Modelos Lineales

Optimización

Para modelizar la relación entre una variable dependiente Y y ciertos predictores X_l asumimos un modelo de la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \eta$$

En realidad tenemos N observaciones y por lo tanto para cada observación $(y_i, \mathbf{x}_i)$ tenemos

$$y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \cdots + \beta_p x_{p_i} + \eta$$

O bien, matricialmente

$$y_i = \beta^T \mathbf{x}_i + \eta$$

El error η es desconocido y, en principio, no es necesario asumir nada sobre este.

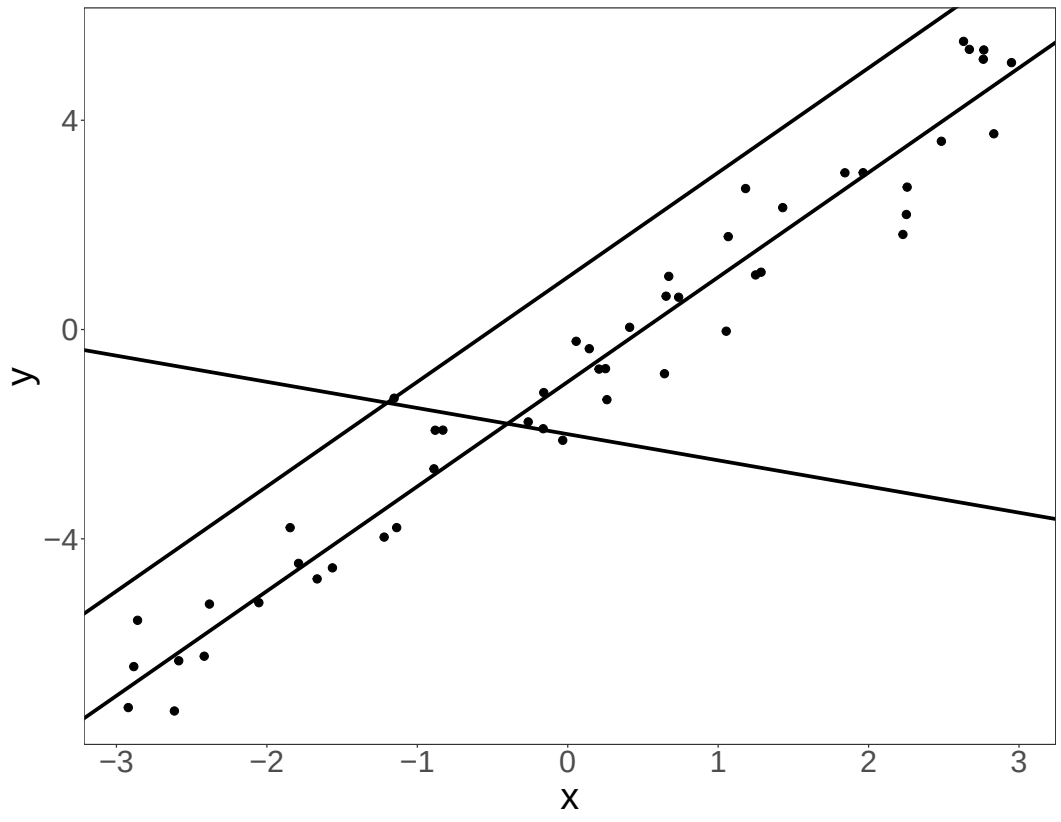
Para predecir valores de y_i es necesario estimar β por $\widehat{\beta}$ dando lugar al siguiente modelo predictivo:

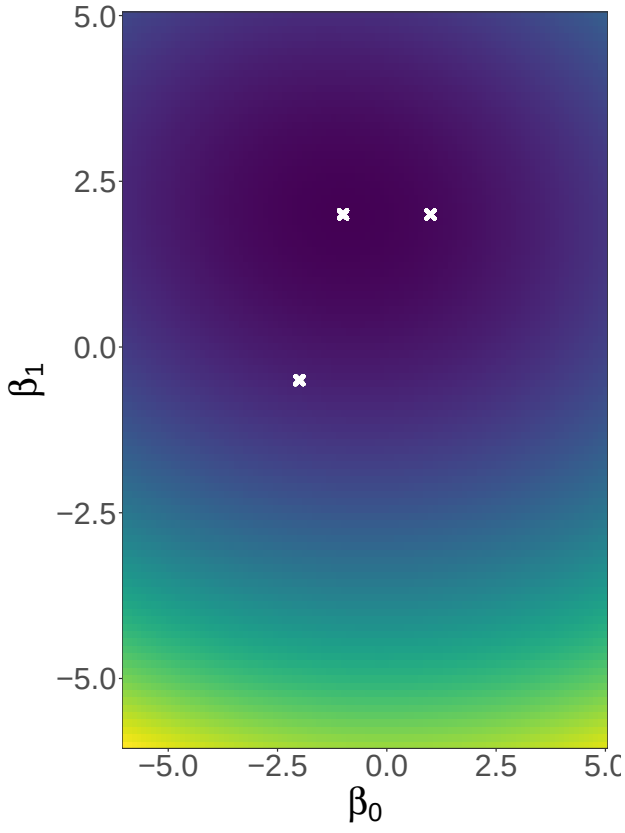
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1_i} + \widehat{\beta}_2 x_{2_i} + \cdots + \widehat{\beta}_p x_{p_i} = \widehat{\beta}^T \mathbf{x}_i$$

Una forma de estimar β es minimizar alguna función del error de aproximar y por $\widehat{y}_i$:

$$\widehat{\beta} = \arg\min_{\beta} \left[J(\beta) \right] = \arg\min_{\beta} \left[\sum_{i=1}^{N} \left(y_i - \beta^T \mathbf{x}_i \right)^2 \right]$$

El β que minimiza el error cuadrático se conoce como **estimador de mínimos cuadrados**. Esto es lo que se conoce como enfoque de optimización.


4
0
−4
y
−3
−2
−1
0
1
2
3
x


5.0
2.5
0.0
−2.5
−5.0
β1
−5.0
−2.5
0.0
2.5
5.0
β0

Estadística clásica

Asumiendo un modelo probabilístico para el error, $\eta \sim \mathcal{N}(0, \sigma^2)$ se puede obtener el estimador de máxima verosimilitud de β.

La función de verosimilitud viene dada por el producto de las funciones de densidad normales:

$$\ell(\beta, \sigma|\mathbf{y}) = \prod_{i=1}^{N} p(y_i|\mathbf{x_i}, \beta^T, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i - \beta^T \mathbf{x}_i)^2}{2\sigma^2}}$$

Maximizar la verosimilitud equivale a minimizar el opuesto de la log-verosimilitud

$$\mathcal{L}(\beta, \sigma|\mathbf{y}) = \log(\ell(\beta, \sigma|\mathbf{y}))$$

$$\widehat{\beta}_{ML} = \arg\min_{\beta} \left[-\sum_{i=1}^{N} \log \left(\left(\frac{1}{2\pi\sigma^2} \right)^{1/2} e^{-\frac{(y_i - \beta^T \mathbf{x}_i)^2}{2\sigma^2}} \right) \right]$$

La expresión anterior puede minimizarse primero respecto de β y luego respecto de σ. Resulta que maximizar la verosimilitud respecto de β equivale a minimizar el error cuadrático.

Estadística bayesiana

En estadística bayesiana, consideramos a los parámetros como variables aleatorias y les asignamos una distribución *a priori*.

Además, contamos con un modelo generativo (probabilístico) para las observaciones: ¿cómo obtendríamos observaciones si conociéramos los parámetros? Es una decisión de la modelización.

Aquí asumimos:

$$Y_i \mid \beta, \sigma \sim \mathcal{N}(\beta^T \mathbf{x}_i, \sigma^2)$$

o bien decimos

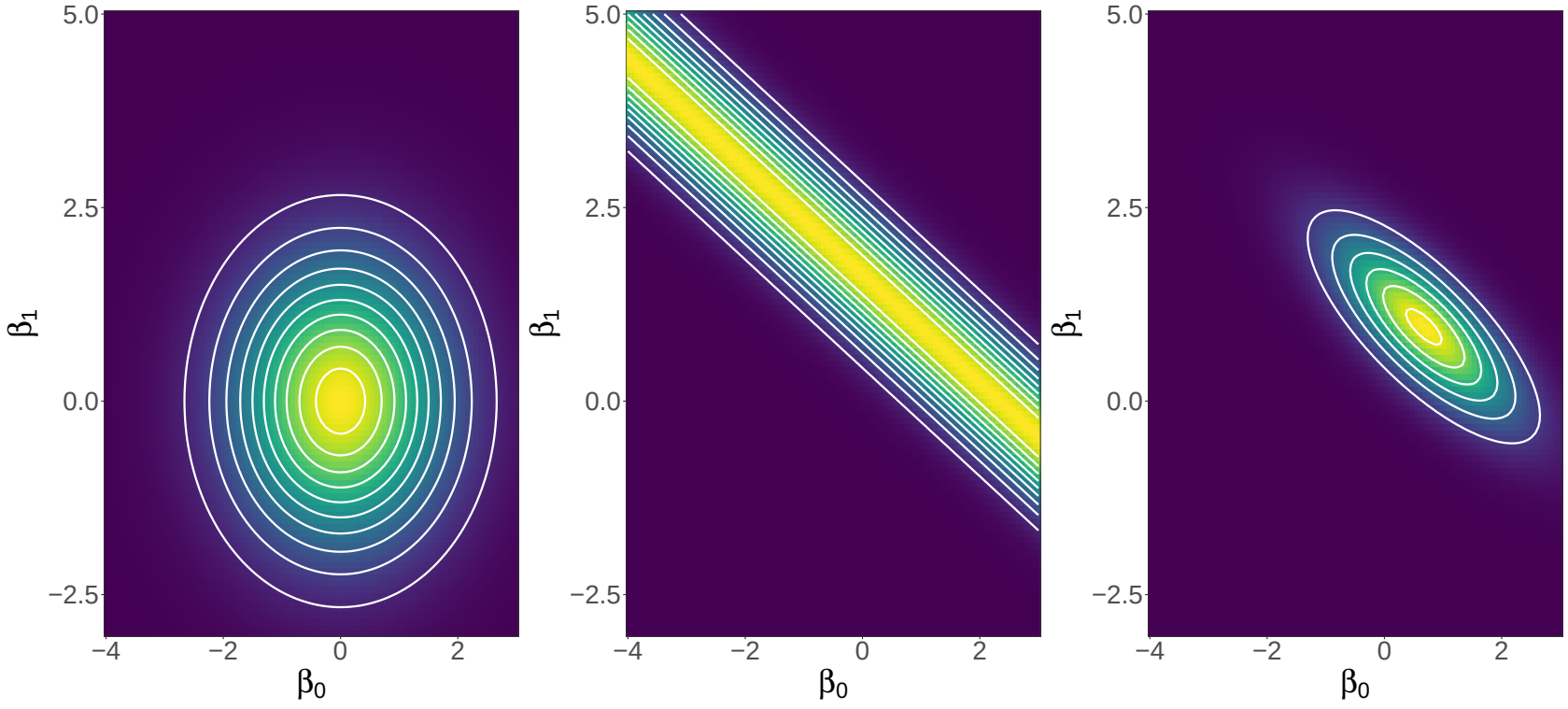
$$\begin{aligned} Y_i &\sim \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \cdots + \beta_p x_{p_i} \end{aligned}$$

Y completamos el modelo especificando una distribución *a priori* $p(\beta, \sigma)$

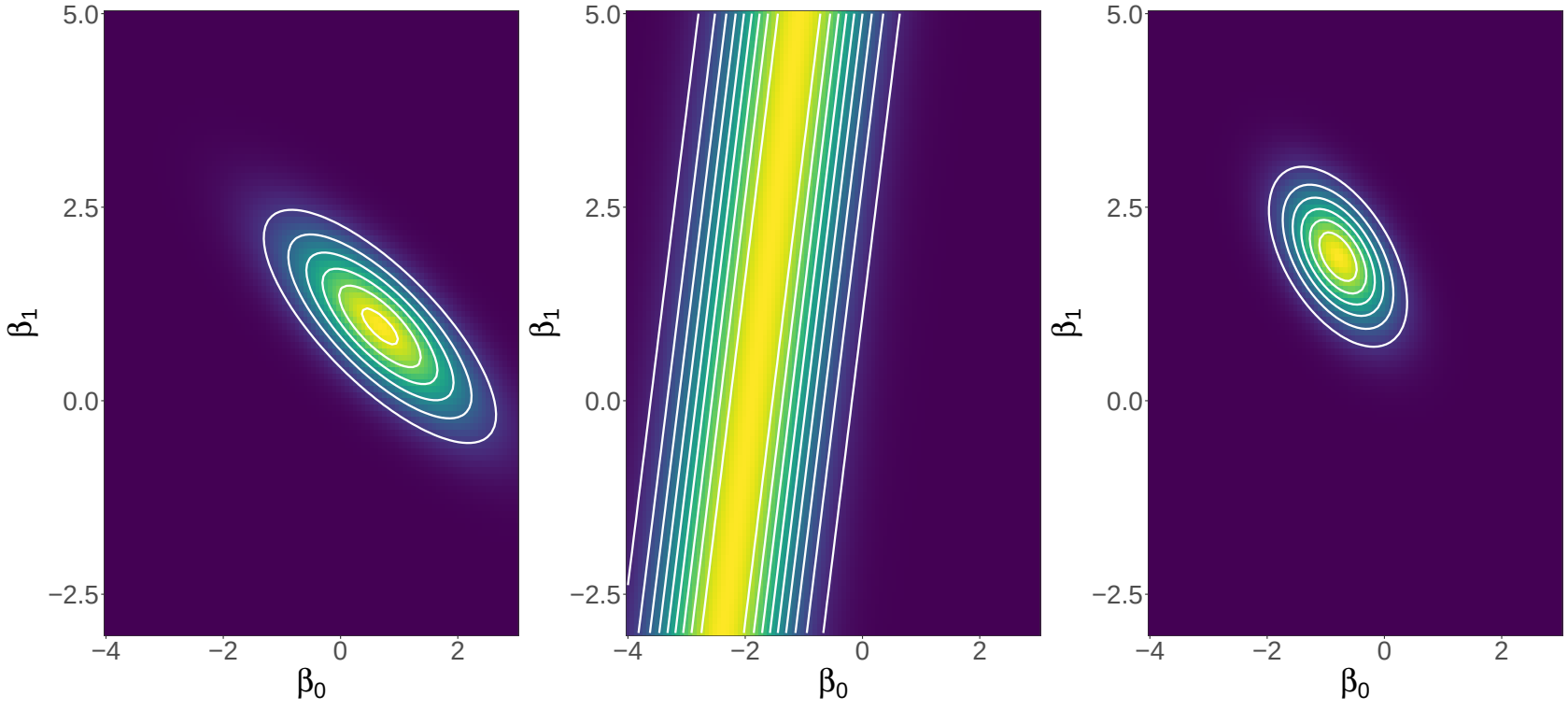
La estimación se hace siempre de la misma manera

$$p(\beta, \sigma \mid \mathbf{y}) \propto p(\mathbf{y}|\beta, \sigma)p(\beta, \sigma)$$

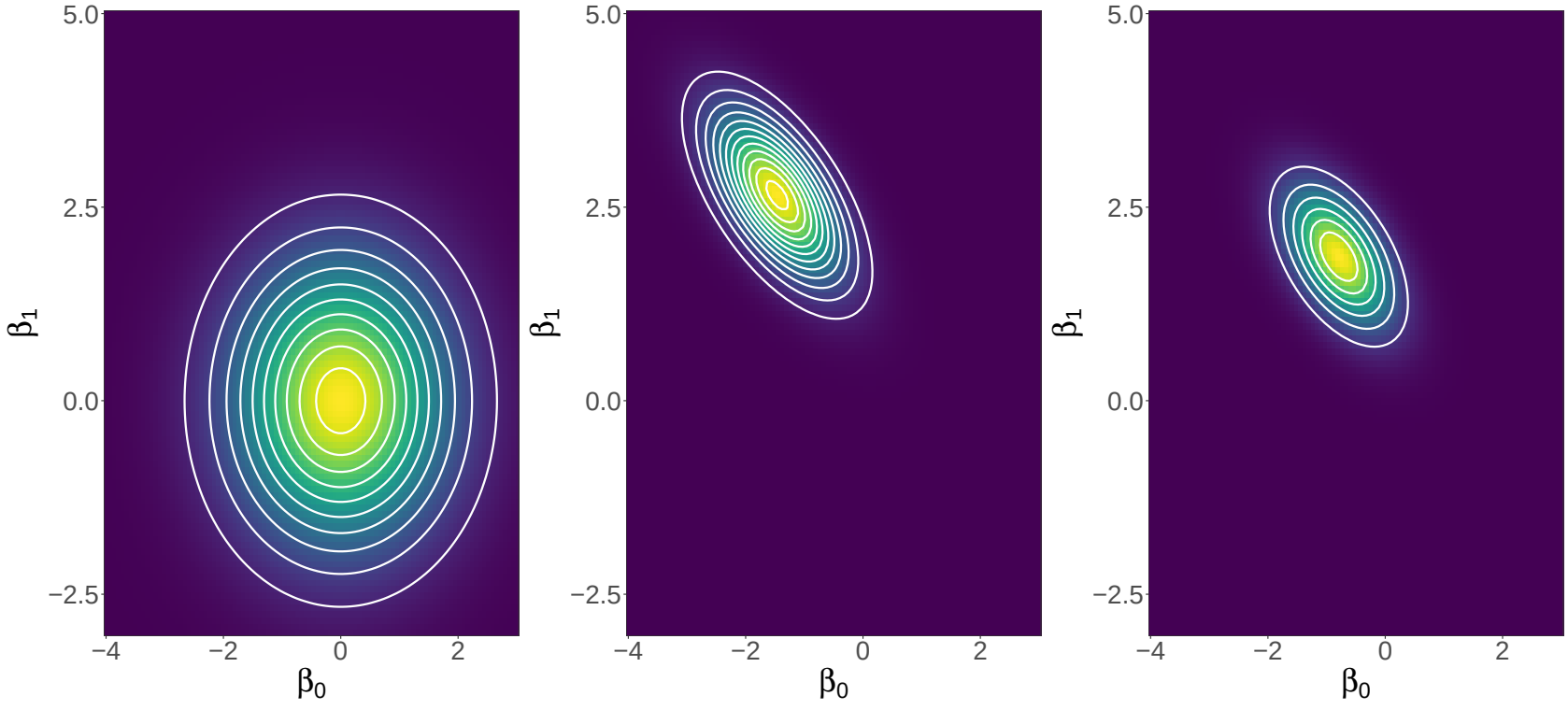
Observando un dato (¿se puede hacer inferencia con un solo punto?)...



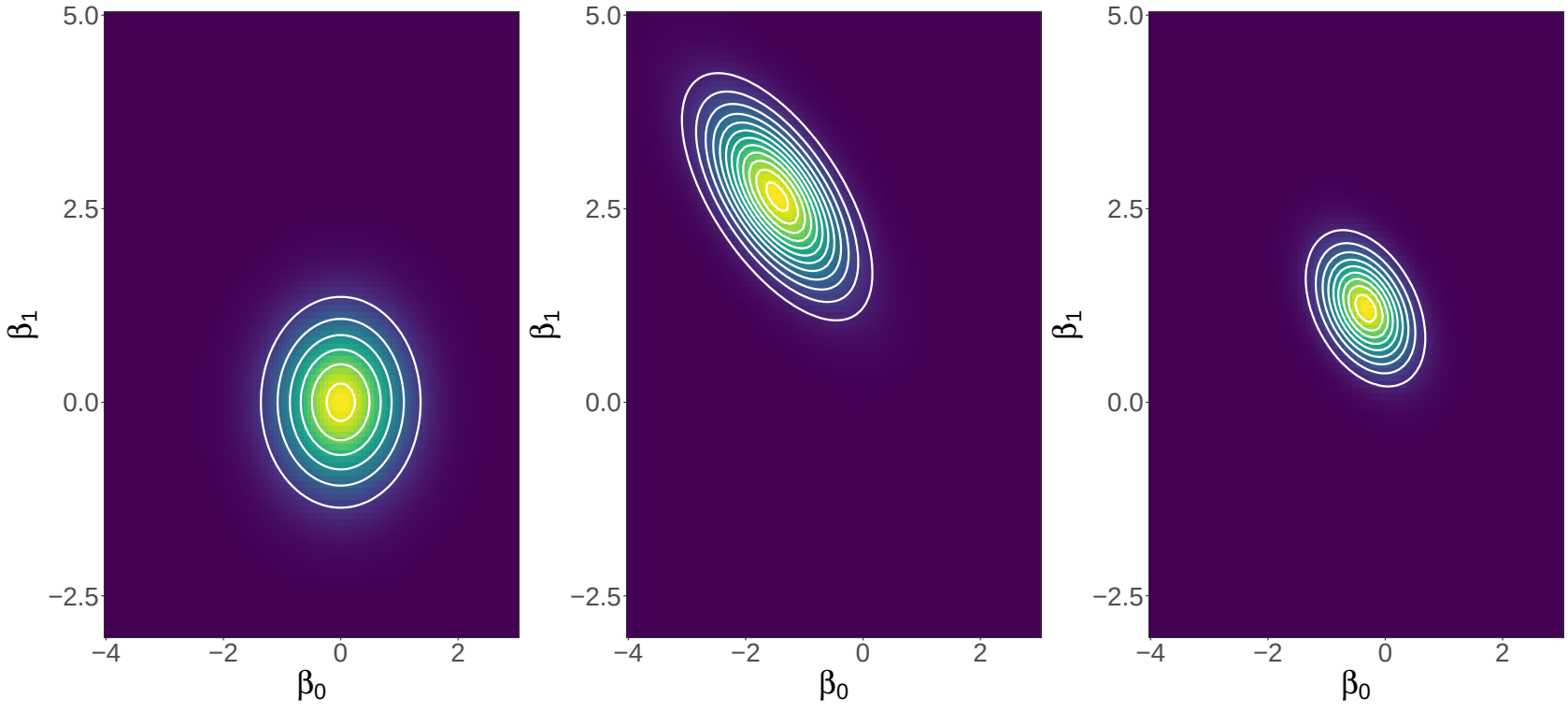
Observando el dato que sigue...



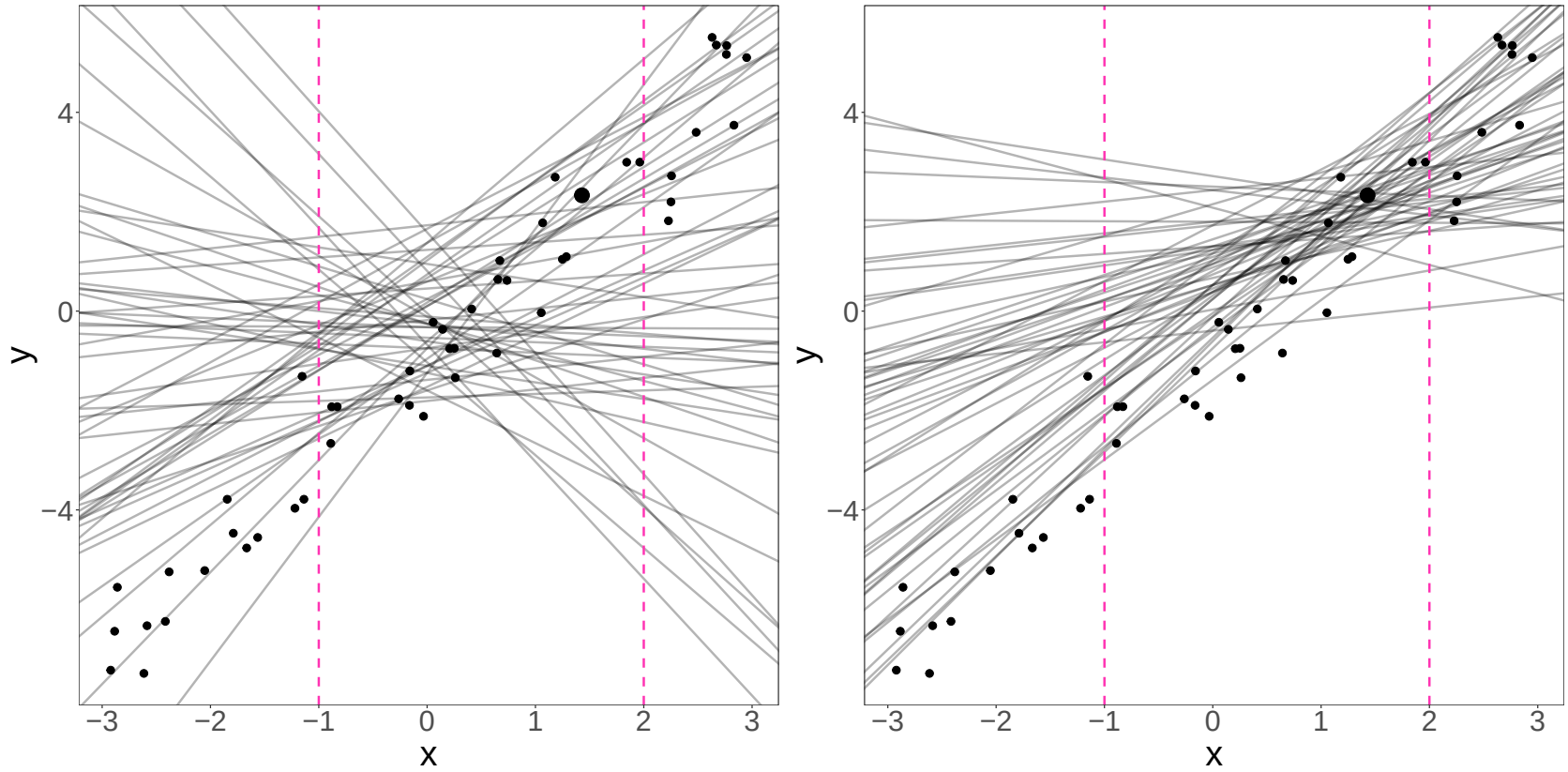
Observando dos puntos juntos...



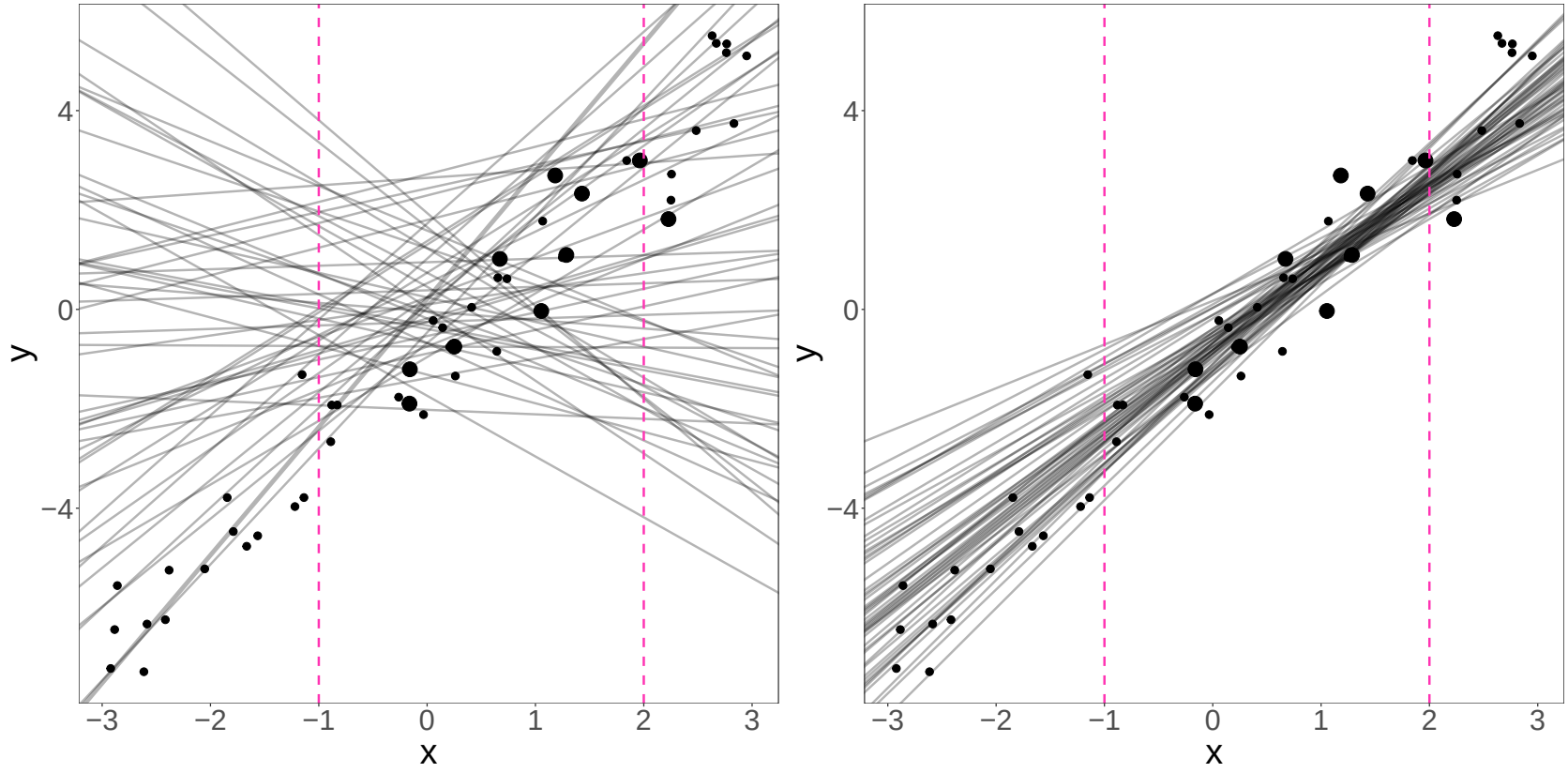
Comparemos con un *prior* más fuerte...



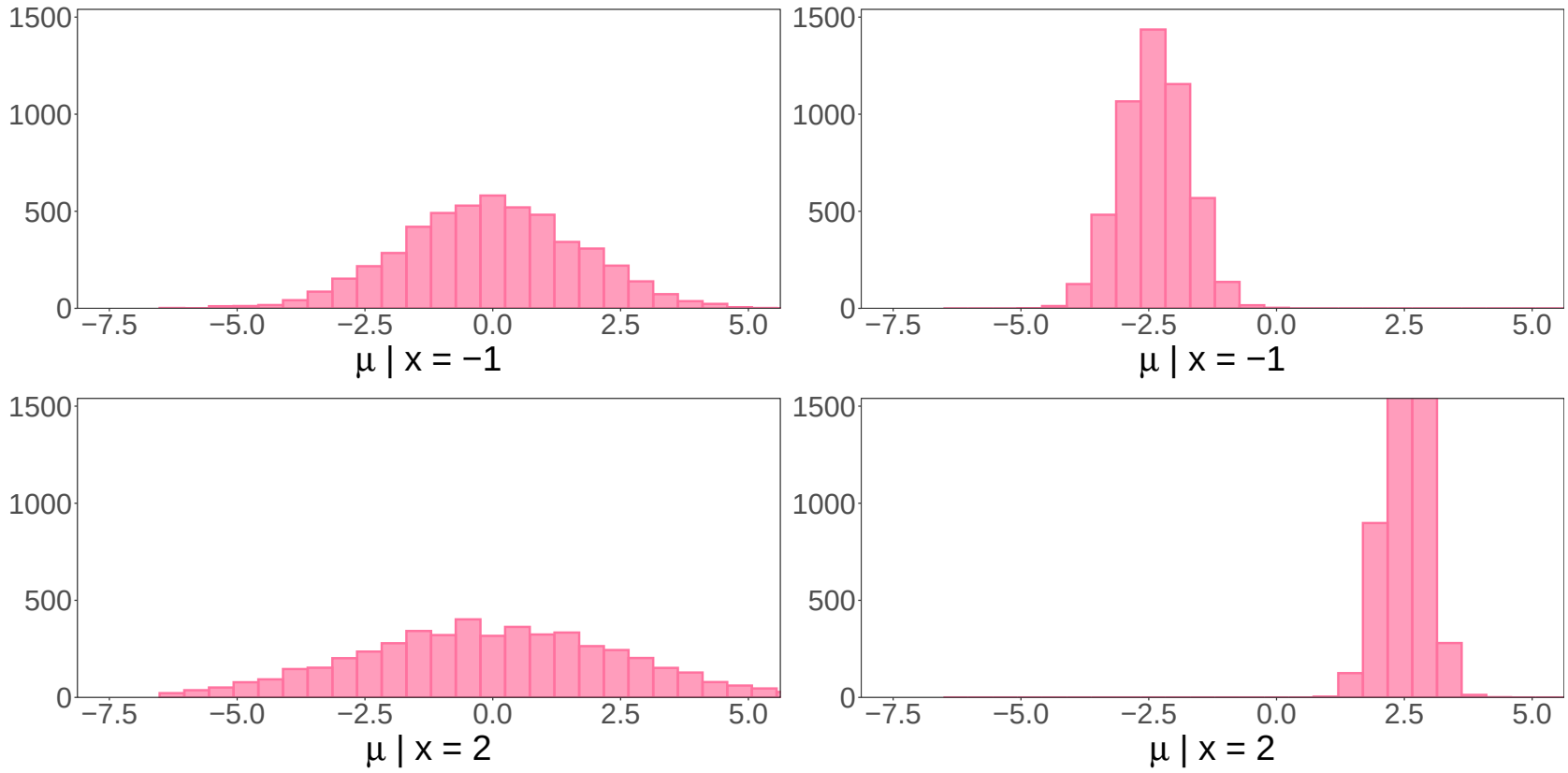
Observando **un punto** (¿se puede hacer inferencia con un solo punto?)



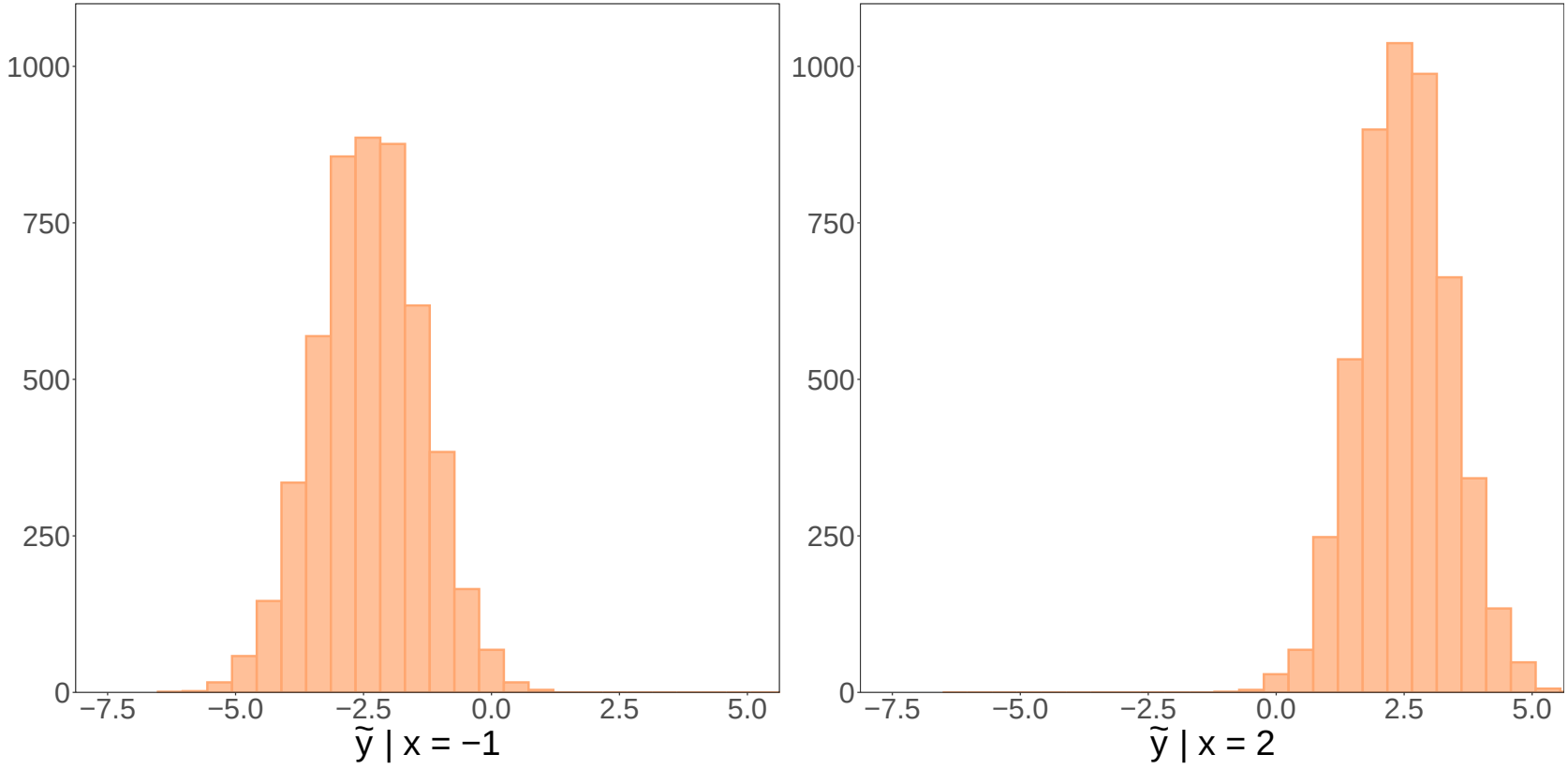
Observando **diez puntos**



μ depende de los parámetros (y por supuesto del valor de x), por lo que tiene una distribución de probabilidad asociada



Por supuesto, las predicciones para y ($\tilde{y}$) también son probabilísticas: distribución predictiva *a posteriori*



Resumen

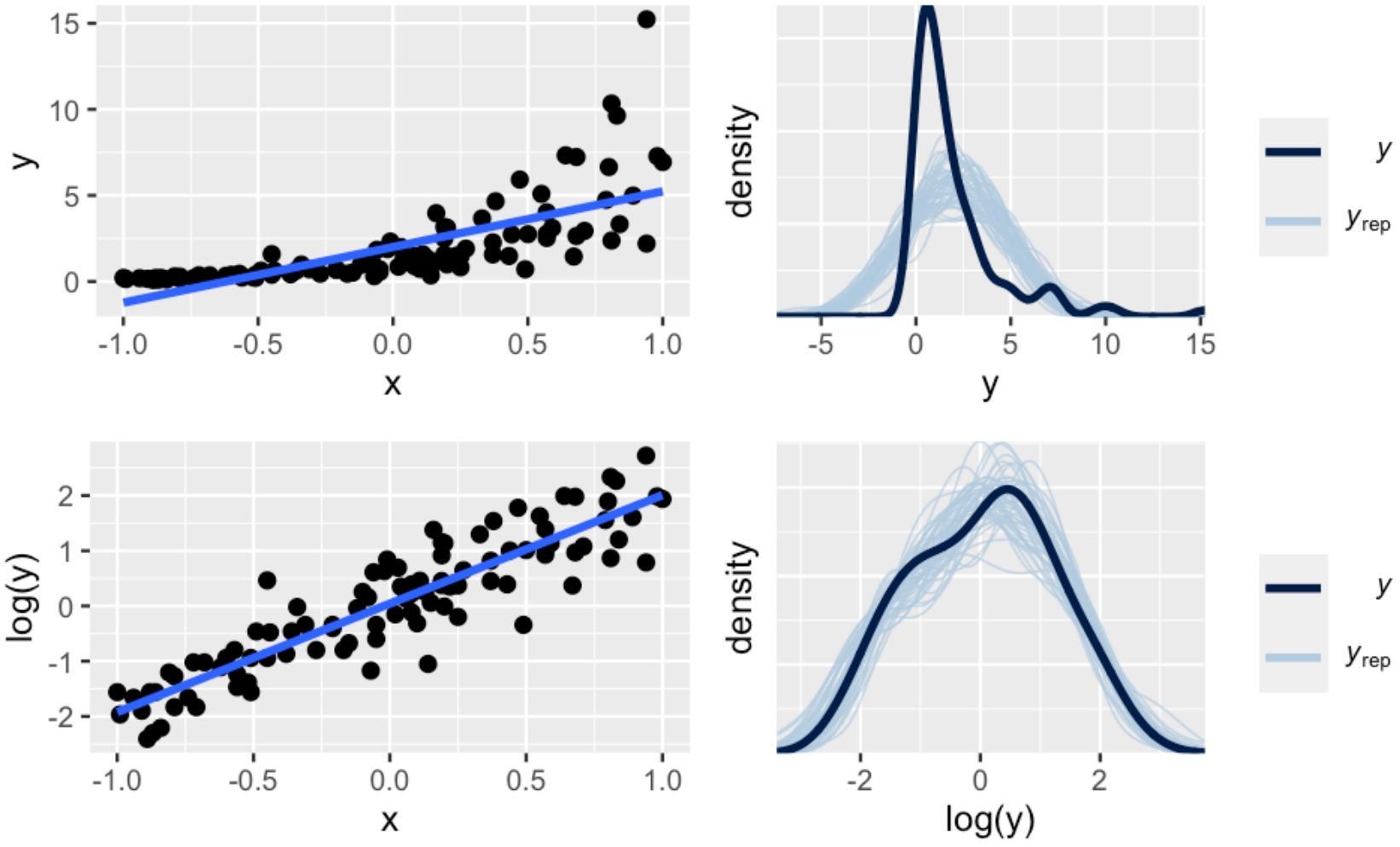
- Tenemos una distribución de probabilidad para los parámetros. Es decir, tenemos incertidumbre en los valores de los parámetros
- Tenemos que trabajar con todo el *posterior* (a través de muestras) y no con estimaciones puntuales
- No confundir predicción de la media (también llamado predictor lineal) con distribución predictiva (para las observaciones)
- A medida que aumenta el tamaño de muestra, los coeficientes de la regresión se estiman cada vez con mayor precisión y la incertidumbre del predictor lineal desaparece (incertidumbre epistémica). No obstante, la incertidumbre en la distribución predictiva no desaparece (siempre quedará σ: incertidumbre aleatoria).

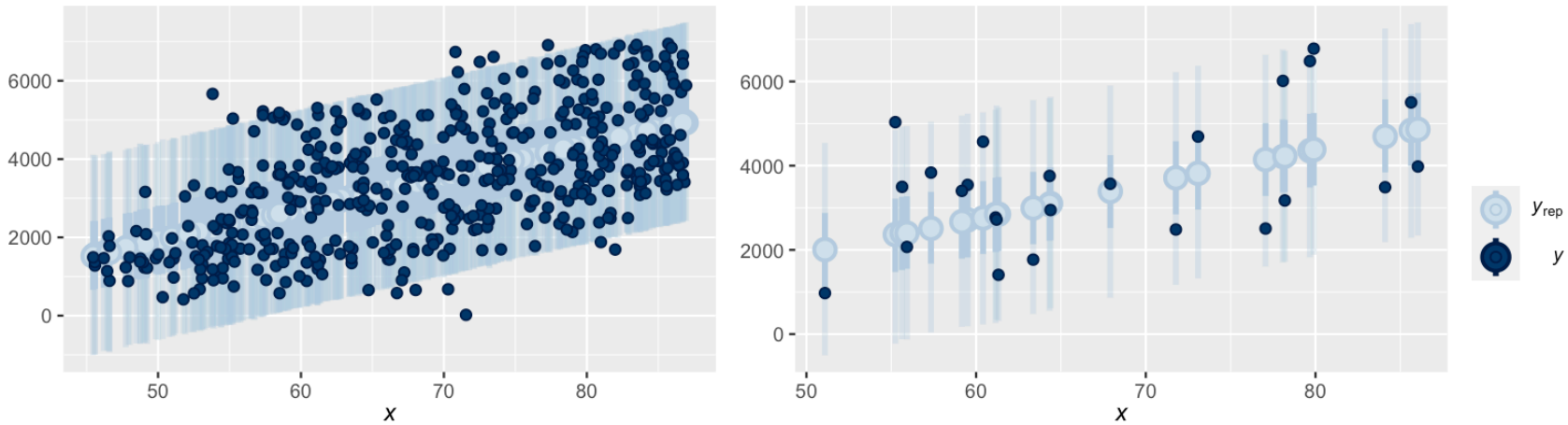
Validación interna

- El modo fundamental de validar el ajuste de un modelo bayesiano es generar réplicas del conjunto de datos (utilizando el modelo ajustado) y compararlas con los datos reales. Esto es lo que se conoce como **validación interna**.
- Para cada muestra de parámetros del *posterior* podemos generar un dataset

$$\left[\begin{array}{c|c|c|c} Y_1^{(1)} & Y_2^{(1)} & \dots & Y_N^{(1)} \\ Y_1^{(2)} & Y_2^{(2)} & \dots & Y_N^{(2)} \\ \vdots & \vdots & & \vdots \\ Y_1^{(S)} & Y_2^{(S)} & \dots & Y_N^{(S)} \end{array}\right]$$

- Esta práctica da lugar a los *posterior predictive checks* (PPC)


15
10
5
0
y
-1.0
-0.5
0.0
0.5
1.0
x
density
-5
0
5
10
15
y
y
y_{rep}
2
1
0
-1
-2
log(y)
-1.0
-0.5
0.0
0.5
1.0
x
density
-2
0
2
log(y)
y
y_{rep}


6000
4000
2000
0
50
60
70
80
x
6000
4000
2000
0
50
60
70
80
x
y_{rep}
y

Validación externa

- Idealmente quisiéramos ver si nuestro modelo tiene capacidad predictiva para datos nuevos (no usados para ajustarlo)
- Antes de preocuparnos por los datos nuevos, pensemos en las predicciones... Las predicciones son probabilísticas
- No podemos simplemente comparar y_i con $\hat{y}_i$
- Debemos utilizar toda la distribución *a posteriori* para evaluar el ajuste (y la capacidad predictiva) del modelo

Un posible *score* predictivo para un determinado valor y_i es la probabilidad que el modelo le asocia (también llamada densidad predictiva),

$$\int p\left(y_i \mid \theta\right) p(\theta \mid y) d\theta \approx \frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^{(s)})$$

El *score* predictivo total (para todas las observaciones) es la *log-posterior pointwise predictive density*. A mayor lppd, mejor es el ajuste del modelo.

$$\mathrm{lppd} = \sum_{i=1}^{N} \log \left(\int p\left(y_i \mid \theta\right) p(\theta \mid y) d\theta \right)$$

El *score* predictivo total (para todas las observaciones) es la *log-posterior pointwise predictive density*. A mayor lppd, mejor es el ajuste del modelo.

$$\mathrm{lppd} = \sum_{i=1}^{N} \log \left(\int p\left(y_i \mid \theta\right) p(\theta \mid y) d\theta \right)$$

Que podemos estimar a través de muestras del *posterior*

$$\mathrm{lppd} = \sum_{i=1}^{N} \log \left(\frac{1}{S} \sum_{s=1}^{S} p\left(y_i \mid \theta^{(s)}\right) \right)$$

La **deviance** de un modelo es

$$D = -2 \ \mathrm{lppd}$$

- La deviance (o la lppd) evalúa las predicciones de un modelo (el ajuste), no nos dice qué tan correcto es...
- Son medidas que siempre mejoran con más parámetros
- En realidad nos importa cómo se desempeña el modelo con datos nuevos.

La lppd predice el i-ésimo valor con un *posterior* que usa todos los datos, incluido el i. Más que la lppd nos interesa su valor esperado en datos nuevos (elppd). Por supuesto, no conocemos datos nuevos.

La lppd predice el i-ésimo valor con un *posterior* que usa todos los datos, incluido el i. Más que la lppd nos interesa su valor esperado en datos nuevos (elppd). Por supuesto, no conocemos datos nuevos.

Podemos aproximar o estimar elppd haciendo *cross-validation* (CV) o, en particular, *leave-one-out cross-validation* (LOO-CV).

$$\mathrm{elppd} \approx \mathrm{lppd}_{LOO} = \sum_{i=1}^{N} \log \left(\frac{1}{S} \sum_{s=1}^{S} p\left(y_i \mid \theta_{-i}^{(s)}\right) \right)$$

donde los $\theta_{-i}^{(s)}$ son muestras del *posterior* de θ obtenido sin considerar la i-ésima observación.

El problema de hacer LOO-CV es que, si tenemos 1000 observaciones, hay que calcular 1000 distribuciones *a posteriori*.

No contamos con muestras de $p(\theta \mid \mathbf{y}_{-i})$ sino simplemente de $p(\theta \mid \mathbf{y})$. No sabemos la distribución *a posteriori* de θ sin considerar la observación i. Hay formas de aproximar el desempeño en LOO-CV sin necesidad de reajustar el modelo. Una forma de hacerlo es usar la "importancia" de cada observación en el *posterior*. Esto da lugar a una técnica que se conoce como *Pareto-smoothed importance sampling cross-validation* (PSIS)

Históricamente se han desarrollado los llamados **criterios de información** que penalizan la verosimilitud con un término adicional para compensar la capacidad de sobreajuste de un modelo que tiene más parámetros.

El **AIC** (*Akaike information criterion*) es

$$AIC = D + 2p = -2\ lppd + 2p$$

donde p es el número de parámetros del modelo y $D = -2\ \mathrm{lppd}$ se conoce como **deviance**. Penalizamos el lppd con la tendencia (o capacidad) del modelo de sobreajustar.

El **WAIC** (*widely applicable information criterion*) es un criterio más general que el AIC (y un poquitito más difícil de calcular):

$$WAIC = -2 \left(\text{lppd} - \sum_{i=1}^{N} \mathbb{V}_\theta \left[\log \left(p(y_i \mid \theta) \right] \right) \right.$$

$\sum_{i=1}^{N} \mathrm{V}_\theta \left[\log \left(p(y_i \mid \theta) \right] \right.$ es un término de penalización que se suele llamar "número efectivo de parámetros". Es la suma de las varianzas en la log-probabilidad de cada observación i (o sea, la varianza total). Si, para un determinado dato i, las diferentes muestras del *posterior* $\theta_{(s)}$ dan como resultado predicciones muy diferentes, es porque el modelo tiene mucha incertidumbre (y es posiblemente muy flexible).