Aprendizaje Supervisado

Enver G. Tarazona

2019-12-07

Introducción

Objetivos de la unidad

- Aprendizaje de máquinas vs. aprendizaje estadístico: ejemplos
- Introducir notación relevante y terminología.
- Estimar f (regresión, clasificación): precisión en la predicción vs. interpretabilidad del modelo
- Balance del Sesgo-Varianza

¿Qué es aprendizaje estadístico?

Aprendizaje estadístico es el proceso de aprendizaje a partir de los datos.

Aplicando *métodos estadísticos* en un *conjunto de datos* (llamado el *conjunto de entrenamiento*), el objetivo es *extraer conclusiones* acerca de las relaciones entre las variables (*inferencia*) o *encontrar una función predictiva* (*predicción*) para nuevas observaciones.

El aprendizaje estadístico juega un rol muy importante en muchas áreas del conocimiento como la ciencia, finanzas y la industria.

Ejemplos de problemas de aprendizaje

Economía:

Para predecir el precio de una acción dentro de 3 meses, en función de las medidas de rendimiento de la empresa y los datos económicos. Aquí la variable de respuesta es cuantitativa (precio).

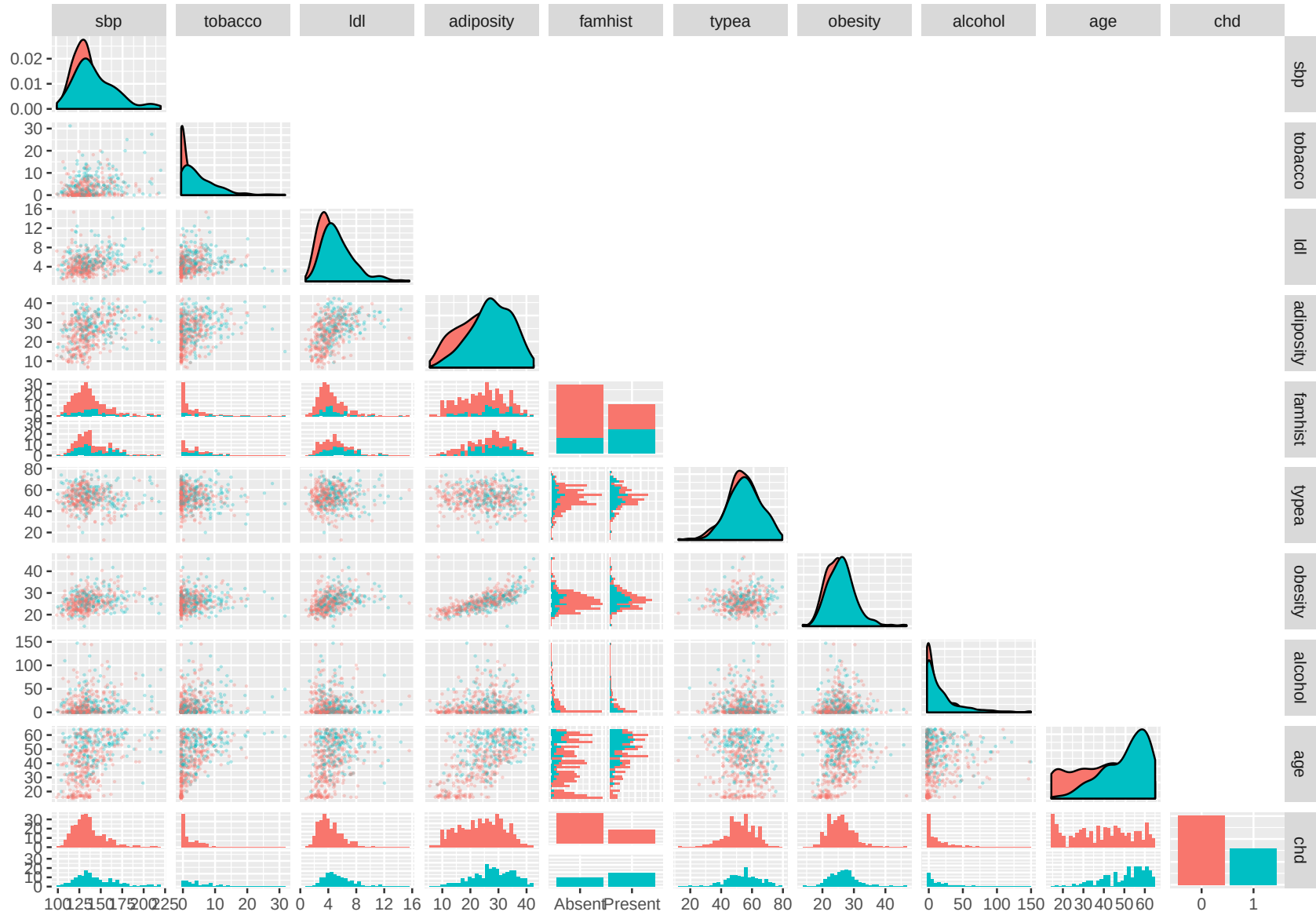
Medicina 1:

Identificar los factores de riesgo para desarrollar diabetes en función de la dieta, la actividad física, los antecedentes familiares y las mediciones corporales. Aquí el objetivo es hacer inferencia de los datos, es decir, encontrar relaciones subyacentes entre las variables.

Medicina 2:

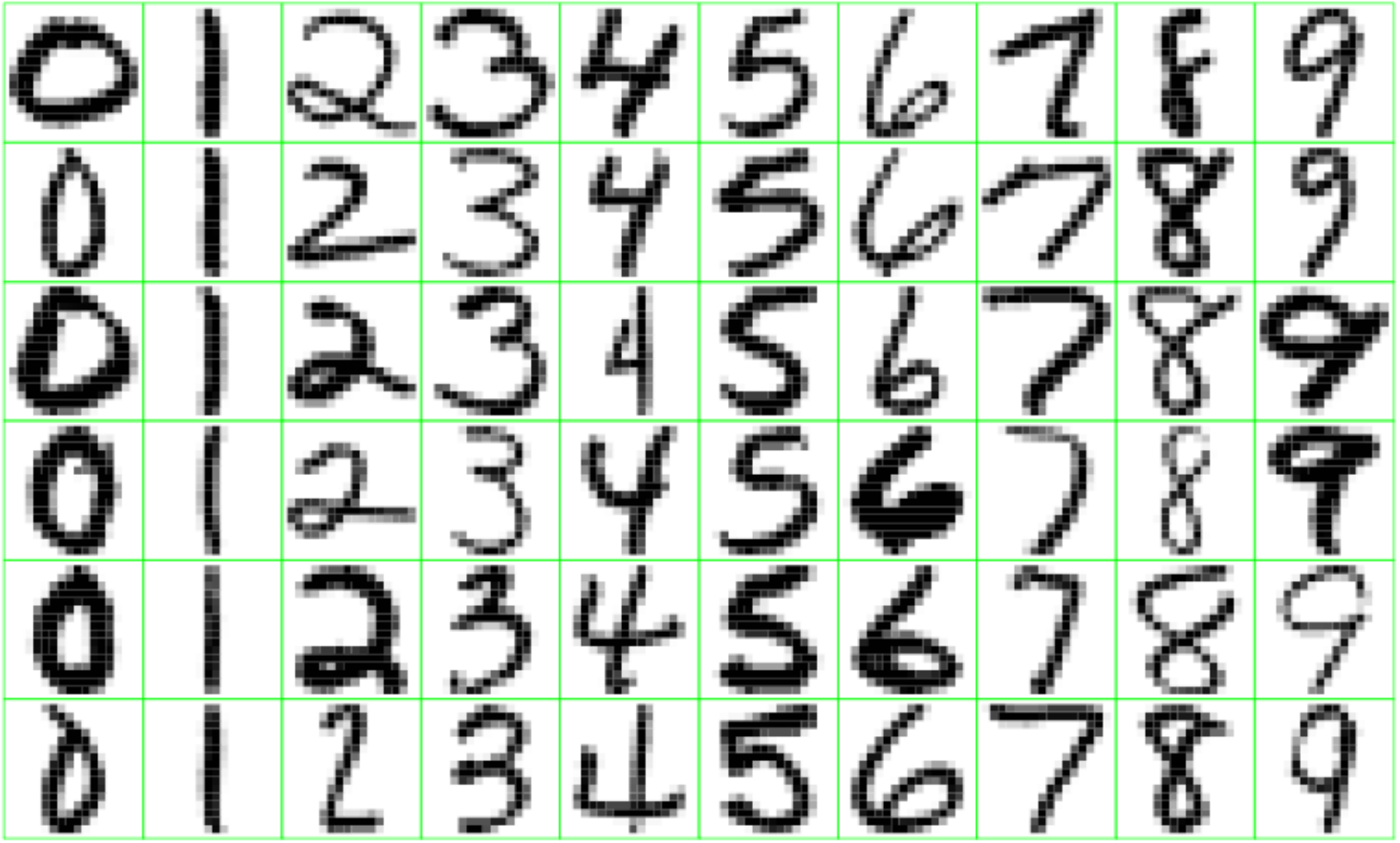
Predecir si alguien sufrirá un ataque al corazón sobre la base de mediciones demográficas, dietéticas y clínicas. Aquí el resultado es binario.(si,no) con variables de entrada tanto cualitativas como cuantitativas.

Datos de Sudáfrica sobre enfermedades cardíacas: 462 observaciones y 10 variables.



Reconocimiento de dígitos escritos a mano:

Identificar los números en un código postal (ZIP) escrito a mano, a partir de una imagen digitalizada. Este es un problema de clasificación, donde la variable de respuesta es categórica con las clases $\{0, 1, 2, ..., 9\}$ y la tarea es predecir correctamente la pertenencia a la clase.



Clasificación de correo electrónico (detección de spam):

El objetivo es construir un filtro de spam. Este filtro puede basarse en las frecuencias de palabras y caracteres en los correos electrónicos. La siguiente tabla muestra el porcentaje promedio de palabras o caracteres en un mensaje de correo electrónico, basado en 4601 correos electrónicos de los cuales 1813 se clasificaron como spam.

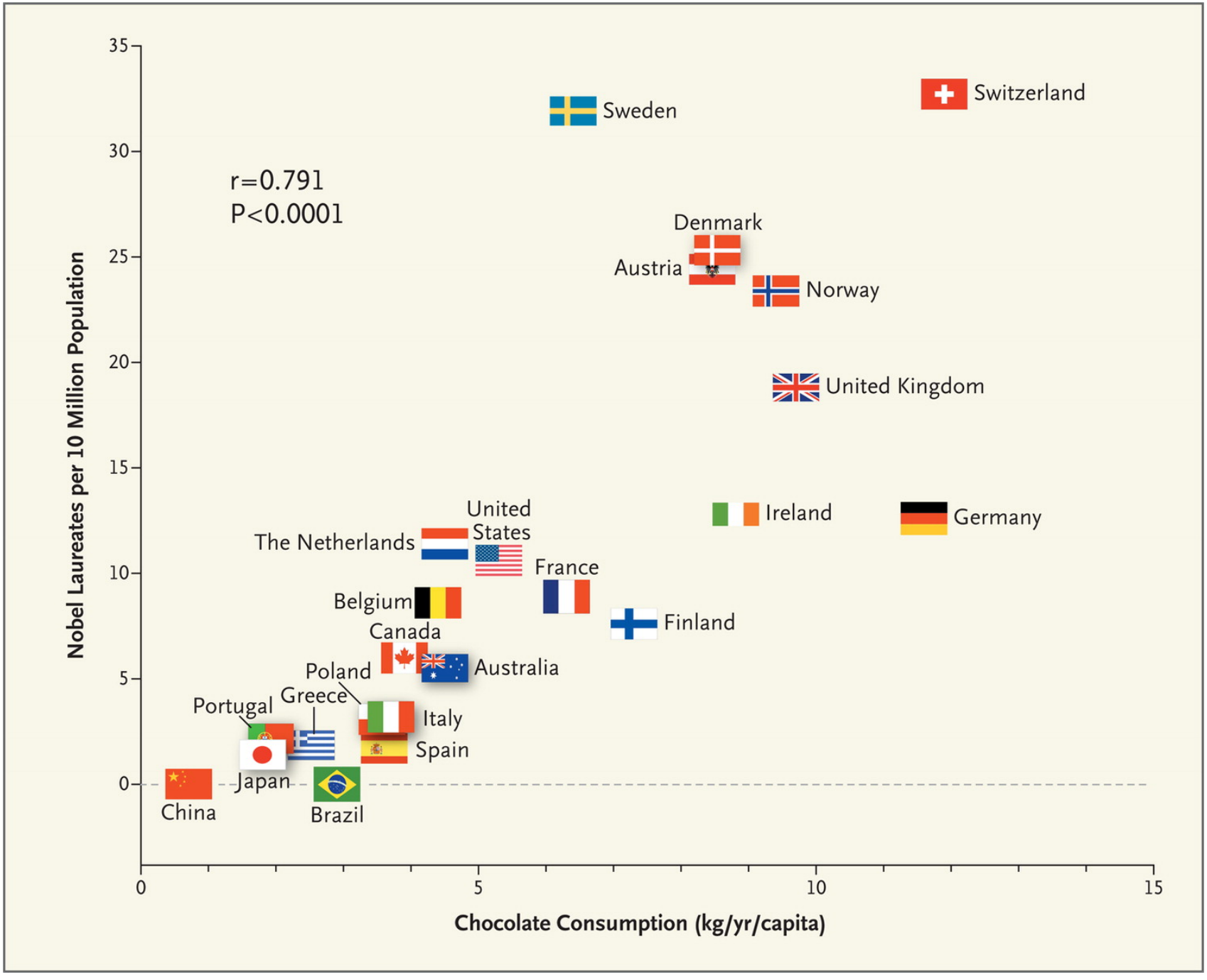
	you	free	george	!	$	edu
not spam	1.27	0.07	1.27	0.11	0.01	0.29
spam	2.26	0.52	0.00	0.51	0.17	0.01

¿Qué hace a un ganador de un Premio Nobel?

¿Perseverancia, suerte, mentores expertos o simplemente consumo de chocolate? Un artículo publicado en el New England Journal of Medicine concluyó con lo siguiente:

> *El consumo de chocolate mejora la función cognitiva, que es una condición sine qua non para ganar el Premio Nobel, y se correlaciona estrechamente con el número de premios Nobel en cada país. Queda por determinar si el consumo de chocolate es el mecanismo subyacente para la asociación observada con una función cognitiva mejorada.*

La figura muestra las correlaciones entre el consumo anual de chocolate per cápita de un país y el número de premios Nobel por cada 10 millones de habitantes.
Puede leer el artículo aqui y una revisión informal del artículo aqui.
Esperemos que no nos quedemos sin chocolate en 2020


r=0.791
P<0.0001
Sweden
Switzerland
Denmark
Austria
Norway
United Kingdom
Ireland
Germany
United States
The Netherlands
France
Belgium
Finland
Canada
Australia
Poland
Greece
Portugal
Italy
Spain
Japan
China
Brazil
Nobel Laureates per 10 Million Population
Chocolate Consumption (kg/yr/capita)
0
5
10
15
20
25
30
35

P: ¿Hubieron objetivos y elementos subyacentes comunes en estos ejemplos de aprendizaje?

- Predecir el precio de una acción dentro de 3 meses a partir de las medidas de desempeño de la compañía y los datos económicos.
- Identificar los factores de riesgo para desarrollar diabetes en función de la dieta, la actividad física, los antecedentes familiares y las mediciones corporales.
- El objetivo es construir un filtro de spam.
- Prededir si alguien sufrirá un ataque cardíaco sobre la base de mediciones demográficas, dietéticas y clínicas.
- Identificar los números en un código postal escrito a mano, a partir de una imagen digitalizada.
- ¿Qué hace que un ganador del Premio Nobel? ¿Perseverancia, suerte, mentores expertos o simplemente consumo de chocolate?

R:

Sí, el objetivo era la explicación o la predicción, algunas variables de salida eran cualitativas (categóricas) y otras cuantitativas.

¿Cuál es el objetivo del aprendizaje supervisado?

Asumir:

- Se observa una variable respuesta *cuantitativa* Y y
- p predictores diferentes $x_1, x_2, ..., x_p$.

Se asume que existe una función f que relaciona la respuesta con las variables predictoras:

$$Y = f(x) + \varepsilon,$$

donde ε es un término aleatorio (error) con media 0 e independiente de x.

Existen dos principales motivos par estimar f: *predicción e inferencia*

Predicción

Basados en los datos observados, el objetivo es construir un modelo que con la mayor precisión posible pueda predecir una respuesta dadas nuevas observaciones de las covariables:

$$\widehat{Y} = \widehat{f}(x).$$

Donde $\widehat{f}$ representa la estimación de f y $\widehat{Y}$ representa la predicción de Y. En este ámbito, nuestra estimación de la función f es tratada como una **caja negra y no es de interés**. Nuestro enfoque está en la predicción de Y, por lo tanto, la precisión de la predicción es importante.

Existen dos términos que influyen en la precisión de $\hat{Y}$ como predicción de Y: el error reducible y el error irreducible.

- El *error reducible* tiene que ver con nuestra estimación $\hat{f}$ de f. Este error puede ser reducirse usando la técnica *apropiada* de aprendizaje supervisado.
- El *error irreducible* proviene del término del error ε y no puede reducirse mejorando f. Esto está relacionado con las cantidades no observadas que influyen en la respuesta y posiblemente la aleatoriedad de la situación.

P: Si hubiera una relación *determinista* entre la respuesta y un conjunto de predictores, ¿habría un error tanto reducible como irreducible?

R:

Si conocemos todos los predictores y la conexión (determinista) a la respuesta, y no se agrega ningún error aleatorio, entonces no tendremos un error irreducible. Si hay una relación determinista, pero no conocemos todos los valores predictores, los predictores no observados nos darán un error irreducible.

Entonces, muy raramente habrá solo un error reducible presente.

Inferencia

En función de los datos observados, el objetivo es *comprender* cómo la variable de respuesta se ve afectada por los diversos predictores (covariables).

En esta situación no usaremos nuestra función estimada $\hat{f}$ para realizar predicciones sino para *entender* como Y cambia como función de $x_1, x_2, ..., x_p$.

Nuestro principal interés es conocer la *forma exacta* de $\hat{f}$:

- ¿Qué predictores están asociados con la respuesta?
- ¿Cuál es la relación entre la respuesta y cada predictor?
- ¿Puede la relación ser lineal o se necesita un modelo más complejo?

Diferencias entre aprendizaje estadístico y aprendizaje automático

Hay mucha superposición entre el aprendizaje estadístico y el aprendizaje automático: el objetivo común es aprender de los datos. Específicamente, encontrar una función f que mejor relacione las variables de entrada x con una variable respuesta Y: $Y = f(x)$. La función f nos permitirá realizar una predicción futura de Y, dadas nuevas observaciones de las variables de entrada x's.

- El aprendizaje automático surgió como un subcampo de la inteligencia artificial y generalmente tiene un mayor énfasis en *aplicaciones a gran escala y precisión de la predicción*, la forma y tipo de f en sí misma (en general) no es interesante. Además, los algoritmos son de primordial importancia.
- El aprendizaje estadístico surgió como un subcampo de estadística con un mayor enfoque en la *interpretabilidad* del modelo más que en una predicción (del tipo de caja negra). Además, está más enfocado en los modelos y métodos que los algoritmos.

Términos usados

Aprendizaje estadístico	Aprendizaje automático
modelo	red, gráfico, mapeo
ajustar, estimar	aprender
covariables, entradas, variables independientes, predictoras	atributos, predictoras
variable respuesta, de salida, dependiente	salida, objetivo
conjunto de datos	datos de entrenamiento

Observación: no es una lista exhaustiva, y muchos términos se usan de la misma manera en ambos campos.

Regresión y Clasificación

Regresión predice un valor numérico.
Ejemplo: Predecir la ganancia dada la cantidad de dinero gastado en publicidad.
Clasificación predice la clase de pertenencia.
Ejemplo: dada la presión arterial, el peso y la proporción de cadera predicen si un paciente sufre de diabetes (sí / no).

P:

Dé un ejemplo de una regresión y un problema de clasificación (problema práctico con el conjunto de datos disponible) que le gustaría estudiar en este curso.

R:

Aprendizaje Supervisado

Nuestro conjunto de datos (datos de entrenamiento) consiste de n mediciones de la variable respuesta Y y p covariables x:

$$(y_1, x_{11}, x_{12}, \dots, x_{1p}), (y_2, x_{21}, \dots, x_{2p}), \dots, (y_n, x_{n1}, x_{n2}, \dots, x_{np}).$$

Objetivo:

- hacer predicciones precisas para nuevas observaciones,
- entender qué entradas afectan a las salidas, y cómo, (no se verá en este curso) y
- para evaluar la calidad de las predicciones e inferencia (no se verá en este curso). Se llama aprendizaje supervisado porque la variable de respuesta *supervisa nuestro análisis*.

Métodos y Modelos

Métodos paramétricos

Los métodos paramétricos se basan en un supuesto sobre la forma y estructura de la función f.
El modelo de regresión lineal múltiple es un ejemplo de modelo paramétrico. Suponemos aquí que la variable de respuesta es una combinación lineal de las covariables con algo de ruido adicional.

$$f(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon.$$

Al hacer esta suposición, la tarea se simplifica en encontrar estimaciones de $p + 1$ coeficientes $\beta_0, \beta_1, .., \beta_p$. Para hacer esto, usamos los datos de entrenamiento para ajustar el modelo, de modo que

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p.$$

La estimación de un modelo paramétrico se divide en dos etapas:

1. Seleccionar una forma para la función f.
2. Estimar los parámetros desconocidos de f usando el conjunto de entrenamiento.

Métodos no paramétricos

Los métodos no paramétricos buscan una estimación de f que se acerca a los datos, pero sin hacer suposiciones explícitas sobre la forma de la función f.

El algoritmo de los K-vecinos más cercanos es un ejemplo de un modelo no paramétrico. Utilizado mayormente en clasificación, este algoritmo predice una membresía de clase para una nueva observación al hacer un voto mayoritario basado en las K observaciones más cercanas.

P: ¿Cuáles son las ventajas y desventajas de los métodos paramétricos y no paramétricos?

Sugerencias: interpretabilidad, cantidad de datos necesarios, complejidad, suposiciones hechas, precisión de predicción, complejidad computacional, sobre/sub ajuste.

R: Métodos paramétricos

Ventajas	Desventajas
Simple de usar y fácil de entender.	La función f está restringida a la forma especificada.
Requiere pocos datos de entrenamiento	La forma de la función f que fue asumida en general no coincidirá con la función verdadera, lo que puede dar una estimación pobre.
Computacionalmente barato	Complejidad limitada

R: Métodos no paramétricos

Ventajas	Desventajas
Flexible: se puede adaptar una gran cantidad de formas funcionales	Puede sobreajustar los datos
No se hacen suposiciones fuertes sobre la función subyacente	Computacionalmente más caro ya que se deben estimar más parámetros
A menudo puede dar buenas predicciones	Se requieren muchos datos para estimar f (compleja).

Precisión de la predicción vs. interpretabilidad

(nos estamos preparando para hablar sobre el balance sesgo/varianza)

Los métodos **inflexibles**, o rígidos, son métodos que tienen fuertes restricciones en la forma de f.

Ejemplos:

- Regresión lineal
- Análisis discriminante lineal
- Selección por subconjuntos y Lasso

Los métodos **flexibles** tienen menos restricciones sobre la forma de f.

Ejemplos:

- Clasificación KNN, regresión KNN, splines de suavizamiento
- Bagging y boosting, máquinas de soporte vectorial
- Redes neuronales

La elección de un método flexible o inflexible depende del objetivo en mente.

Si el objetivo es la inferencia, se preferirá un modelo inflexible, que sea fácil de entender. Por otro lado, si queremos hacer predicciones tan precisas como sea posible, no nos preocupa la forma de f. Se puede elegir un método flexible, a costa de la interpretación del modelo, y tratamos f como una caja negra.

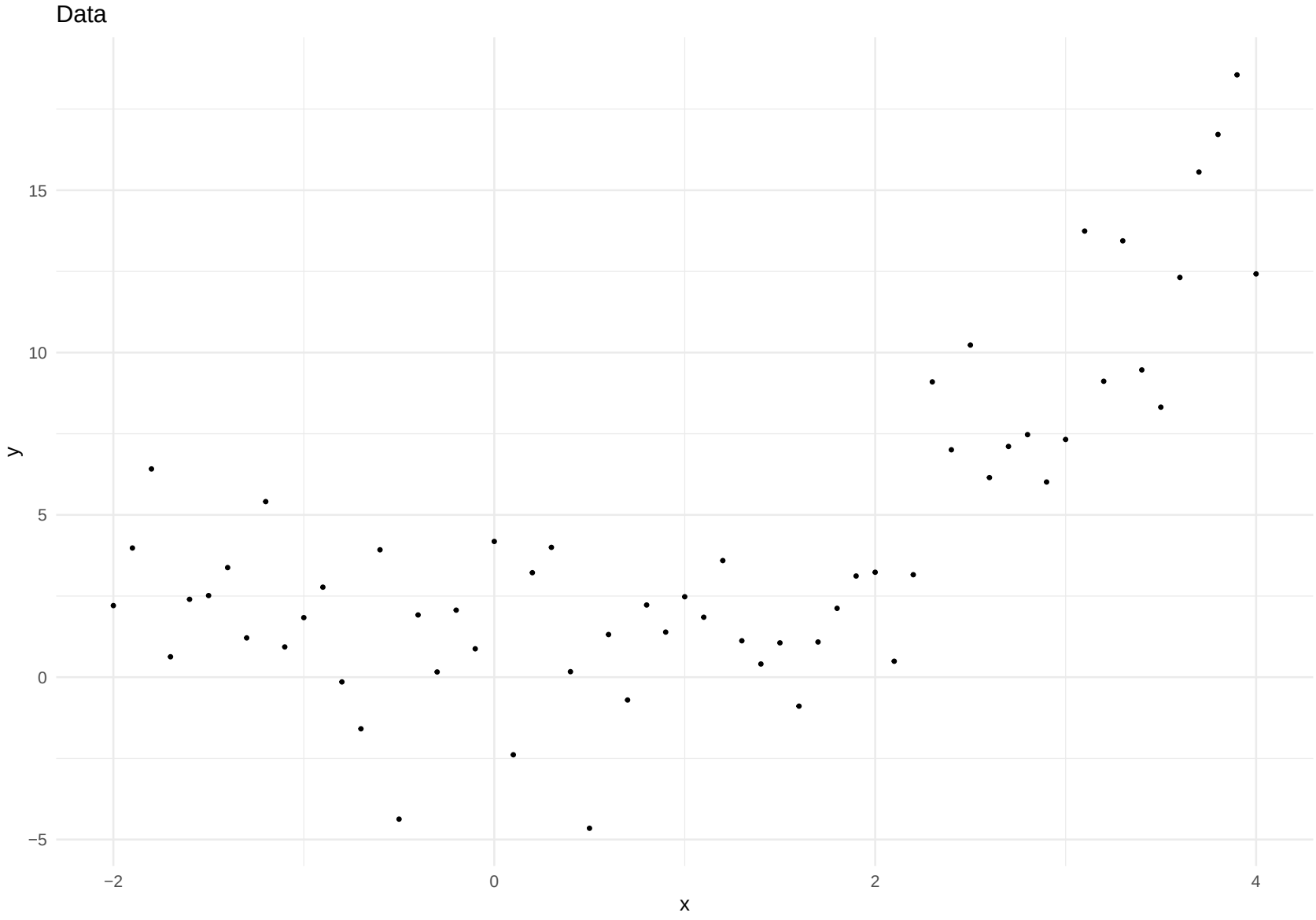
Sobreajuste ocurre cuando la función estimada f se ajusta demasiado a los puntos de datos observados.

Subajuste ocurre cuando la función estimada f es demasiado rígida para capturar la estructura subyacente de los datos.

Ilustramos esto con un ejemplo usando regresión polinomial.

Ejemplo de Regresión Polinomial

Consideremos la covariable x observada en la cuadrícula de la recta de los reales de -2 a 4, igualmente espaciada en 0.1, dando $n = 61$ observaciones.



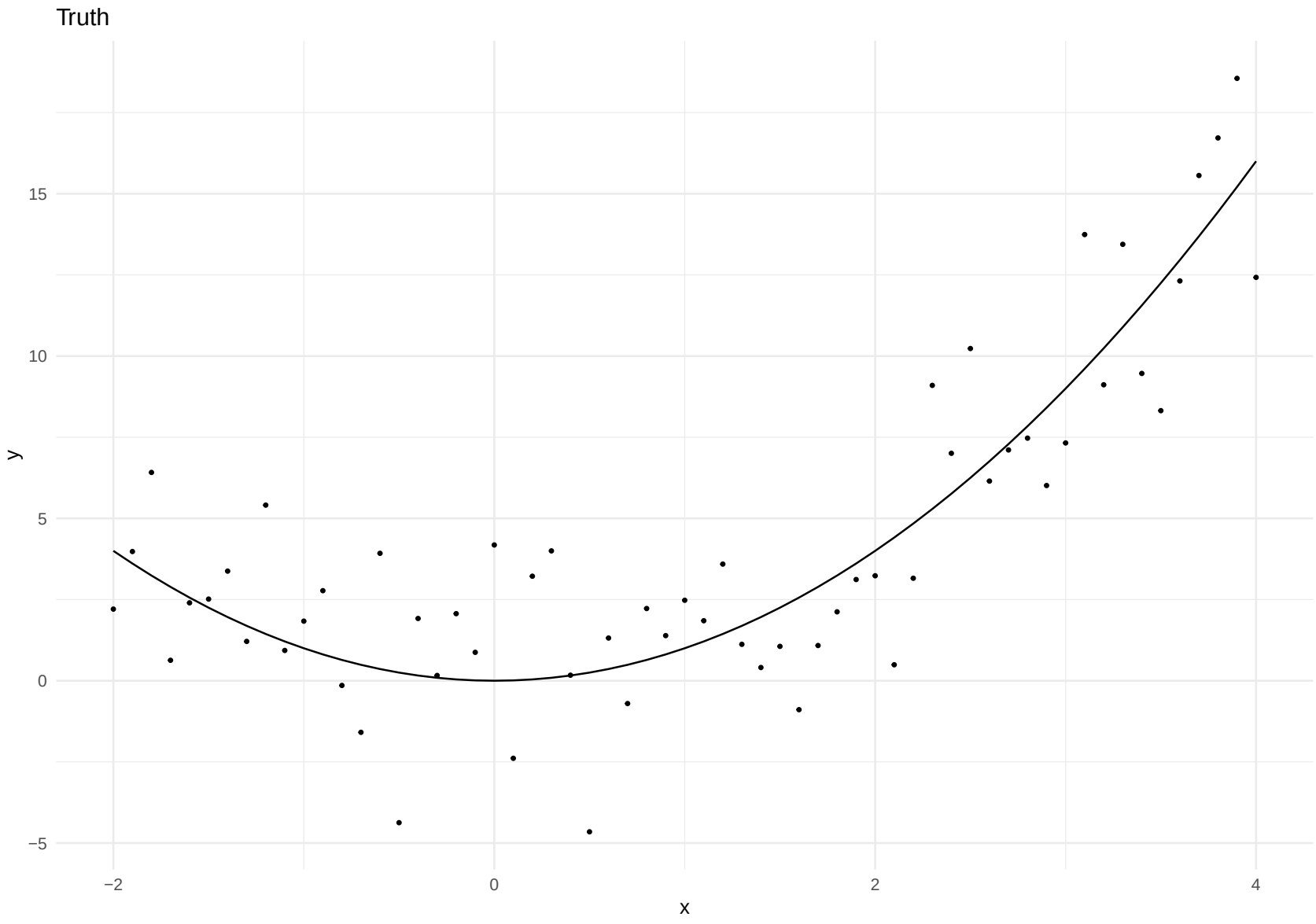
Asumiremos que la relación teórica entre la respuesta Y y la covariable x:

$$Y = x^2 + \varepsilon$$

Donde ε es llamado el término de error (o ruido), y es simulado de una distribución normal con media 0 y desviación estándar 2.

Llamaremos a $Y = x^2$ la *relación verdadera*.

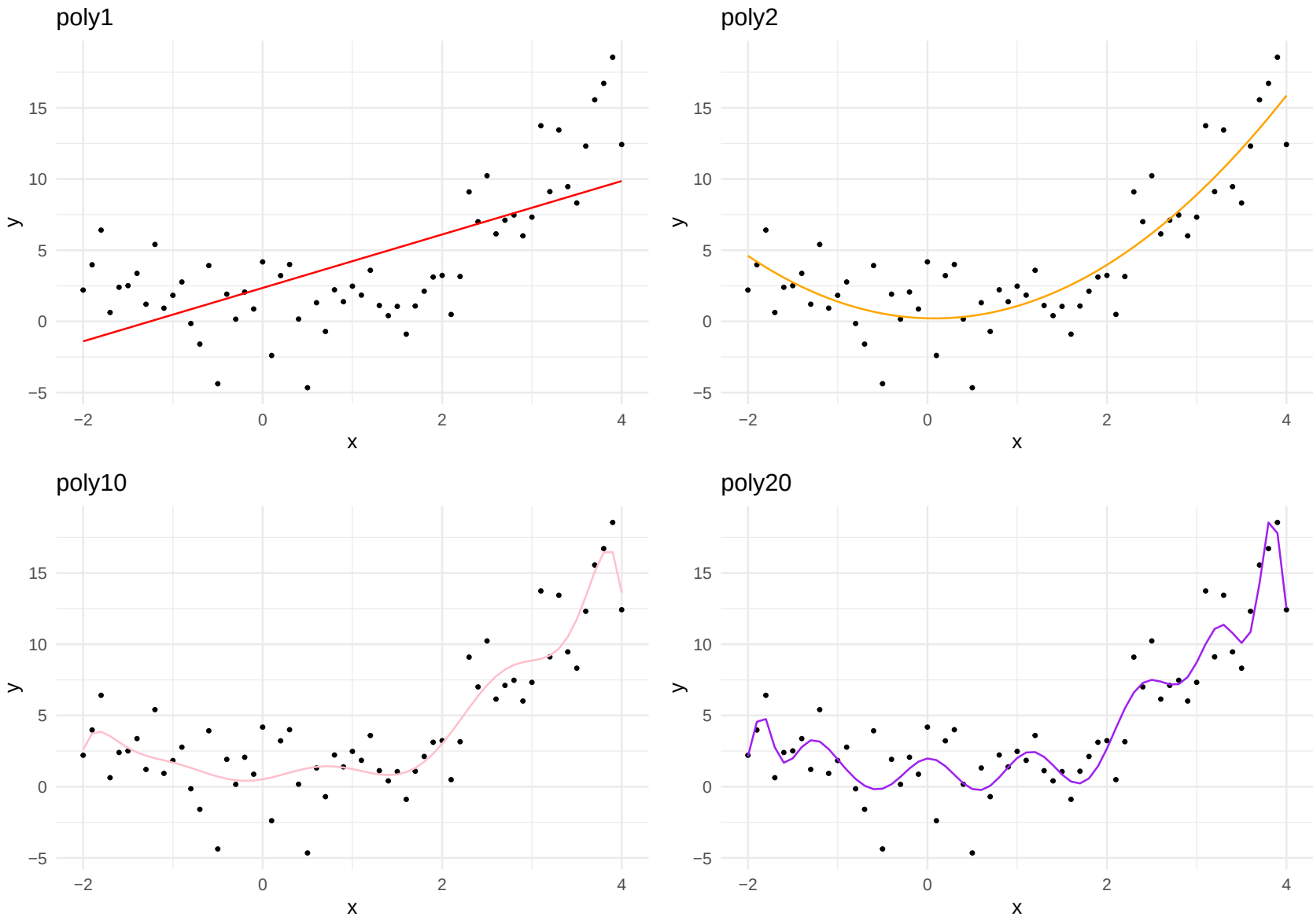
El error agregado se usa como un sustituto de todas las variables no observadas que no están en nuestra ecuación, pero que pueden influir en Y. Esto significa que no estamos viendo una mera relación *determinística* entre x y Y, sino una que permita la aleatoriedad.


Truth
15
10
5
0
−5
y
−2
0
2
4
x

A continuación, queremos ajustar una función a las observaciones *sin conocer* la verdadera relación, y probamos diferentes funciones polinómicas paramétricas.

- [poly1: izquierda superior]: La linea roja muestra un modelo lineal simple de la forma $\beta_0 + \beta_1 x$ ajustada a las observaciones. Esta línea claramente *subajusta* los datos. Vemos que esta función no puede capturar esa naturaleza cuadrática de los datos.
- [poly2: derecha superior]: La línea naranja muestra un ajuste polinómico cuadrático a los datos, de la forma $\beta_0 + \beta_1 x + \beta_2 x^2$. Vemos que esta función se ajusta bien y se ve casi idénticamente como la verdadera función.

- [poly10: inferior izquierda]: La línea rosada muestra un polinomio de grado 10 ajustado a los datos, de la forma $\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{10} x^{10}$. La función captura el ruido en lugar de la estructura subyacente de los datos. La función *sobreajusta* los datos.
- [poly20: inferior derecha]: La línea púrpura muestra un polinomio de grado 20 ajustado a los datos, de la forma $\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{20} x^{20}$. La función captura el ruido en lugar de la estructura subyacente de los datos. La función *sobreajusta* los datos.


poly1
poly2
poly10
poly20
y
x
15
10
5
0
−5
−2
0
2
4

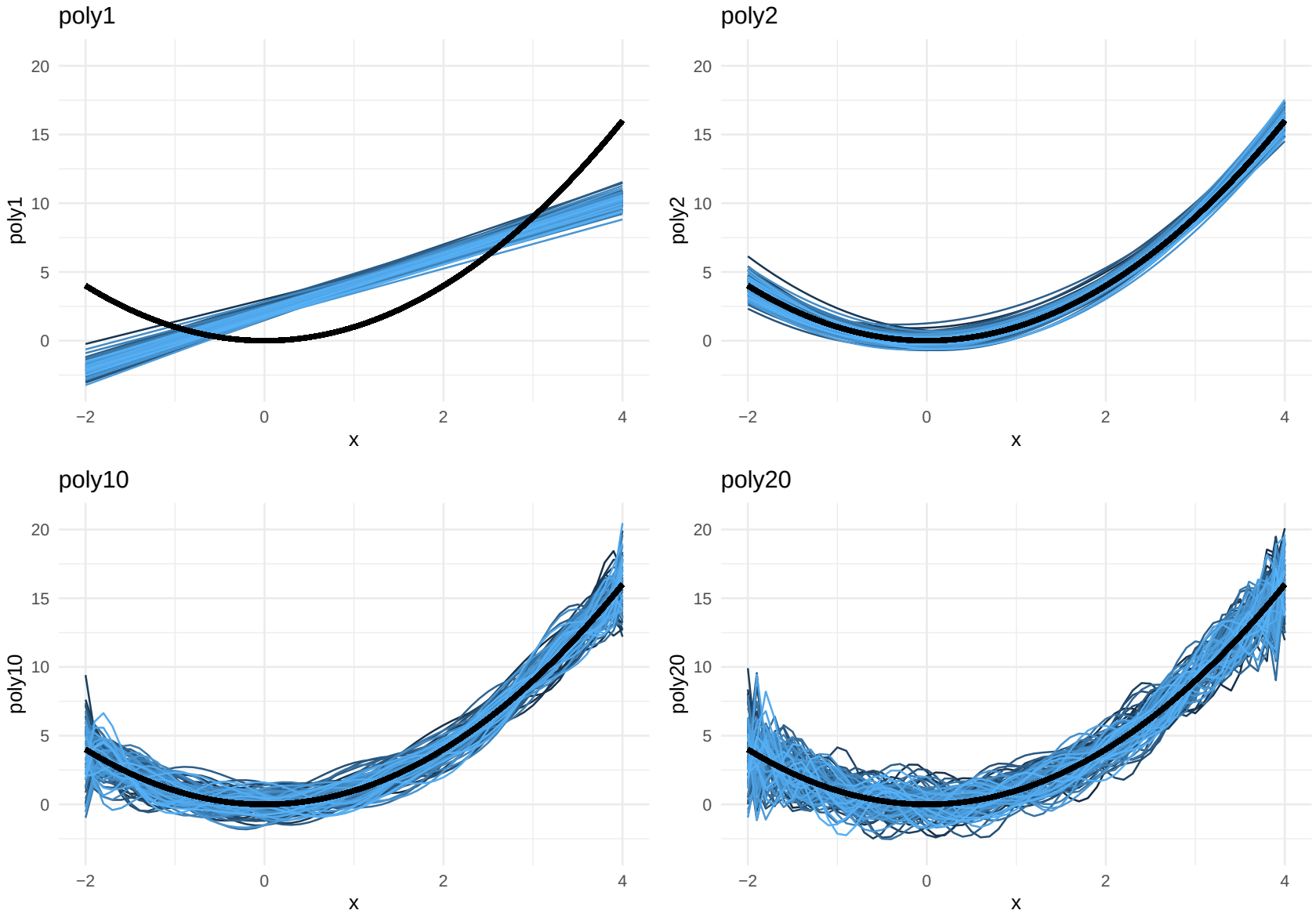
¿Por qué comparar regresiones con diferentes grados de polinomios? Porque en aprendizaje supervisado varios métodos incluyen un parámetro que controla la flexibilidad del ajuste del modelo, por lo que podemos hacer algunas generalizaciones con nuestro ejemplo con grados para los polinomios. El K en el vecino K más cercano es un parámetro de este tipo. Necesitamos saber cómo elegir este parámetro de flexibilidad.

P: Las curvas de colores son nuestras estimaciones para la relación funcional entre x y Y. A continuación trabajaremos con las siguientes preguntas.

- ¿Cuál de las curvas de colores hace el mejor trabajo? Clasifica las curvas.

Ahora, ignore la curva naranja poly2, y solo considere las curvas roja, rosa y morada.

- Supongamos que recopilamos datos nuevos de Y (pero con nuevos errores agregados distribuidos normalmente) y estimamos nuevas curvas. ¿Cuál de las curvas de color tendría *en promedio* el mejor rendimiento?
- ¿Qué elegiste aquí para definir como "mejor rendimiento"?



Se fijo x y generaron nuevos errores 100 veces. Se muestran las 100 curvas ajustadas. La línea negra es la verdadera curva $y = x^2$

Función pérdida

Ahora nos centramos en la predicción.

P: ¿Cómo podemos medir la *pérdida* (error) entre una respuesta pronosticada $\hat{y}_i$ y la respuesta observada y_i?

R: Posibles funciones de pérdida:

- Pérdida absoluta (norma L1): $\mid y_i - \hat{y}_i \mid$
- Pérdida cuadrática (norma L2): $(y_i - \hat{y}_i)^2$
- Pérdida 0/1 (y categórico): Pérdida=0 si $\hat{y}_i = y_i$ y 1 caso contrario

Problemas: robustez, estabilidad, aspecto matemático.

Usaremos la pérdida cuadrática en regresión.

Evaluar la precisión del modelo y la calidad de ajuste

P: Para la regresión (y clasificación) en general: ¿habrá *un método* que domine a todos los demás?

R: Ningún método domina a todos los demás sobre todos los conjuntos de datos posibles.

- Es por eso que necesitamos aprender sobre muchos métodos diferentes.
- Para un conjunto de datos dado, necesitamos saber cómo decidir qué método produce los mejores resultados.
- ¿Qué tan cerca está la respuesta pronosticada al verdadero valor de respuesta?

MSE de entrenamiento

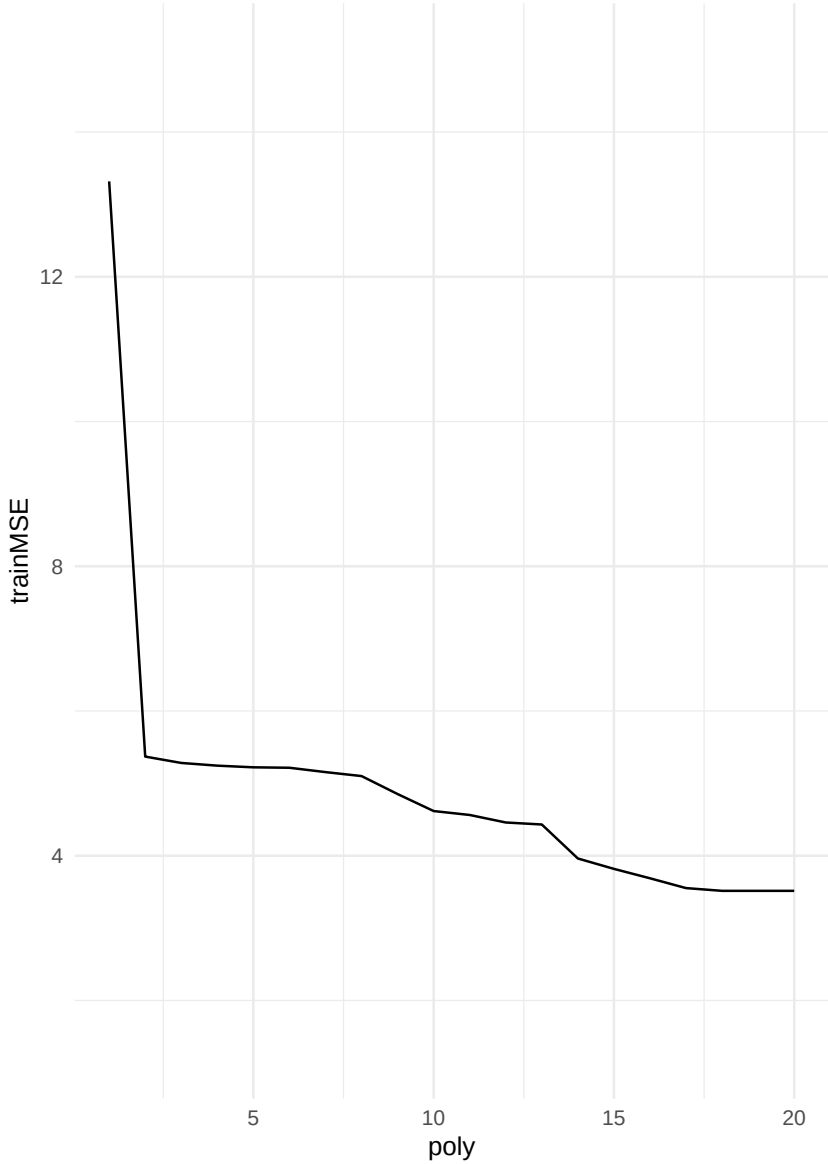
En regresión, se asume que $Y = f(x) + \varepsilon$, y $\hat{f}(x_i)$ da la respuesta predicha para x_i, una medida popular es el *MSE de entrenamiento* (error cuadrático medio): media de las diferencias al cuadrado entre predicción y el valor verdadero para los datos de entrenamiento (los mismos valores que se usaron para estimar f):

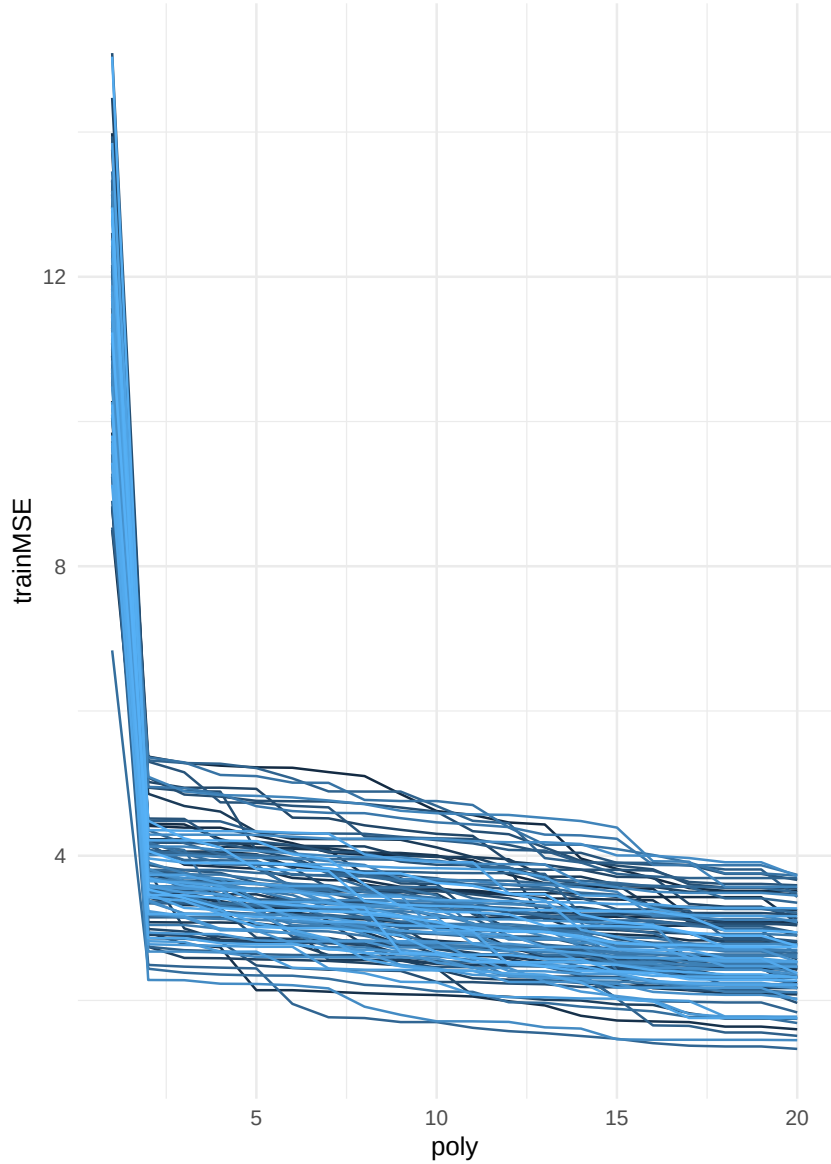
$$\text{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Pero, realmente - *no* estamos interesados en cómo funciona el método en los datos de entrenamiento (y, a menudo, ya hemos diseñado el método para que funcione bien en los datos de entrenamiento), queremos saber qué tan bueno es el método cuando lo usamos en *datos de prueba no vistos*, es decir, datos que podamos observar en el futuro.

Ejemplo:

- no queremos predecir el precio de las acciones de la semana pasada, queremos predecir el precio de las acciones la próxima semana.
- no queremos predecir si un paciente en los datos de entrenamiento tiene diabetes, queremos predecir si un nuevo paciente tiene diabetes.





** P **: Según el MSE de entrenamiento, ¿qué modelo se ajusta mejor a los datos?

Ejemplo polinomio:polinomios ajustados de orden 1-20 cuando la

MSE de prueba

Solución simple: Ajustamos (estimamos $\widehat{(f)}$) de diferentes modelos usando los datos de entrenamiento (tal vez minimizando el MSE), pero elegimos el *mejor* modelo usando un *conjunto de prueba* separado -calculando el *MSE prueba* para un conjunto de n_0 observaciones de prueba (x_{0j}, y_{0j}):

$$\text{MSE}_{\text{test}} = \frac{1}{n_0} \sum_{j=1}^{n_0} (y_{0j} - \hat{f}(x_{0j}))^2$$

Notación alternativa:

$$\text{Ave}(y_0 - \hat{f}(x_0))^2$$

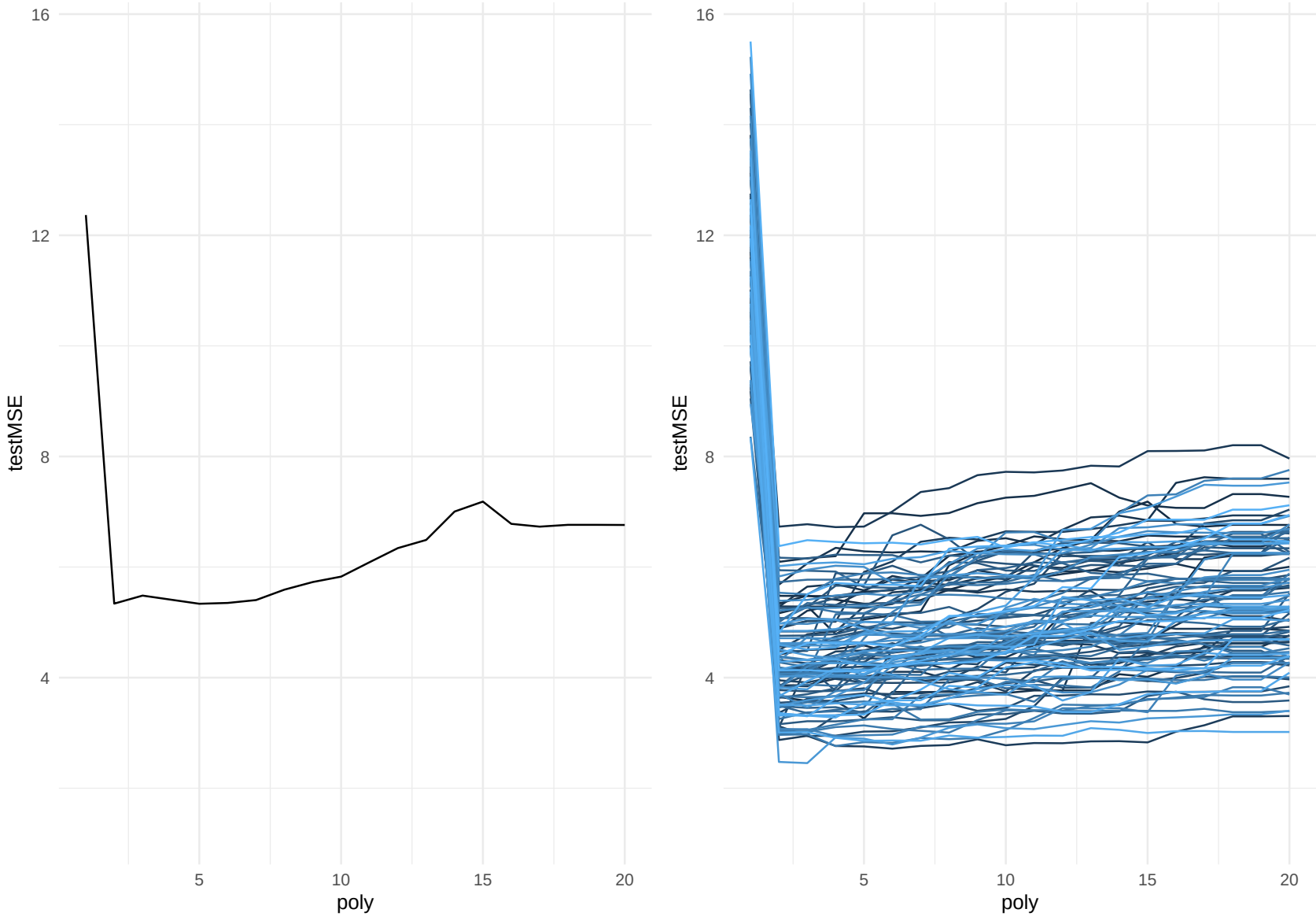
(tomando el promedio sobre todas las observaciones de prueba disponibles).

P: ¿Qué pasa si no tenemos acceso a los datos de prueba?

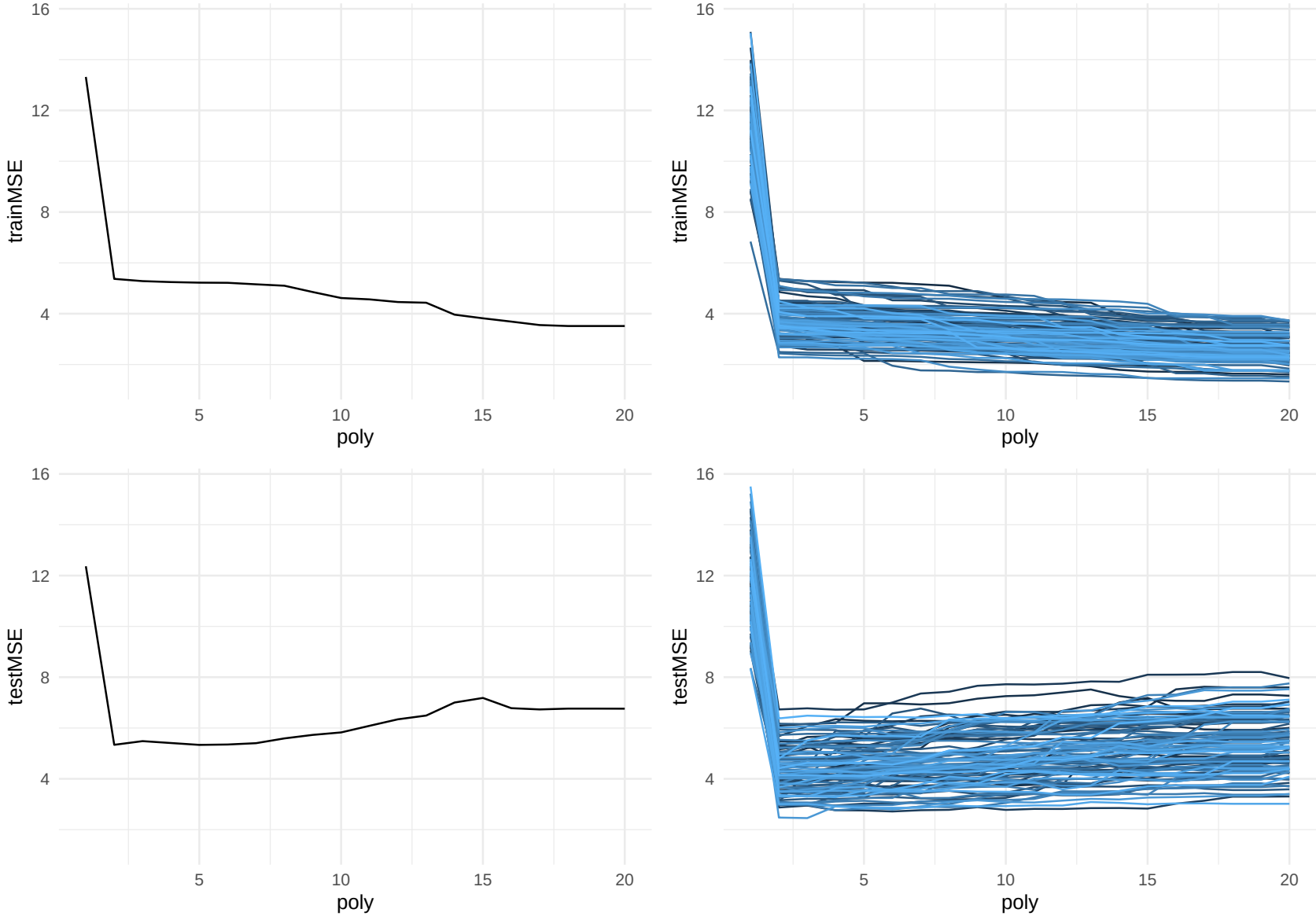
R: En una siguiente clase analizaremos el uso de *validación cruzada* para imitar el uso de un conjunto de prueba.

P: Pero, ¿podemos usar los datos de capacitación MSE para elegir un modelo? ¿Un error de entrenamiento bajo también debería dar un error de prueba bajo?

R: Lamentablemente, no, si utilizamos un modelo flexible, veremos varios casos en los que un error de entrenamiento bajo es un signo de sobreajuste, y daremos un error de prueba alto. Por lo tanto, el error de entrenamiento no es un buen estimador para el error de prueba porque no explica adecuadamente la complejidad del modelo.

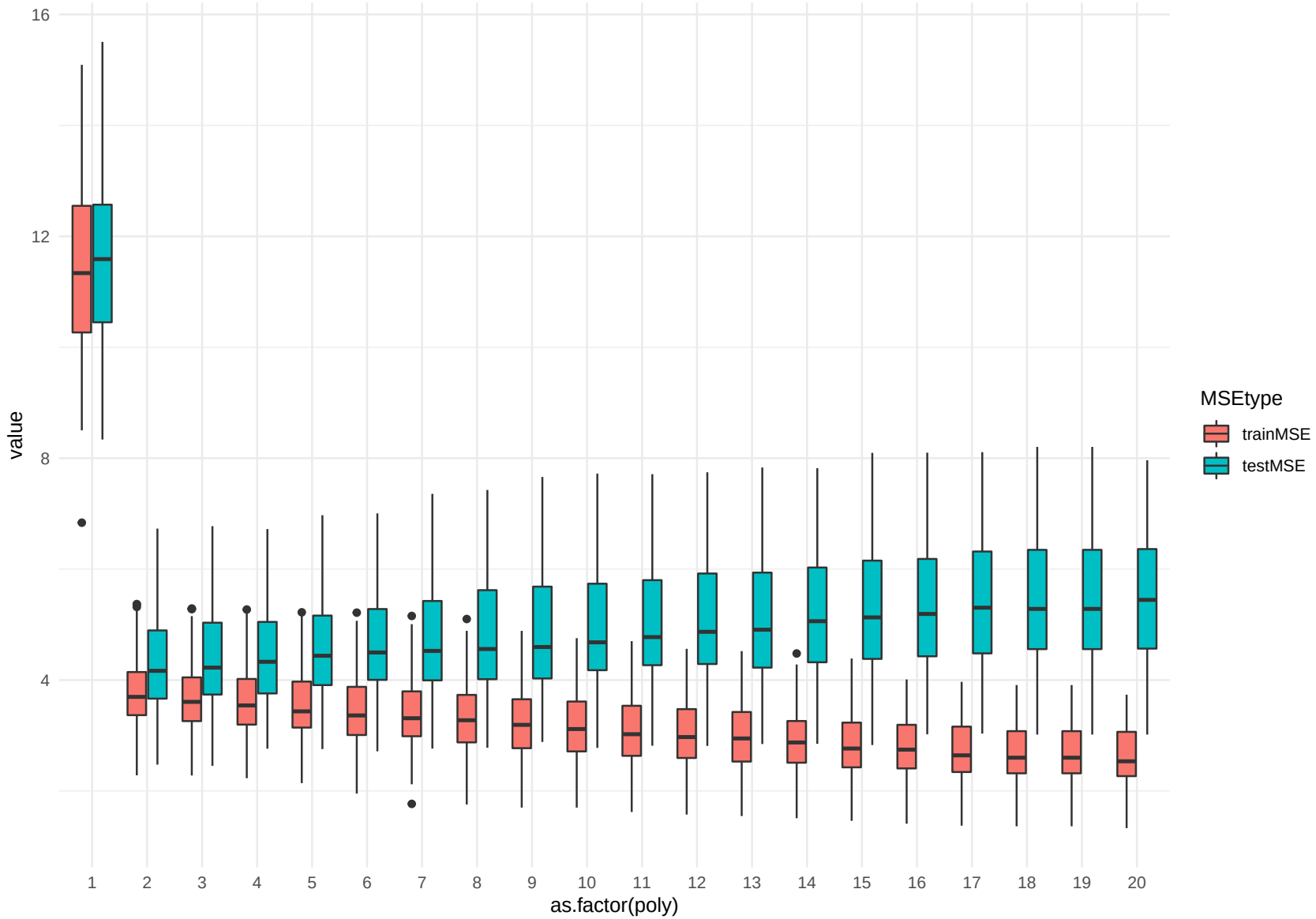


Ejemplo polinomio: Ajuste orden 1-20 cuando el verdadero es 2. Izquierda: una repetición, derecha: 100 repeticiones para el MSE de prueba.



P: Basado en el MSE prueba - ¿que modelo se ajusta mejor?

R: Si se elije flexibilidad basado en trMSE=poly20 gana, si se elije testMSE=poly 2 gana.



Boxplot de las 100 repeticiones (experimento polinomial). Observar la forma de U del error de test.

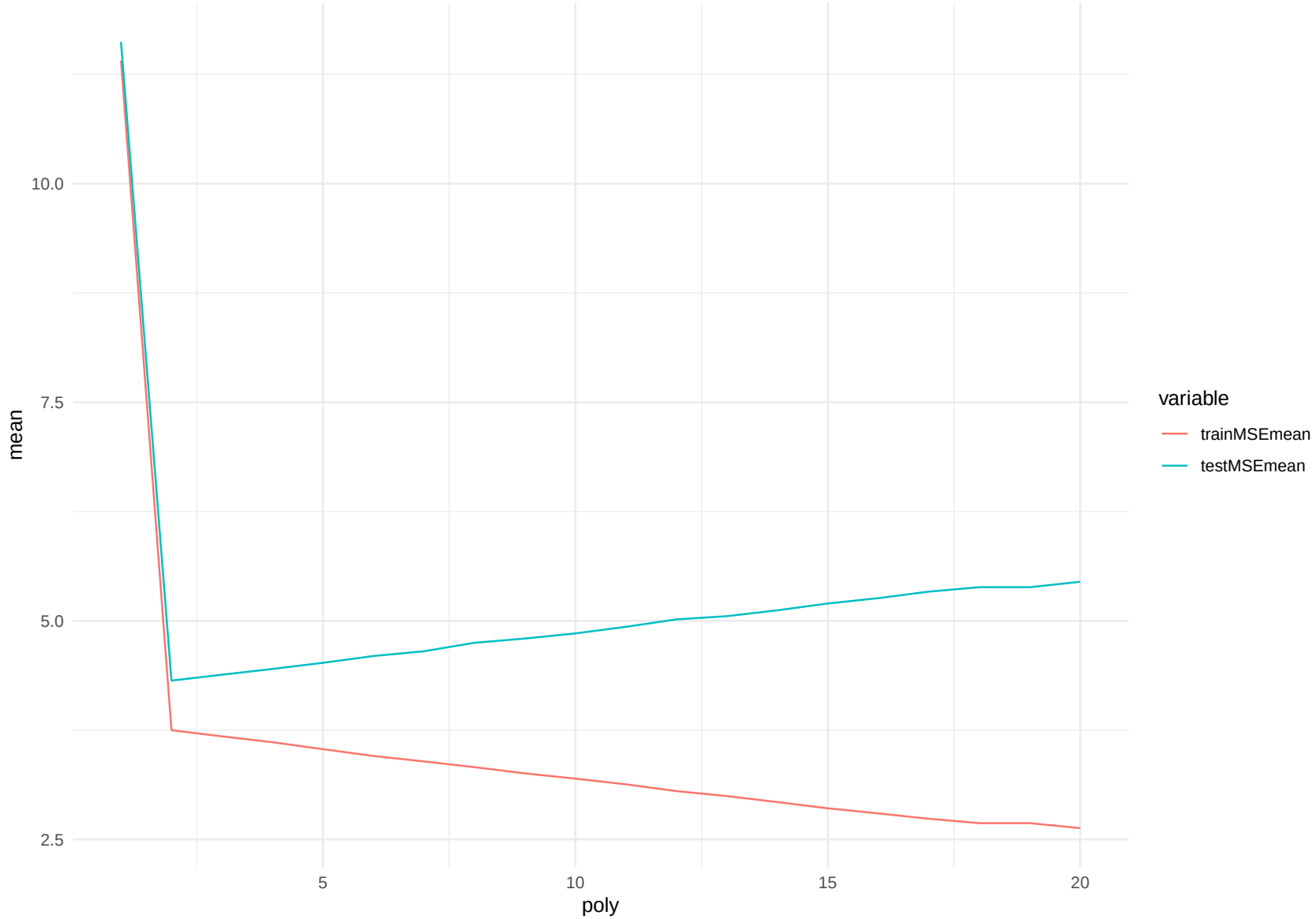
P: ¿Qué puedes leer sobre el diagrama de cajas? ¿Algo nuevo en

R:

Los mismos datos que antes, pero ahora se presentan conjuntamente para el MSE de entrenamiento y prueba, para centrarse en la ubicación y la variabilidad:

Boxplot:

- linea negra=mediana,
- caja del 1r al 3r quartil,
- IQR= Rango Intercuartílico= ancho de la caja
- Bigote de min al max, excepto cuando un dato sea 1.5 veces IQR de la caja, será marcado como outlier.



Media de 100 repeticiones (ejemplo). Observe la forma de U.

Siguiente: Dejamos el trainMSE e intentamos comprender la curva testMSE: ¡dos propiedades en competencia!

El balance entre el sesgo y la varianza

Supongamos que hemos estimado una curva de *regresión* $Y = f(x) + \varepsilon$ en nuestros datos de entrenamiento, que consisten en pares de observaciones independientes $\{x_i, y_i\}$ para $i = 1, .., n$. (Si, solo una covariable x.)

Asumimos que ε ies una variable aleatoria no observada que agrega ruido a la relación entre la variable de respuesta y las covariables y se llama error aleatorio, y que los errores aleatorios tienen media cero y varianza constante σ^2 para todos los valores de x.

Este ruido se usa como un sustituto de todas las variables no observadas que no están en nuestra ecuación, pero que influyen Y.

Supongamos que hemos usado nuestros datos de entrenamiento para producir una curva ajustada, denotada por $\widehat{f}$.

Queremos usar $\hat{f}$ para realizar predicciones con una nueva observación x_0, y estamos interesados en el error asociado con esa predicción. El valor predicho de la variable respuesta será $\hat{f}(x_0)$.

El *error cuadrático medio esperado en la prueba* (MSE) para x_0 está definido como:

$$\mathrm{E}[Y - \hat{f}(x_0)]^2$$

Observación 1: sí, podríamos haber llamado a la nueva respuesta Y_0 en lugar de Y.

Observación 2: compare esto con el MSE de prueba para el ejemplo polinómico; observe que aquí tenemos el *la versión teórica* donde hemos reemplazado el promedio con la media matemática (valor esperado)

Este MSE de prueba esperado puede ser descompuesto en tres términos:

$$\mathsf{E}[Y - \hat{f}(x_0)]^2 = \mathsf{E}[Y^2 + \hat{f}(x_0)^2 - 2Y\hat{f}(x_0)]$$

$$= \mathsf{E}[Y^2] + \mathsf{E}[\hat{f}(x_0)^2] - \mathsf{E}[2Y\hat{f}(x_0)]$$

$$= \mathsf{Var}[Y] + \mathsf{E}[Y]^2 + \mathsf{Var}[\hat{f}(x_0)] + \mathsf{E}[\hat{f}(x_0)]^2 - 2\mathsf{E}[Y]\mathsf{E}[\hat{f}(x_0)]$$

$$= \mathsf{Var}[Y] + f(x_0)^2 + \mathsf{Var}[\hat{f}(x_0)] + \mathsf{E}[\hat{f}(x_0)]^2 - 2f(x_0)\mathsf{E}[\hat{f}(x_0)]$$

$$= \mathsf{Var}[Y] + \mathsf{Var}[\hat{f}(x_0)] + (f(x_0) - \mathsf{E}[\hat{f}(x_0)])^2$$

$$= \mathsf{Var}(\varepsilon) + \mathsf{Var}[\hat{f}(x_0)] + [\mathsf{Bias}(\hat{f}(x_0))]^2.$$

$$\mathsf{E}[(Y - \hat{f}(x_0))^2] = \cdots = \mathsf{Var}(\varepsilon) + \mathsf{Var}[\hat{f}(x_0)] + [\mathsf{Bias}(\hat{f}(x_0))]^2$$

- Primer término: error irreducible, $\mathsf{Var}(\varepsilon) = \sigma^2$, siempre está presente a menos que tengamos mediciones sin error. Este término no se puede reducir independientemente de qué tan bien nuestro modelo estadístico se ajuste a los datos.
- Segundo término: Varianza de la predicción en x_0 o la desviación esperada alrededor de la media en x_0. Si la varianza es alta, existe una gran incertidumbre asociada con la predicción.
- Tercer término: sesgo al cuadrado. El sesgo da una estimación de cuánto difiere la predicción de la media real. Si el sesgo es bajo, el modelo proporciona una predicción cercana al valor verdadero.

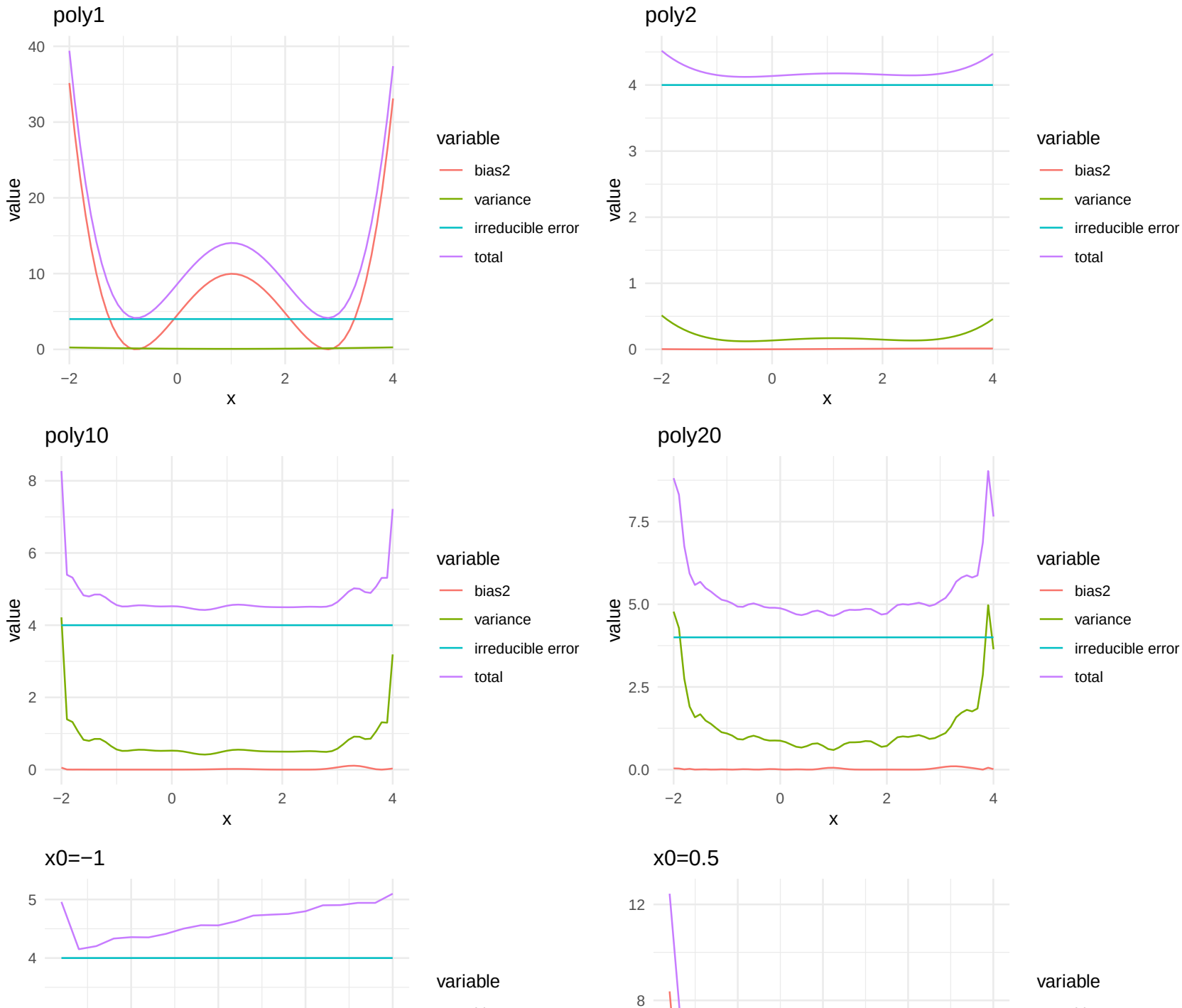
$$\mathrm{E}[(Y - \hat{f}(x_0))^2] = \cdots = \mathrm{Var}(\varepsilon) + \mathrm{Var}[\hat{f}(x_0)] + [\mathrm{Bias}(\hat{f}(x_0))]^2$$

Este es el **MSE de prueba esperado**. odemos considerar esto como el MSE promedio de prueba que obtendríamos si estimáramos repetidamente f usando muchos conjuntos de entrenamiento (como lo hicimos en nuestro ejemplo), y luego probamos esta estimación en x_0.

Entonces, esto es realmente $\mathrm{E}[(Y - \hat{f}(x_0))^2 \mid X = x_0]$ si asumimos que X es una variable aleatoria.

El **MSE de prueba esperado total** * puede ser calculado promediando el MSE de prueba esperado en x_0 sobre todos los valores posibles de x_0 (promedio con respecto a la frecuencia en el conjunto de prueba), o matemáticamente por la ley de esperanza total $\mathrm{E}\{\mathrm{E}[(Y - \hat{f}(X))^2 \mid X]\}$ (a veces también se conoce como la ley de esperanza doble).

Polynomial example (cont.)



Elegir el mejor modelo: observaciones

Cuando se ajusta un modelo estadístico, el objetivo suele ser obtener el modelo más predictivo. A menudo hay muchos modelos candidatos, y la tarea es decidir qué modelo elegir.

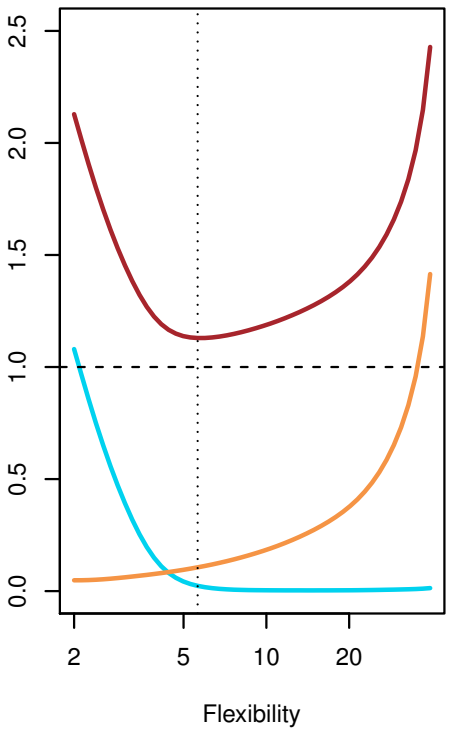
- Las observaciones utilizadas para ajustar el modelo estadístico conforman el conjunto de entrenamiento. El error de entrenamiento es la pérdida promedio sobre la muestra de entrenamiento.
- A medida que aumenta la complejidad (y, por lo tanto, la flexibilidad) de un modelo, el modelo se vuelve más adaptable a las estructuras subyacentes y cae el error de entrenamiento.
- El error de prueba es el error de predicción sobre una muestra de prueba.
- La muestra de prueba tendrá nuevas observaciones que no se utilizaron al ajustar el modelo. Uno quiere que el modelo capture las relaciones importantes entre la variable de respuesta y las covariables, de lo contrario, lo adaptaremos. Recuerde la línea roja en la figura correspondiente al ejemplo de simulación anterior.

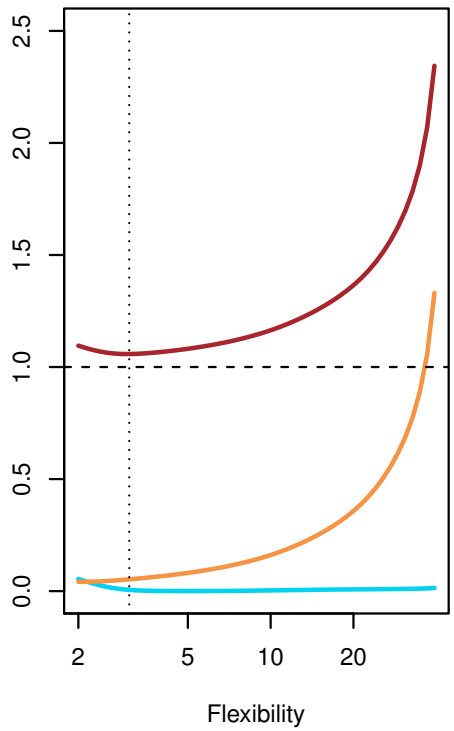
Esta compensación en la selección de un modelo con la cantidad correcta de complejidad/flexibilidad es el **balance entre el sesgo-varianza**.

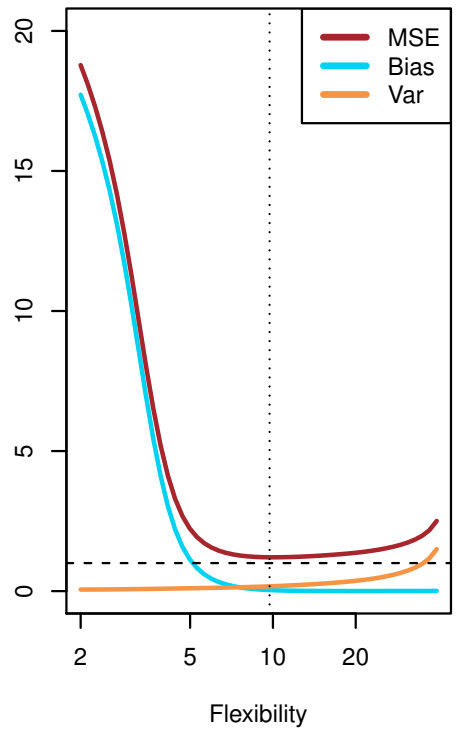
Para resumir:

- los modelos inflexibles (con pocos parámetros para ajustarse) son fáciles de calcular pero pueden dar lugar a un ajuste deficiente (alto sesgo)
- los modelos flexibles (complejos) pueden proporcionar ajustes más imparciales pero pueden sobreajustar los datos (alta varianza)
- habrá errores irreducibles presentes

Elegir un estimador sesgado puede ser mejor que uno insesgado debido a las diferencias en las variaciones, y los métodos como bagging y boosting pueden reducir la variación mientras prevalece un sesgo bajo.


0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility


0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility


MSE
Bias
Var
0
5
10
15
20
2
5
10
20
Flexibility