

MAT-266: Análisis de Regresión

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM



Horario:

Clases: Lunes y Martes, bloque 1-2 (08:15-09:25 hrs.), salas M201 y M203.

Contacto:

E-mail: felipe.osorios@usm.cl.

Material de clases:

Página del curso (GitHub): <https://github.com/faosorios/Curso-Regresion>

Página personal: <http://fosorios.mat.utfsm.cl/teaching.html#MAT266>

El material también estará disponible en AULA.

Evaluación:

Se realizará **3 Certámenes**.



Criterio de aprobación:

Sea NP el promedio de los Certámenes. Aquellos estudiantes que obtengan NP mayor o igual a 55 y **todos** los certámenes sobre 40, **aprobarán la asignatura** con nota final, $NF = NP$.

Criterio para rendir global:

En caso contrario, y siempre que $NP \geq 45$,¹ los estudiantes podrán rendir el **certamen global** (CG), en cuyo caso la **nota final** (NF) es calculada como sigue:

$$NF = 0.6 \cdot NP + 0.4 \cdot CG.$$

¹Si $NP < 45$ usted ha reprobado la asignatura.



Reglas adicionales

- ▶ Se llevará un **control de asistencia**.
- ▶ Se puede realizar **preguntas** sobre la materia en **cualquier momento**.
- ▶ Los alumnos deben **apagar/silenciar** sus **teléfonos celulares** durante clases.
- ▶ Conversaciones sobre asuntos ajenos a la clase no serán tolerados. Otros estudiantes tiene derecho a **asistir clases en silencio**.
- ▶ Al enviar algún **e-mail al profesor**, identificar el código de la asignatura en el asunto (**MAT266**).
- ▶ **E-mail** será el canal de **comunicación oficial** entre el profesor y los estudiantes.



Reglas: sobre los certámenes

- ▶ Es derecho del estudiante conocer la **pauta de corrección** la que será publicada en la **página web del curso**.
- ▶ Use principalmente **lapiz pasta** (no utilice lapiz rojo).
- ▶ Pedidos de corrección **deben ser argumentados por escrito**.
- ▶ En modalidad online, **Certámenes** deben ser enviados en formato **PDF**.²
- ▶ **Cualquier tipo de fraude** en prueba (copia, uso de WhatsApp, suplantación, etc.) será llevado a **Comisión Universitaria**.

²En un único archivo, orientado en una dirección legible.



Orientaciones de estudio

- ▶ Mantener la frecuencia de estudio de inicio a final del semestre. El ideal es estudiar el contenido luego de cada clase.
- ▶ Estudiar primeramente el contenido dado en clases, buscando apoyo en las referencias bibliográficas.
- ▶ Las referencias son fuentes de ejemplos y ejercicios. Resuelva una buena cantidad de ejercicios. No deje esto para la víspera de la prueba.
- ▶ Buscar las referencias bibliográficas al inicio del semestre, dando preferencia a las principales y complementarias.



- ▶ El requisito formal es [MAT-041: Probabilidad y Estadística](#).³
- ▶ Adicionalmente [*el profesor tiene la muy mala costumbre de..*] usaremos algunas ideas desde [MAT-206: Inferencia Estadística](#).
- ▶ Se asume un conocimiento básico de los siguientes aspectos:
 - ▶ Variables aleatorias.
 - ▶ Convergencia de variables aleatorias.
 - ▶ Manipulación de matrices y vectores aleatorios.

³O sus equivalentes MAT-031 o MAT-042.



1. Preliminares.
2. Inferencia en el modelo de regresión lineal.
3. Análisis de los supuestos del modelo.
4. Identificación del mejor conjunto de regresores.
5. Alternativas a mínimos cuadrados.
6. Tópicos adicionales.

⁴ Este es un curso **fundamental** donde exploramos métodos para abordar el proceso de inferencia en modelos de regresión, **no** es un curso **enfocado** en el análisis de datos.





Hocking, R. (2013).

Methods and Applications of Linear Models: Regression and the analysis of variance, 3rd Edition.

Wiley, New York.



Seber, G.A.F., Lee, A.J. (2007).

Linear Regression Analysis, 2nd Edition.

Wiley, New York.



Weisberg, S. (2013).

Applied Linear Regression, 4th Edition.

Wiley, New York.



"Todos los modelos son errados, pero algunos son útiles."

– George Box.

"Aunque puede parecer una paradoja, toda la ciencia exacta está dominada por la idea de aproximación."

– Bertrand Russell.

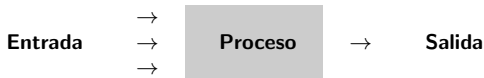
Principio KISS: "Keep It Simple, Stupid."

– Clarence "Kelly" Johnson.



Objetivo del análisis de regresión

Estudiar una variable de **respuesta**, y [asumiendo continua] como función de algunas variables explicativas o **regresores**, x_1, x_2, \dots [pueden ser discretas y/o continuas].



En ocasiones la relación funcional es **conocida** salvo algunos coeficientes (**parámetros**).

Es decir, la relación es gobernada por un **proceso físico** o por leyes bien aceptadas

$$Y \approx f(x_1, \dots, x_p; \theta),$$

en cuyo caso, el interés recae en **estimar el vector de parámetros** $\theta = (\theta_1, \dots, \theta_p)^\top$.



Asumiremos una muestra aleatoria Y_1, \dots, Y_n , tal que

$$Y_i = \mu_i + \epsilon_i, \quad E(\epsilon_i) = 0, \quad i = 1, \dots, n,$$

esto es,

respuesta = parte sistemática + error aleatorio

Idea del modelamiento:

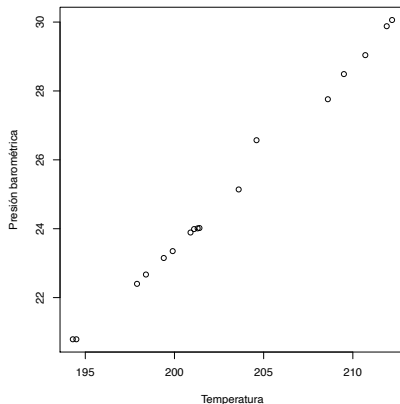
“estructurar” la media como $\mu_i = \mu_i(\beta)$,

para $i = 1, \dots, n$.



Datos de Forbes (1857)⁵

Presión barométrica en pulgadas de mercurio y temperatura de ebullición del agua en grados Fahrenheit para 17 diferentes altitudes.



⁵Transactions of the Royal Society of Edinburgh **21**, 235-243.

Para describir la relación entre la temperatura y la media de la presión barométrica, podemos considerar

$$\mu = \beta_0 + \beta_1 x,$$

note que

$$\mu = \mathbf{x}^\top \boldsymbol{\beta}, \quad \mathbf{x} = (1, x)^\top, \quad \boldsymbol{\beta} = (\beta_0, \beta_1)^\top,$$

y $\mu = \mathbf{x}^\top \boldsymbol{\beta}$ se denomina **predicador lineal**.

El conjunto de datos consiste del **vector de respuestas**.

$$\mathbf{Y} = (Y_1, \dots, Y_n)^\top,$$

y una **matriz de diseño** $n \times 2$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}.$$

Convención:

Todos los vectores siempre serán **columna**.



Se dice que un vector aleatorio \mathbf{Y} sigue un **modelo lineal general** si,

$$\mathbb{E}(\mathbf{Y}) = \sum_{j=1}^p \beta_j \mathbf{x}_j = \mathbf{X}\boldsymbol{\beta},$$

$$\text{Cov}(\mathbf{Y}) = \sum_{t=1}^k \phi_t \mathbf{V}_t = \mathbf{V}(\boldsymbol{\phi}),$$

y decimos que el modelo es **lineal simple** si $\text{Cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}$.

Observación:

Es usual hacer la **suposición distribucional**:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{V}(\boldsymbol{\phi})).$$



Para los datos de Forbes podemos considerar un **modelo lineal** (simple) definido como

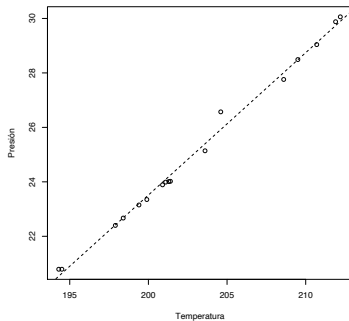
$$\mathbf{Y} \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

donde

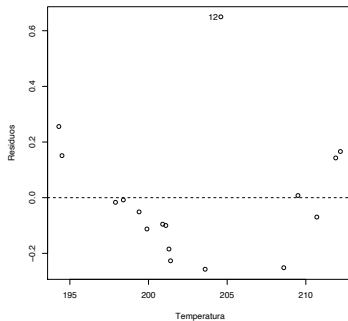
$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 x_i, \quad \text{var}(Y_i) = \sigma^2, \quad \text{Cov}(Y_i, Y_j) = 0,$$

para $i, j = 1, \dots, n$.





(a) recta ajustada



(b) residuos vs. ajuste

Ahora consideramos el modelo

$$100 \times \log_{10}(\text{Presión}_i) = \alpha + \beta \text{Temperatura}_i + \epsilon_i,$$

para $i = 1, \dots, n$.

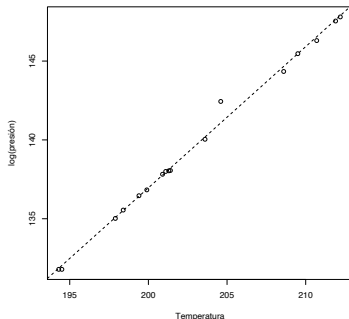
Se obtuvo (usando función `lm` de **R**)

$$\hat{\beta} = (-42.1378, 0.8955)^\top \quad \text{y} \quad s^2 = 0.1438$$

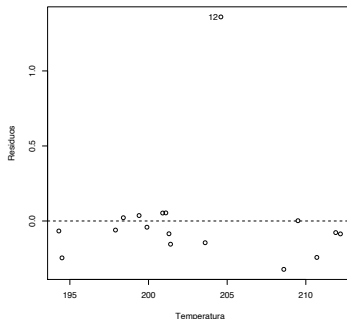
Además, $R^2 = 0.9950$.



Recta de regresión y gráfico de residuos para los datos de Forbes⁶.



(a) recta ajustada



(b) residuos vs. ajuste

⁶datos transformados

Datos de Huber (1981)⁷

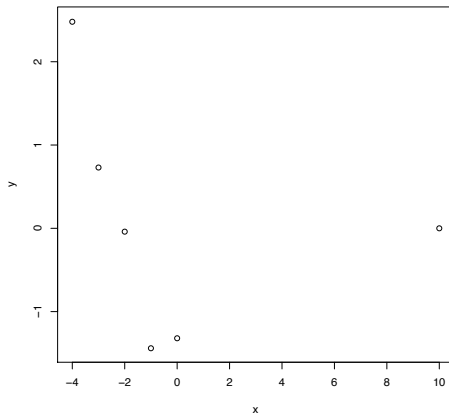
Considere el conjunto de **datos hipotéticos** de Huber.

Y	2.48	0.73	-0.04	-1.44	-1.32	0.00
x	-4	-3	-2	-1	0	10

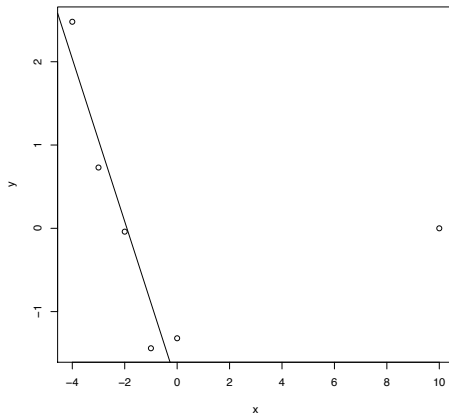
⁷*Robust Statistics*. Wiley, New York



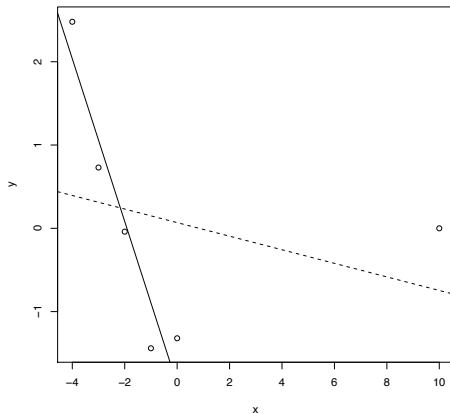
Diagrama de dispersión para los datos de Huber.



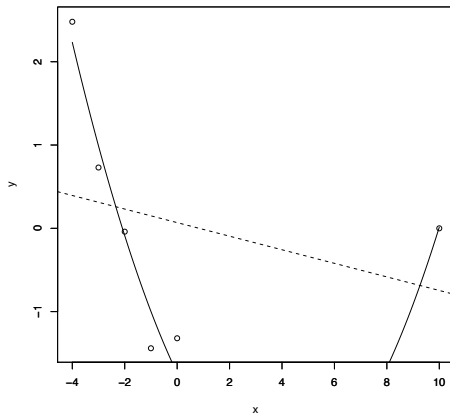
¿Qué opina de la recta de regresión?



¿Y ahora?

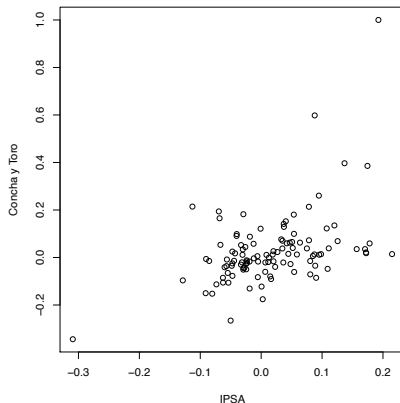


¿Cuál modelo prefiere?



Datos de Concha y Toro (Osorio y Galea, 2006)⁸

Rentabilidades mensuales de Concha y Toro vs. IPSA, ajustados por bonos de interés del Banco Central entre marzo/1990 a abril/1999.



⁸Statistical Papers 47, 31-38

Modelo CAPM (Valoración de Activos de Capital), Sharpe (1964)⁹

$$E(r) = r_f + \beta(E(r_m) - r_f),$$

usando datos observados, podemos escribir

$$R_t = \alpha + \beta \times IPSA_t + \epsilon, \quad t = 1, \dots, T.$$

Características del problema:

- ▶ Relación **lineal** entre las variables.
- ▶ Posibles periodos de **alta volatilidad**.

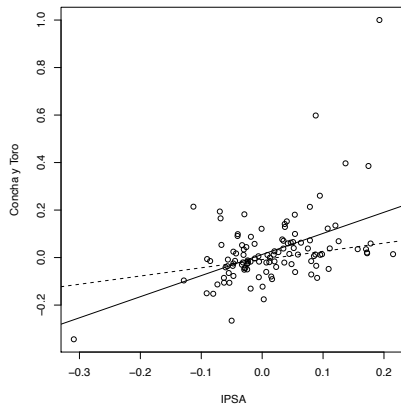
Hipótesis de interés:

- ▶ $H_0 : \beta > 1$ (**Amante del riesgo**).
- ▶ $H_0 : \beta = 1$ (**Neutral al riesgo**).
- ▶ $H_0 : \beta < 1$ (**Averso al riesgo**).

⁹ Journal of Finance **19**, 425-442



Datos de Concha y Toro



Ajuste usando errores **normales** (—) y **Cauchy** (---).



Estudio experimental relacionando la emisión de calor durante la producción y endurecimiento de 13 muestras de cementos Portland. Woods et al. (1932) consideraron cuatro compuestos para los clinkers desde los que se produce el cemento.

La respuesta (Y) es la emisión de calor después de 180 días de curado, medido en calorías por gramo de cemento. Los regresores son los porcentajes de los cuatro compuestos: aluminato tricálcico (X_1), silicato tricálcico (X_2), ferrito aluminato tetracálcico (X_3) y silicato dicálcico (X_4).

¹⁰Industrial and Engineering Chemistry 24, 1207-1214.



Cemento Portland (Woods, Steinour y Starke, 1932)

Y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Observación:

En efecto, existe una **relación lineal aproximada**, pues $x_1 + x_2 + x_3 + x_4 \approx 100$.



Mediciones de Radiación Solar (Davies y Gather, 1993)¹¹

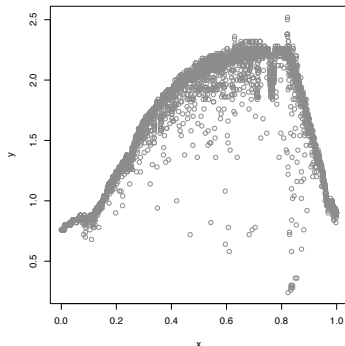


¹¹ Journal of the American Statistical Association **88**, 782-801.



Mediciones de Radiación Solar (Davies y Gather, 1993)

- ▶ Mediciones de la radiación del sol tomadas durante el vuelo de un globo meteorológico.
- ▶ Una gran cantidad de observaciones son consideradas outliers.
- ▶ Se ha sugerido usar métodos robustos (Kovac y Silverman, 2000;¹⁰ Lee y Oh, 2007;¹¹ Tharmaratnam et al., 2010¹²).
- ▶ Diagnóstico de influencia en splines penalizados (Osorio, 2016)¹³.



¹⁰ Journal of the American Statistical Association **95**, 172-183.

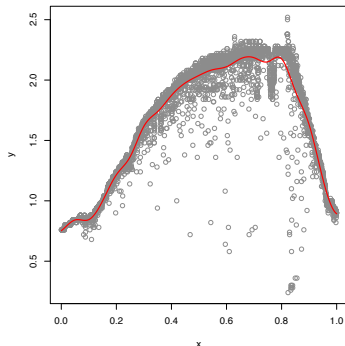
¹¹ Computational Statistics **22**, 159-171.

¹² Journal of Computational and Graphical Statistics **19**, 609-625.

¹³ Annals of the Institute of Statistical Mathematics **68**, 589-619.

Mediciones de Radiación Solar (Davies y Gather, 1993)

- ▶ Mediciones de la radiación del sol tomadas durante el vuelo de un globo meteorológico.
- ▶ Una gran cantidad de observaciones son consideradas outliers.
- ▶ Se ha sugerido usar métodos robustos (Kovac y Silverman, 2000;¹⁰ Lee y Oh, 2007;¹¹ Tharmaratnam et al., 2010¹²).
- ▶ Diagnóstico de influencia en splines penalizados (Osorio, 2016)¹³.



¹⁰ Journal of the American Statistical Association **95**, 172-183.

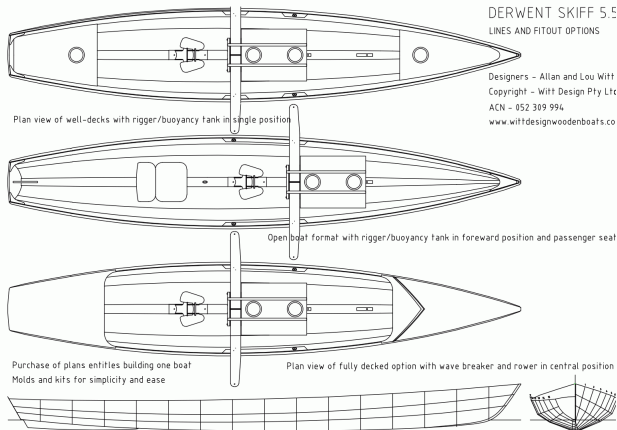
¹¹ Computational Statistics **22**, 159-171.

¹² Journal of Computational and Graphical Statistics **19**, 609-625.

¹³ Annals of the Institute of Statistical Mathematics **68**, 589-619.



Métodos Spline: Diseño técnico





¹⁴ Imágenes extraídas desde página web del Prof. [Carl de Boor](#), University Wisconsin-Madison

Considere el modelo de **regresión no paramétrica**

$$Y_i = g(t_i) + \epsilon_i, \quad i = 1, \dots, n,$$

donde las respuestas Y_i son medidas en los puntos de diseño t_i , g es una función suave definida en $[a, b]$ y $\{\epsilon_i\}$ representa disturbios aleatorios.

Típicamente, \hat{g}_λ puede ser obtenido como solución de un problema de **mínimos cuadrados penalizados (PLS)**

$$S(\lambda) = \sum_{i=1}^n \{Y_i - g(t_i)\}^2 + \lambda \int_a^b \{g''(t)\}^2 dt, \quad \lambda > 0,$$

sobre la clase de todas las funciones dos veces diferenciables.



El criterio PLS puede ser adaptado para [ajuste de curvas usando P-splines](#):

$$S(\lambda) = (\mathbf{Y} - \mathbf{B}\mathbf{a})^\top (\mathbf{Y} - \mathbf{B}\mathbf{a}) + \lambda \mathbf{a}^\top \mathbf{K}^\top \mathbf{K} \mathbf{a}, \quad \lambda > 0,$$

donde $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, $\mathbf{B} = (B_j(t_i))$ es matriz $n \times p$, $\mathbf{K}^\top \mathbf{K}$ es una representación matricial de la penalidad descrita por Eilers y Marx (1996) y

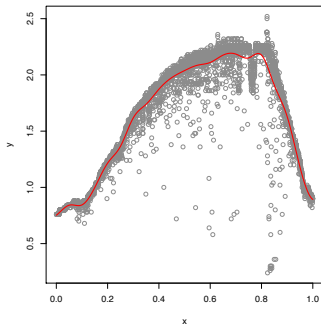
$$g(t) = \sum_{j=1}^p a_j B_j(t),$$

con $\mathbf{a} = (a_1, \dots, a_p)^\top$ y p el número de funciones base conocidas $B_1(t), \dots, B_p(t)$. Una elección común para las funciones base es [B-splines](#).

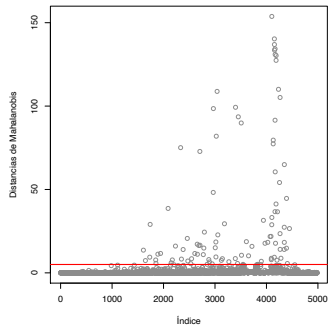
En este contexto λ denota un [parámetro de suavizamiento](#).



Modelo ajustado y distancias de Mahalanobis

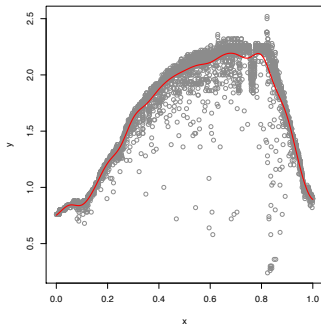


(a) curva ajustada

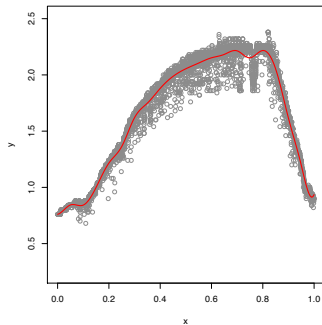


(b) distancias de Mahalanobis

Modelo ajustado con todos los datos y “outliers” removidos

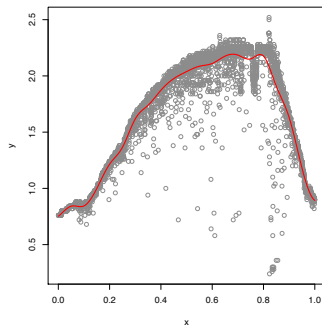


(a) curva ajustada

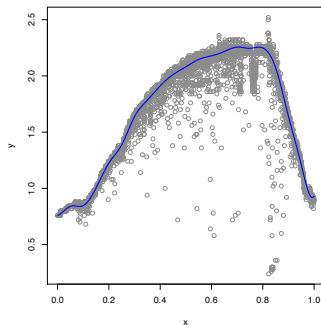


(b) eliminando 'outliers'

Modelo ajustado usando distribuciones con colas pesadas



(a) curva ajustada



(b) ajuste robusto

Modelos lineales son los *bloques de construcción* para metodologías más complejas, tales como:

- ▶ Modelos lineales generalizados.
- ▶ Modelos no lineales.
- ▶ Regresión multivariada.
- ▶ Datos longitudinales, GMANOVA.
- ▶ Regresión semiparamétrica.
- ▶ Modelos con efectos mixtos.

