

## 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  $\eta$  es desconocido y, en principio, no es necesario asumir nada sobre este.

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

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

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

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

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



# 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  $\beta$ .

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

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

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

La expresión anterior puede minimizarse primero respecto de  $\beta$  y luego respecto de  $\sigma$ . Resulta que maximizar la verosimilitud respecto de  $\beta$  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

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

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

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



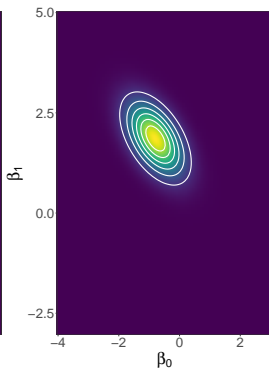
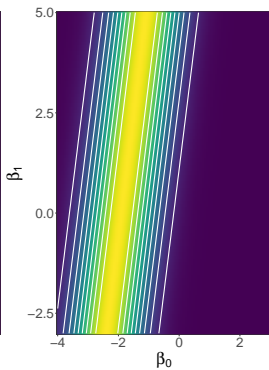
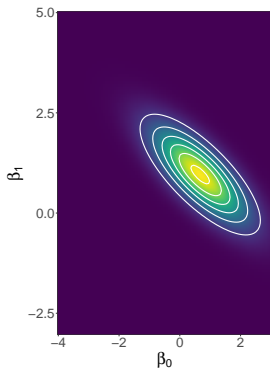
La estimación se hace siempre de la misma manera

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

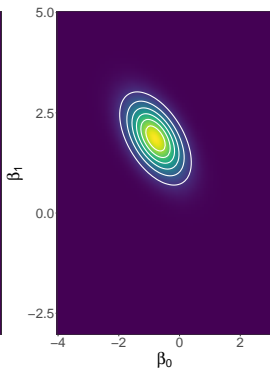
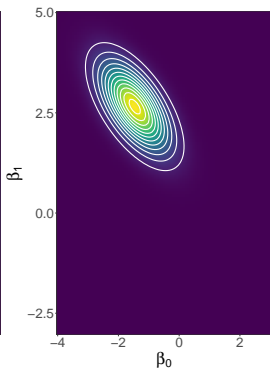
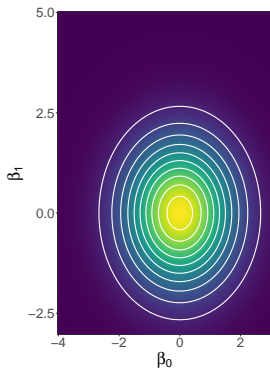
Observando un dato (¿se puede hacer inferencia con un solo punto?)...



Observando el dato que sigue...



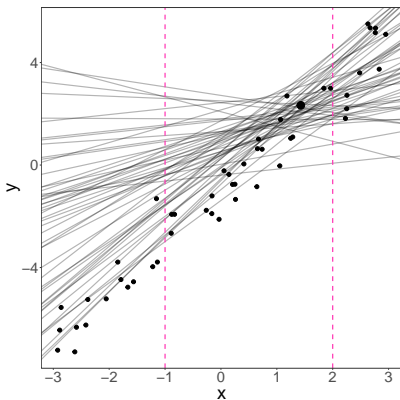
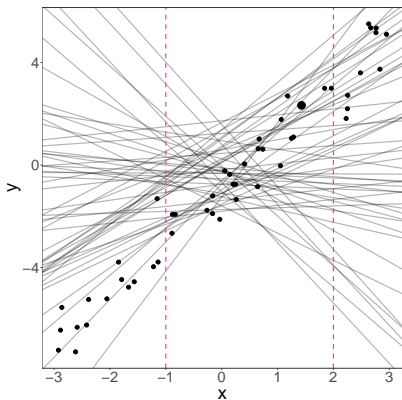
Observando dos puntos juntos...



Comparemos con un *prior* más fuerte...



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



## Observando **diez puntos**



$\mu$  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

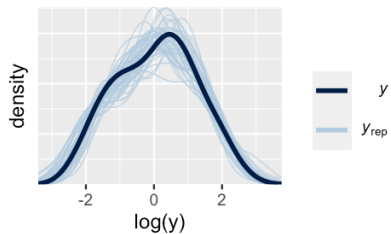
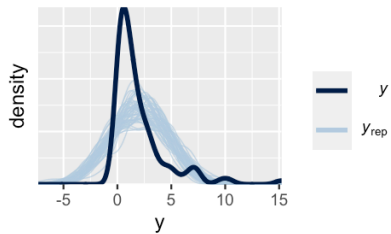
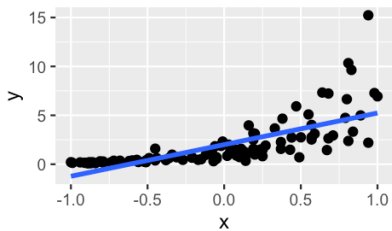
- ▶ Tenemos una distribución de probabilidad para los parámetros. Es decir, tenemos incertidumbre en los valores de los parámetros
- ▶ Tenemos que trabajar con todo el *posterior* (a través de muestras) y no con estimaciones puntuales
- ▶ No confundir predicción de la media (también llamado predictor 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  $\sigma$ : 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)





## 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),

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

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

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



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

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

Que podemos estimar a través de muestras del *posterior*

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

La **deviance** de un modelo es

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

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

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

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

donde los  $\theta_{-i}^{(s)}$  son muestras del *posterior* de  $\theta$  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  $\theta$  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

$$AIC = D + 2p = -2 \text{ lppd} + 2p$$

donde  $p$  es el número de parámetros del modelo y  $D = -2 \text{ lppd}$  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):

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

$\sum_{i=1}^N \mathbb{V}_{\theta} [\log (p(y_i | \theta))]$  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).