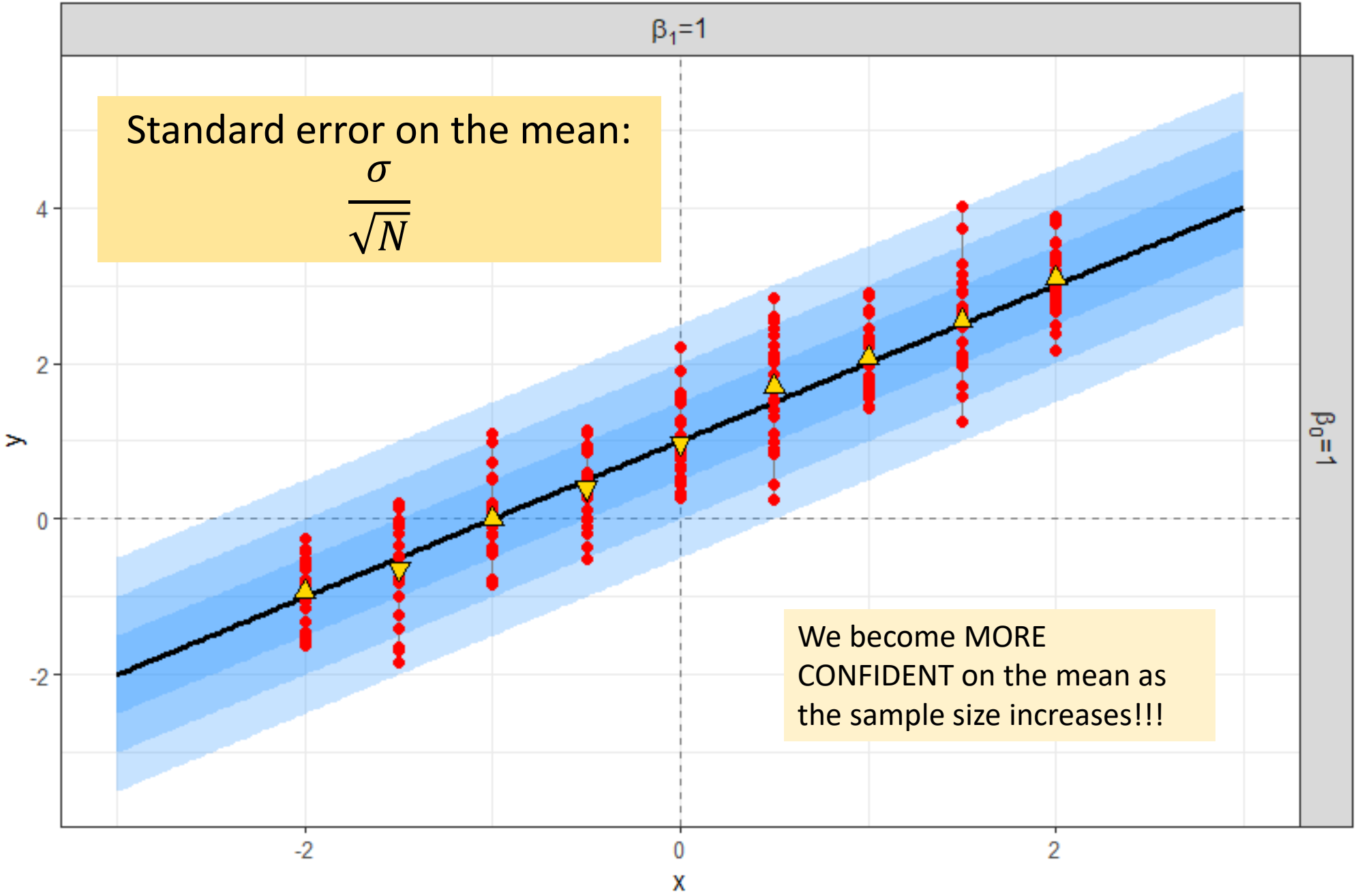
Include the response sample average at each input location



25 observations per input location…what happened to the sample averages?



Remember the CENTRAL LIMIT THEOREM!

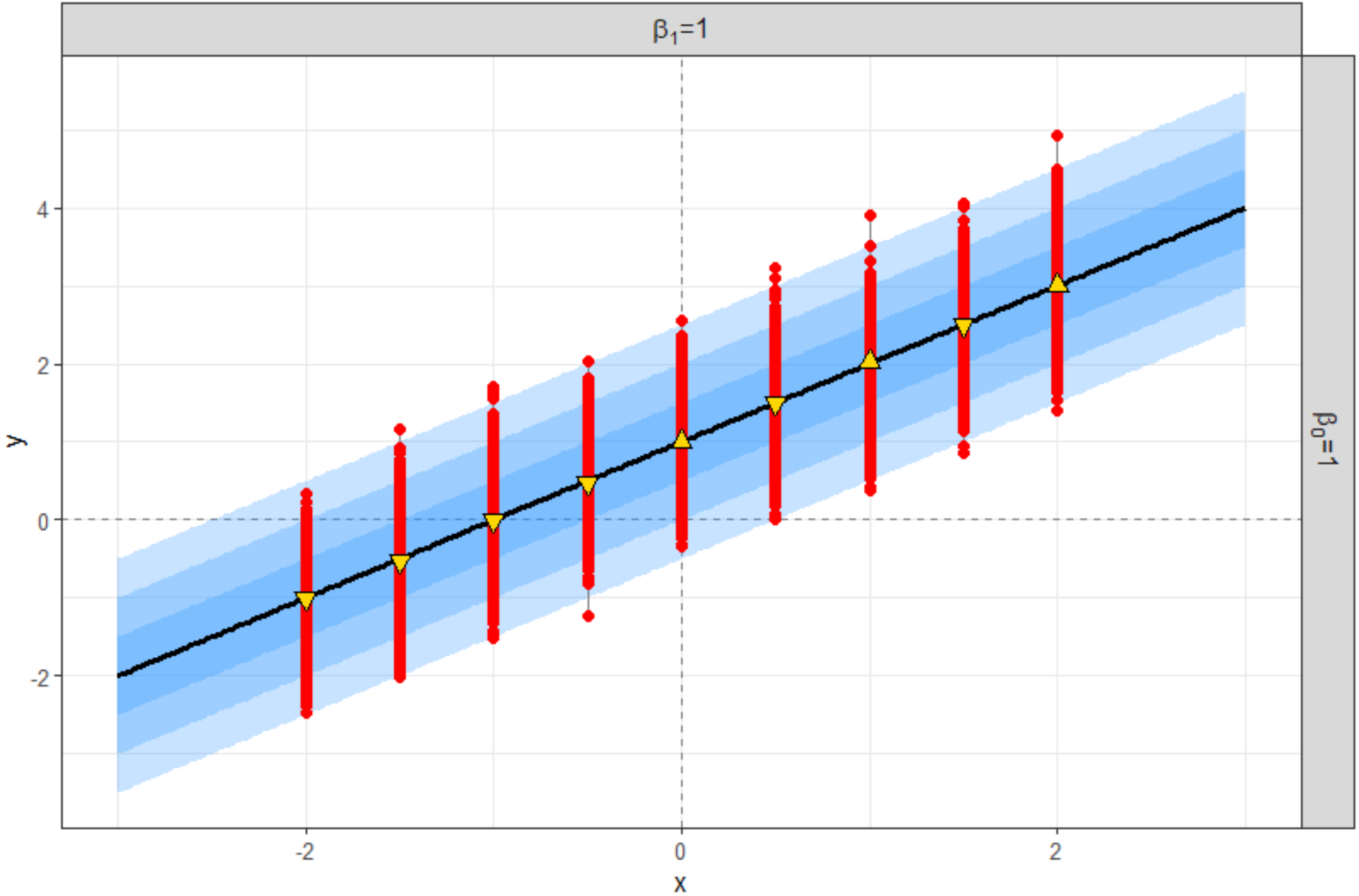


Standard error on the mean:

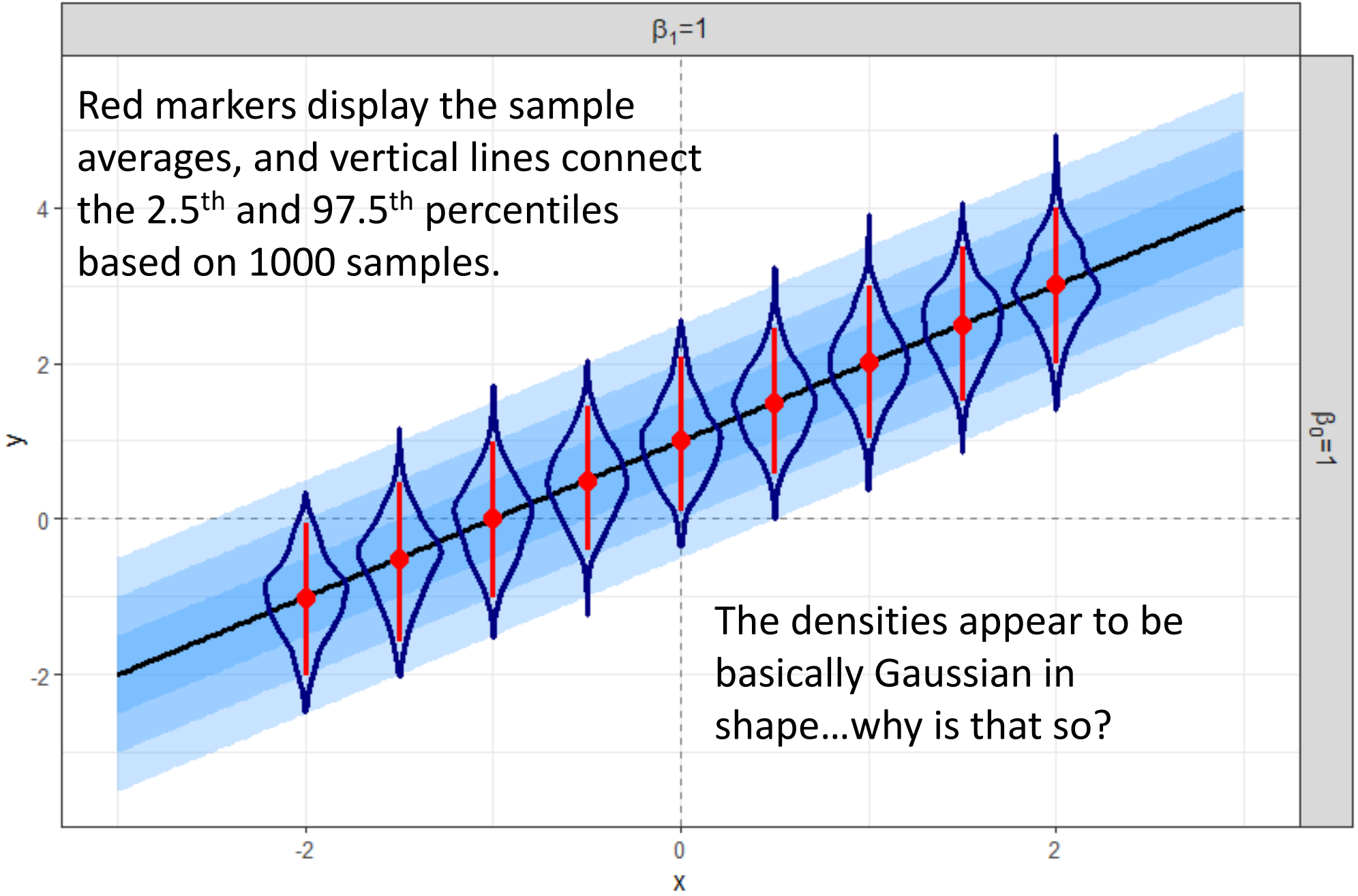
$$\frac{\sigma}{\sqrt{N}}$$

We become MORE CONFIDENT on the mean as the sample size increases!!!

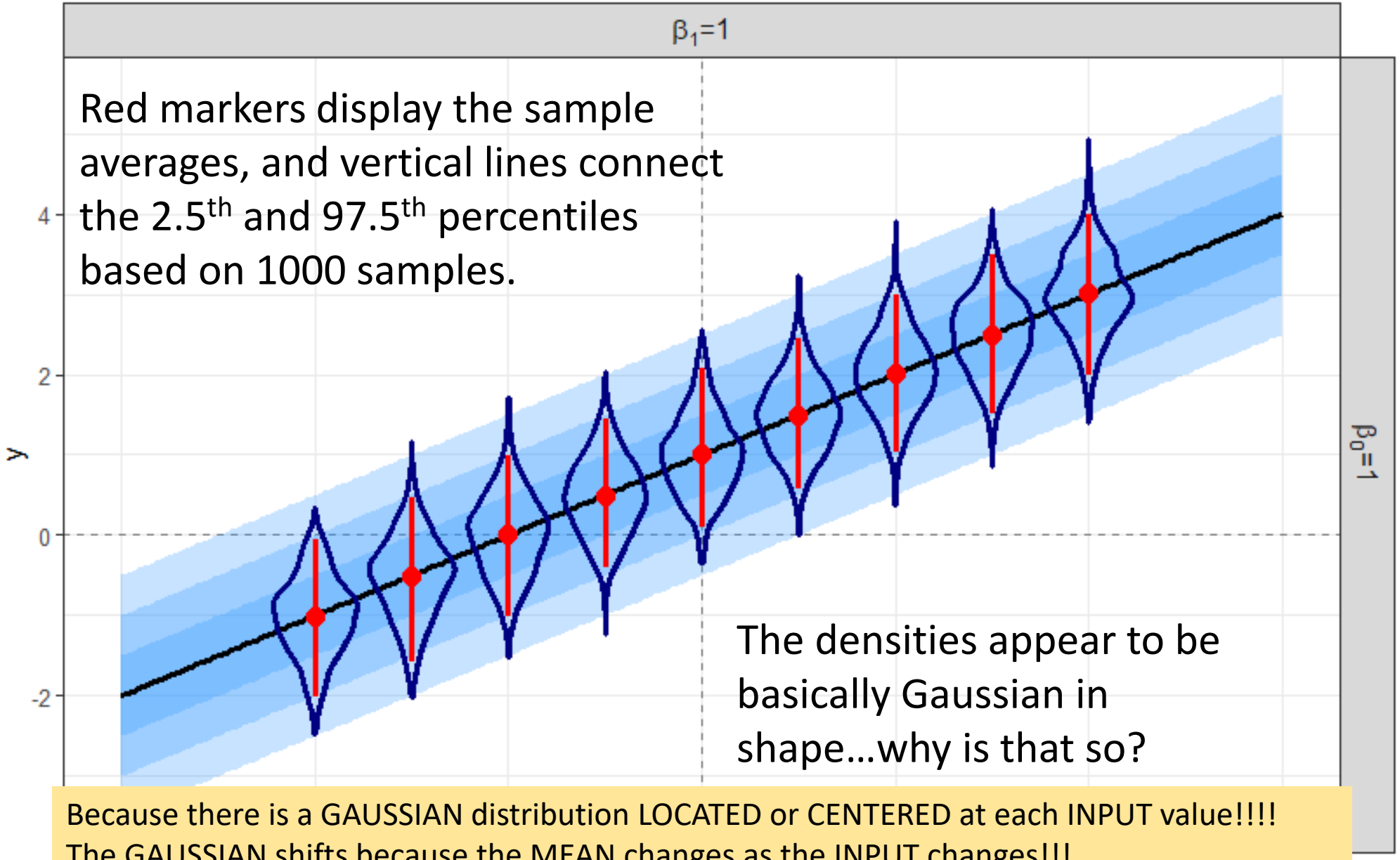
1000 observations per input location…why is the red so “wide”?



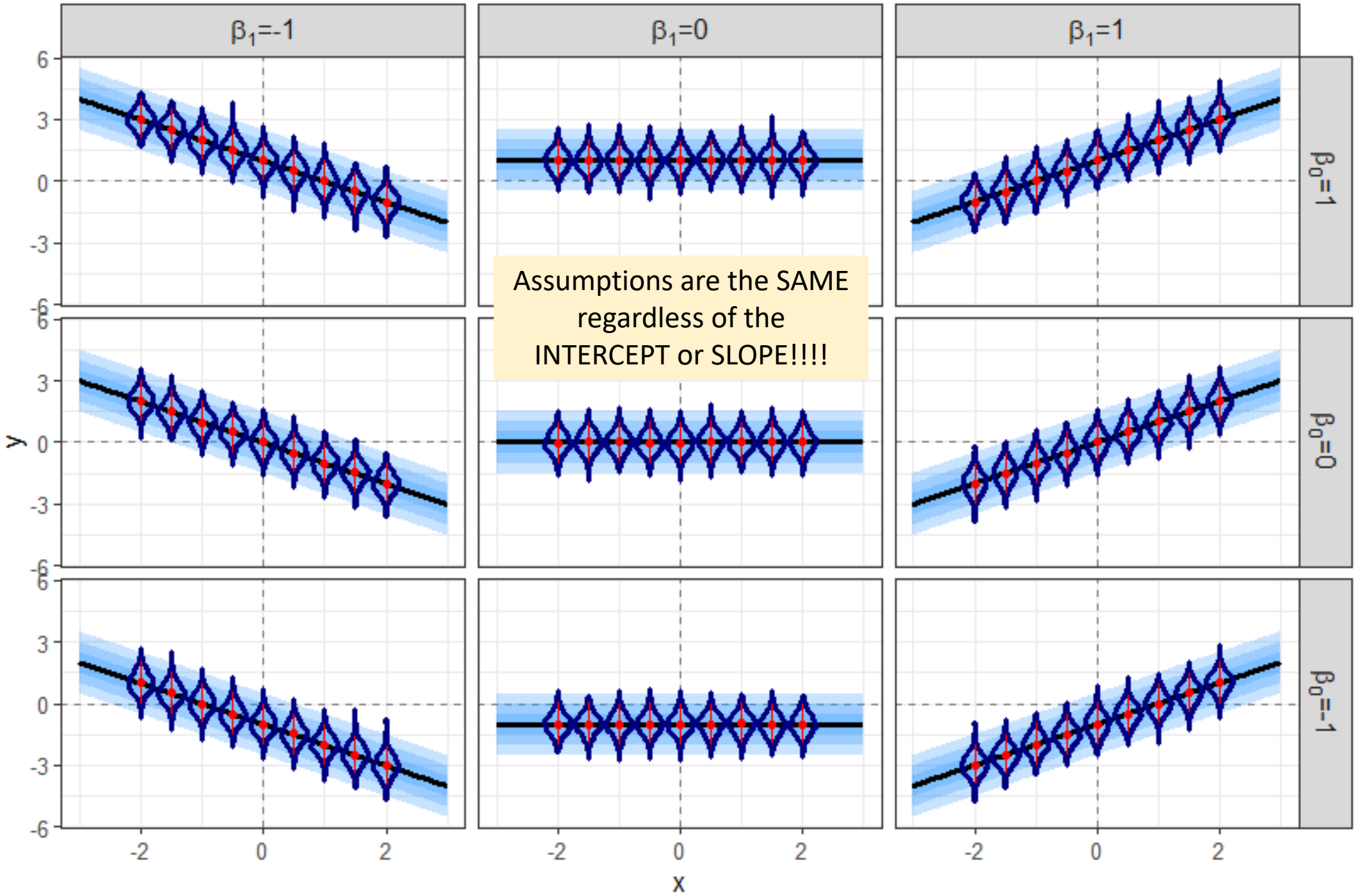
Violin plots represent the DENSITY at each input location



Violin plots represent the DENSITY at each input location



Because there is a GAUSSIAN distribution LOCATED or CENTERED at each INPUT value!!!!
The GAUSSIAN shifts because the MEAN changes as the INPUT changes!!!
The WIDTH of the GAUSSIAN is the SAME because linear models assume constant variance!!



Assumptions are the SAME regardless of the INTERCEPT or SLOPE!!!!

What makes a linear model…linear…?

Which of the following are linear models?

$$\mu_n = \beta_0 + \beta_1 x_n$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2}$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \beta_3 x_{n,1} x_{n,2}$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,1}^2$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,1}^2 + \beta_3 x_{n,1}^3 + \beta_4 x_{n,1} x_{n,2}^2$$

$$\mu_n = \beta_0 + \beta_1 \sin(x_n)$$

$$\mu_n = \beta_0 + \beta_1 \sin(1 + x_n^2)$$

$$\mu_n = \beta_0 \exp(\beta_1 x_n)$$

HINT: Only **ONE** is **NOT** a linear model…

$$\mu_n = \beta_0 + \beta_1 x_n$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2}$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \beta_3 x_{n,1} x_{n,2}$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,1}^2$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,1}^2 + \beta_3 x_{n,1}^3 + \beta_4 x_{n,1} x_{n,2}^2$$

$$\mu_n = \beta_0 + \beta_1 \sin(x_n)$$

$$\mu_n = \beta_0 + \beta_1 \sin(1 + x_n^2)$$

$$\mu_n = \beta_0 \exp(\beta_1 x_n)$$

HINT: Only **ONE** is **NOT** a linear model…

$$\mu_n = \beta_0 + \beta_1 x_n$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2}$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \beta_3 x_{n,1} x_{n,2}$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,1}^2$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,1}^2 + \beta_3 x_{n,1}^3 + \beta_4 x_{n,1} x_{n,2}^2$$

$$\mu_n = \beta_0 + \beta_1 \sin(x_n)$$

$$\mu_n = \beta_0 + \beta_1 \sin(1 + x_n^2)$$

$$\mu_n = \beta_0 \exp(\beta_1 x_n)$$

NOT a linear model!

A linear model is not linear because of the relationship between the response and the input

- What matters is the relationship between the response and the parameters or coefficients!

- If the parameters are linearly related to the average response, then the model is a linear model!

Consider $\beta_0 + \beta_1 \sin(x_n)$

- When we collect data, we observe N input-output pairs $\{x_n, y_n\}$.
- The $\sin(x_n)$, although non-linear, is just a function of the KNOWN input!
- We do not need to **learn** the value of $\sin(x_n)$, we just need to calculate it!

However, the parameters are unknown and must be learned from the data!

- The n-th observation's likelihood can be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

This is the same as in the linear relationship case!

However, the parameters are unknown and must be learned from the data!

- The n-th observation's likelihood can be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

This is the same as in the linear relationship case!

- We've changed the deterministic relationship between the mean trend and the input:

$$\mu_n = \beta_0 + \beta_1 \sin(x_n)$$

However, the parameters are unknown and must be learned from the data!

- The n-th observation's likelihood can be written as:

$$y_n \mid \mu_n, \sigma \sim \text{normal}(y_n \mid \mu_n, \sigma)$$

This is the same as in the linear relationship case!

- We've changed the deterministic relationship between the mean trend and the input:

$$\mu_n = \beta_0 + \beta_1 \sin(x_n)$$

- If we define a **FEATURE**, $\phi_n = \sin(x_n)$, the deterministic mean trend can be rewritten as:

$$\mu_n = \beta_0 + \beta_1 \phi_n$$

This should look familiar!

FEATURES allow us to extend the linear model framework to capture highly non-linear input-to-output relationships

- A simple set of features are the polynomial features:

$$\phi_j(x) = x^j$$

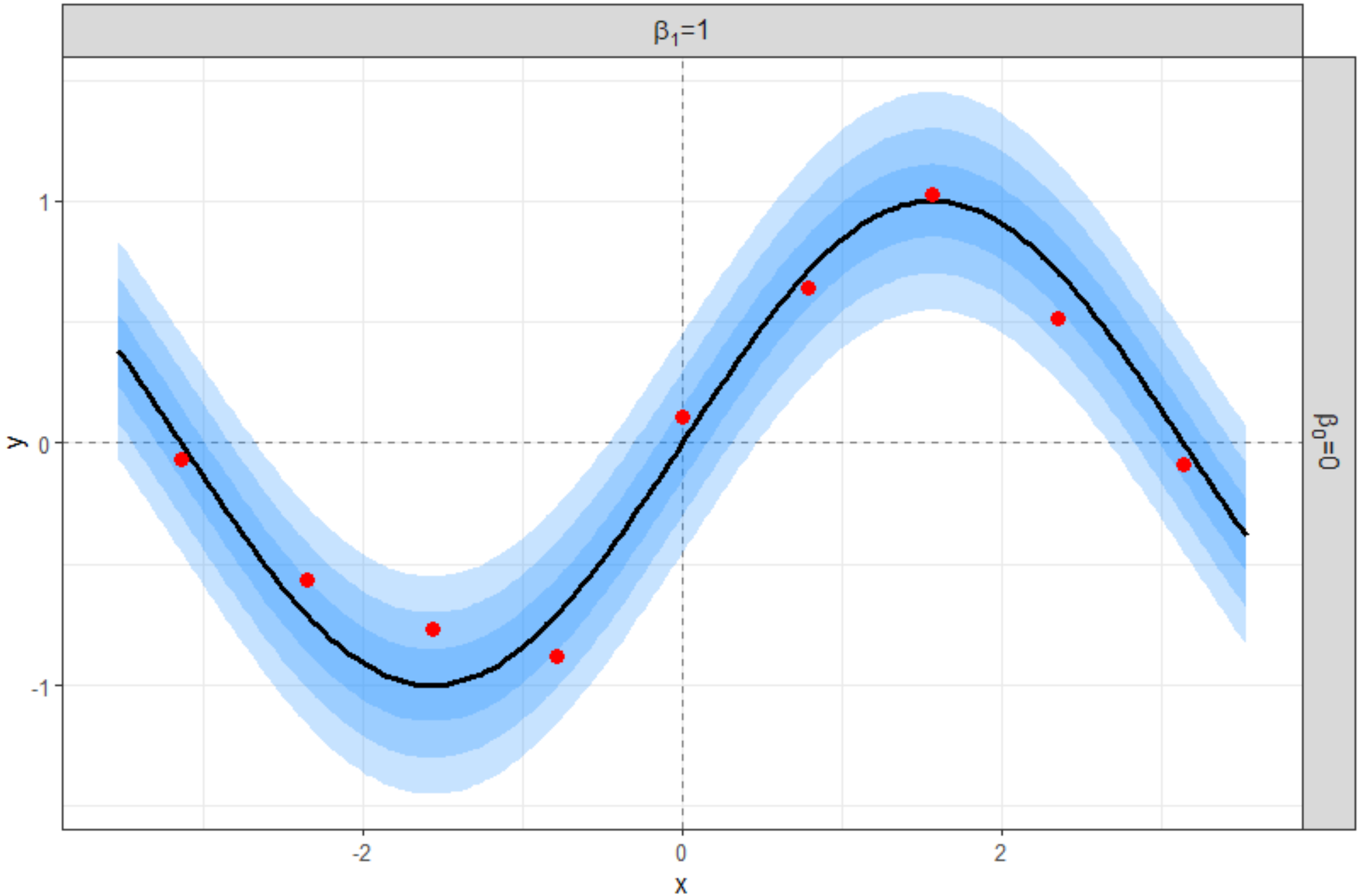
- A third order polynomial or cubic function can be written as:

$$\mu_n = \beta_0 + \sum_{j=1}^{J=3} \left(\beta_j \phi_j(x)\right) = \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \beta_3 x_n^3$$

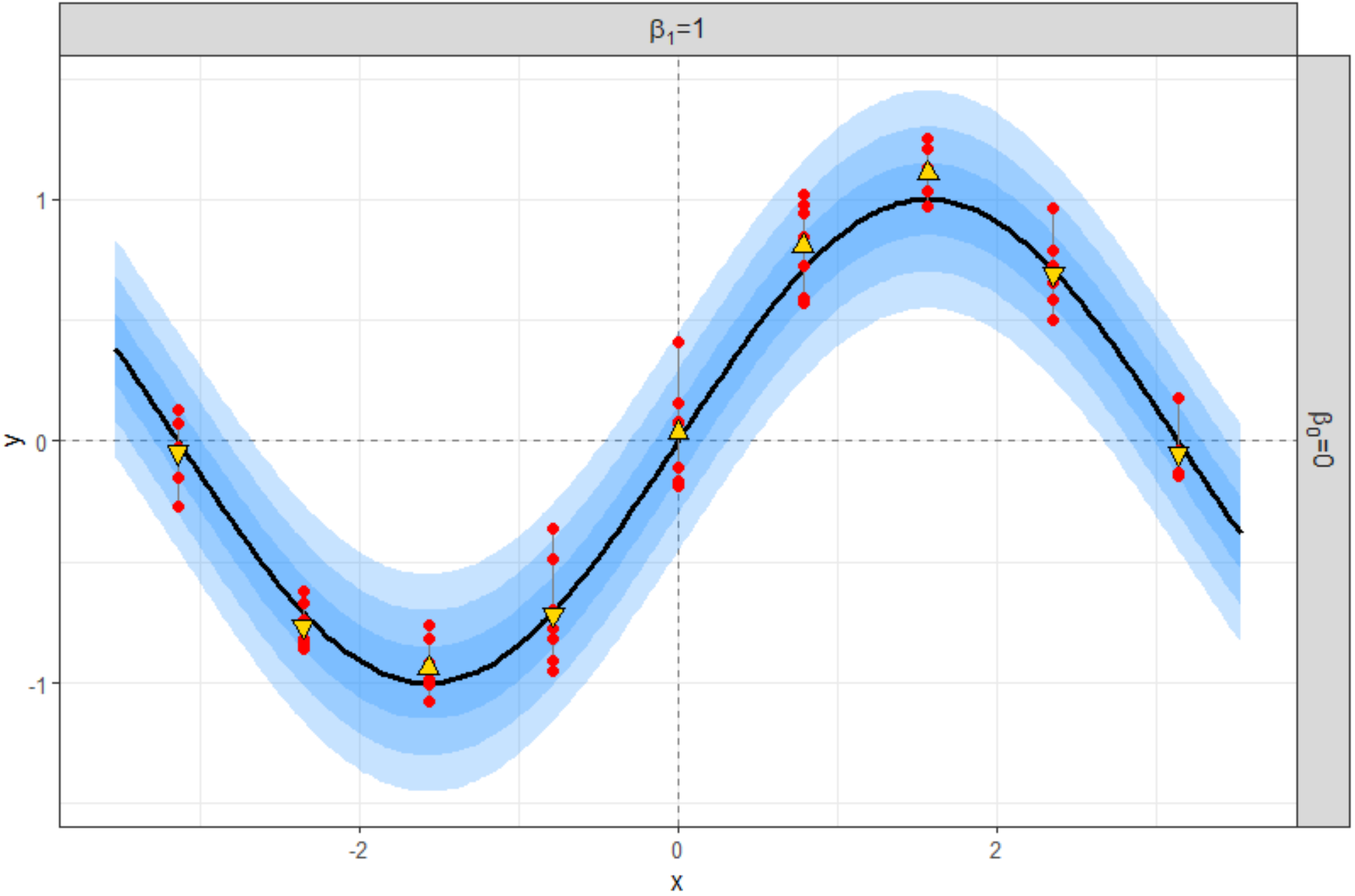
- Other common **features**: cosines/sines, splines, kernels.

Demonstrate on a simple sine function

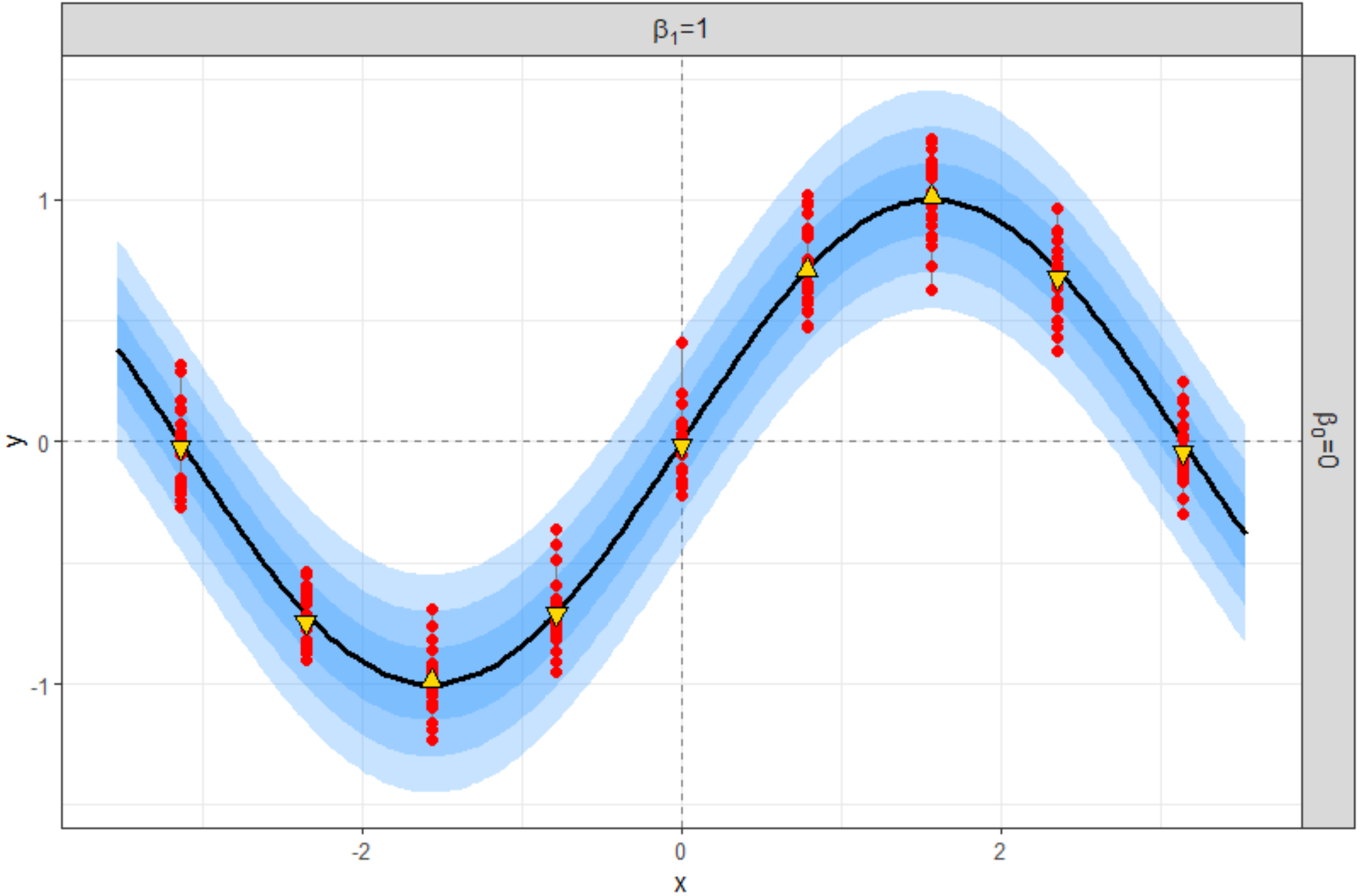
- Specify the coefficients to be $\beta_0 = 0$ and $\beta_1 = 1$.
- Use a relatively small likelihood noise, $\sigma = 0.15$.
- Specify 9 evenly spaced points between $-\pi$ and $+\pi$.
- Evaluate the mean trend with the sine function.
- Generate random draws around the mean trend.


β_1=1
β_0=0
y
x
1
0
-1
-2
0
2

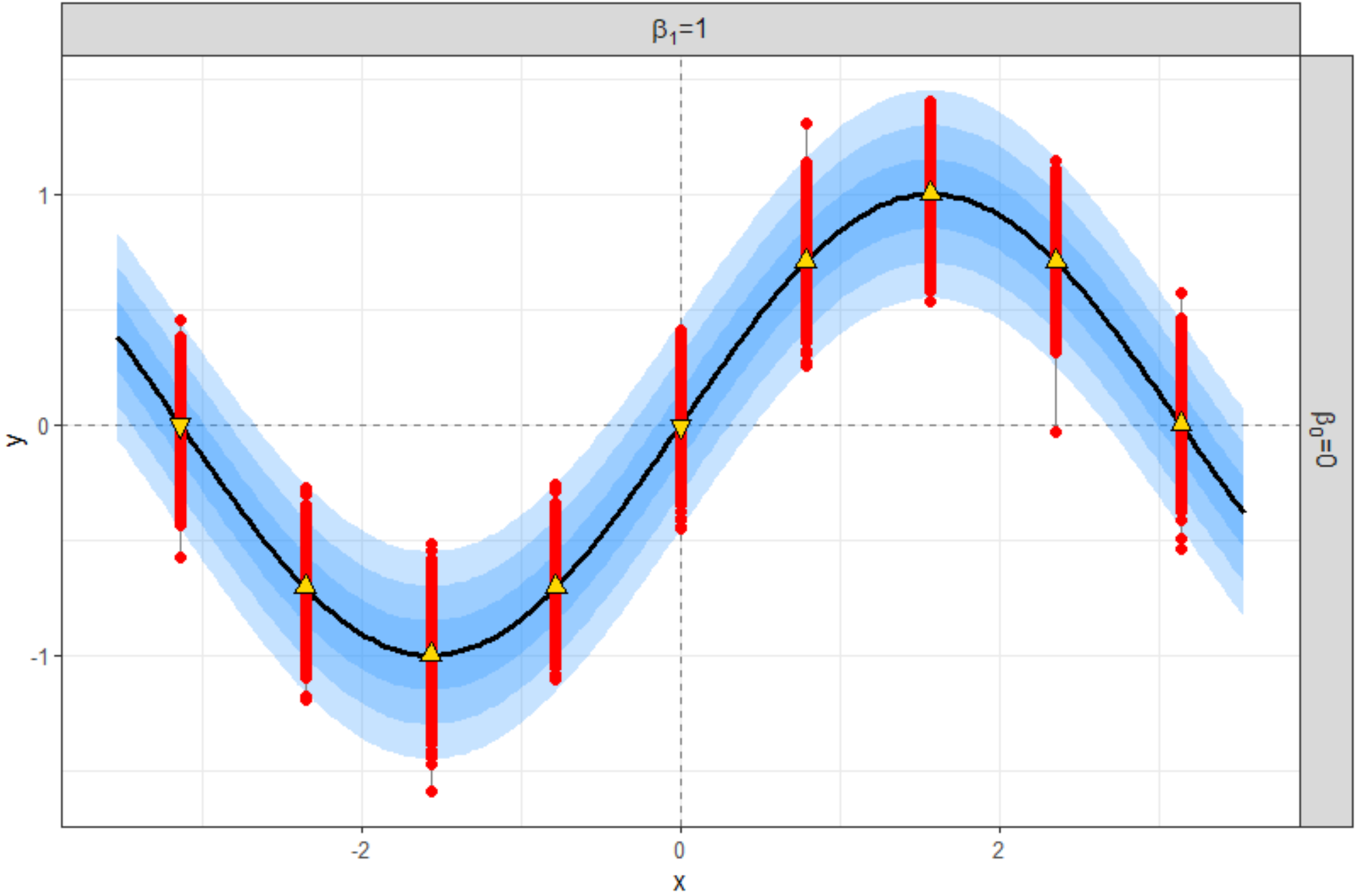
Increase sample size to 7 replications per input location



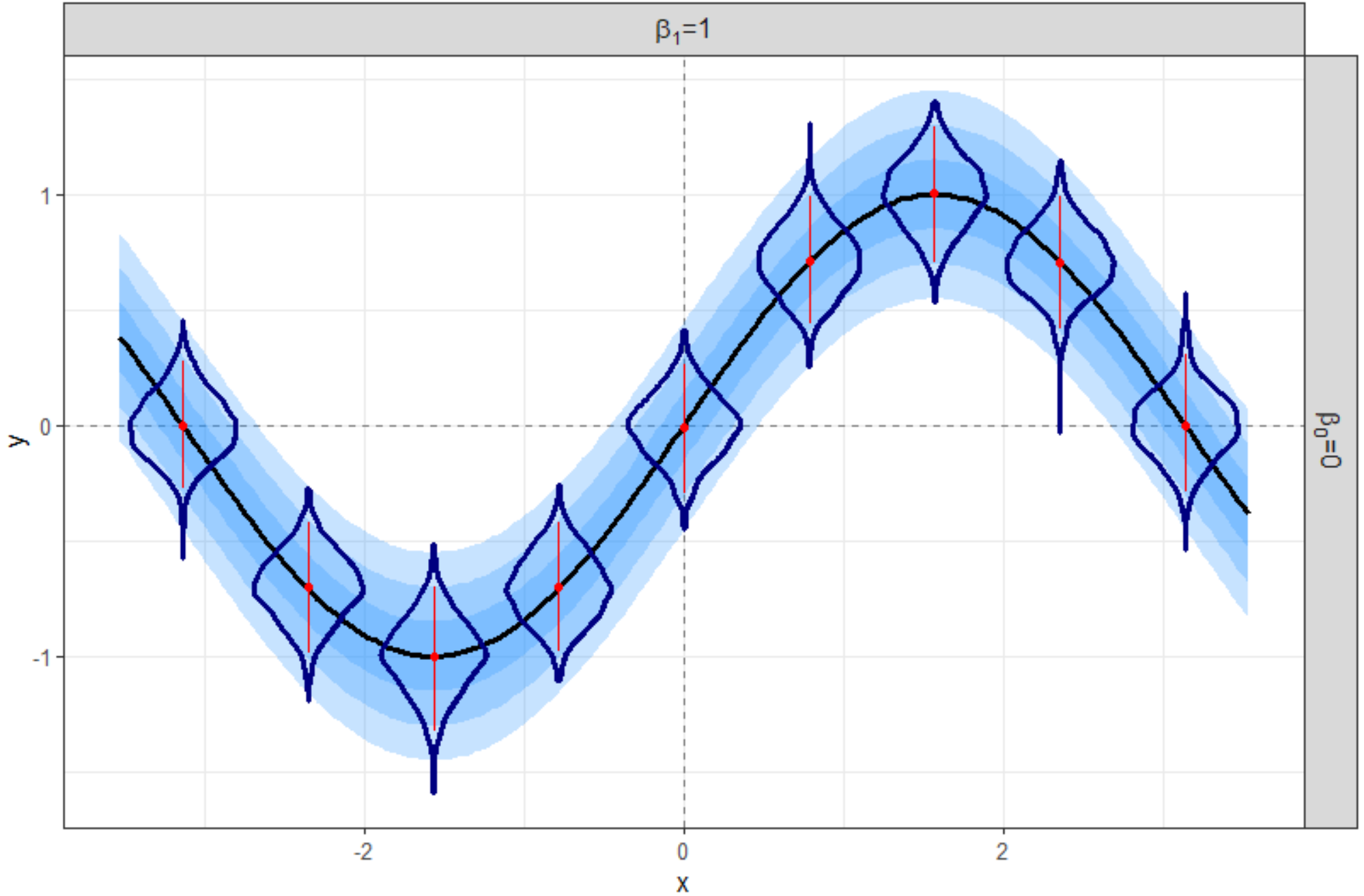
Increase sample size to 25 replications per input location



Increase sample size to 1000 replications per input location



Violin plots demonstrate the response density at each input location is Gaussian!



What about the example that was not a linear model…

$$\mu_n = \beta_0 \exp(\beta_1 x_n)$$

- The parameter β_1 is non-linearly related to μ_n!

- Would we have to change to a non-linear model?

Apply a natural log function…

$$\log[\mu_n] = \log[\beta_0] + \beta_1 x_n$$

- Define: $\eta_n = \log[\mu_n]$ and $b_0 = \log[\beta_0]$

- Now we have a new transformed model with all unknown parameters linearly related to (transformed) trend!

$$\eta_n = b_0 + \beta_1 x_n$$

Why is this useful?

- We have discussed the importance of visually exploring your data BEFORE doing anything else.

- If your exploration reveals the output varies by several orders of magnitude, you may consider applying the natural log or log-10 transformation to the output!

- You can then fit a linear model using the transformed output!

Transforming the output is common in applications involving predicted home sale prices

- Home prices can be on the order tens of thousands of dollars to multiple millions of dollars.
- 10% of a million dollar house is more than the price of a $75,000 house!
- Thus, if the error of our linear model is 10% of million dollar houses we may completely miss predicting behavior of less expensive homes!
- Applying the log transformation to the sale price “stabilizes the variance” and makes the modeling situation more consistent with the assumptions of the linear model!