Más Allá de la Linealidad

Expansiones de Base y Splines
2025

Hoja de Ruta

Donde estamos en el curso:

- Clases 1-3: Bayesiano, decision theory, shrinkage, regularizacion
- Clase 4: Regresion penalizada (Ridge, Lasso)
- **Clase 5: Mas alla de la linealidad**
- Proximas: Arboles, seleccion de modelos

Hoy cubriremos:

1. El trade-off sesgo-varianza formalmente
2. Regresion polinomial y splines
3. Regularizacion en espacio de funciones
4. Aplicacion: Perfil de ingresos por edad

Motivacion: Perfiles de Ingresos por Edad

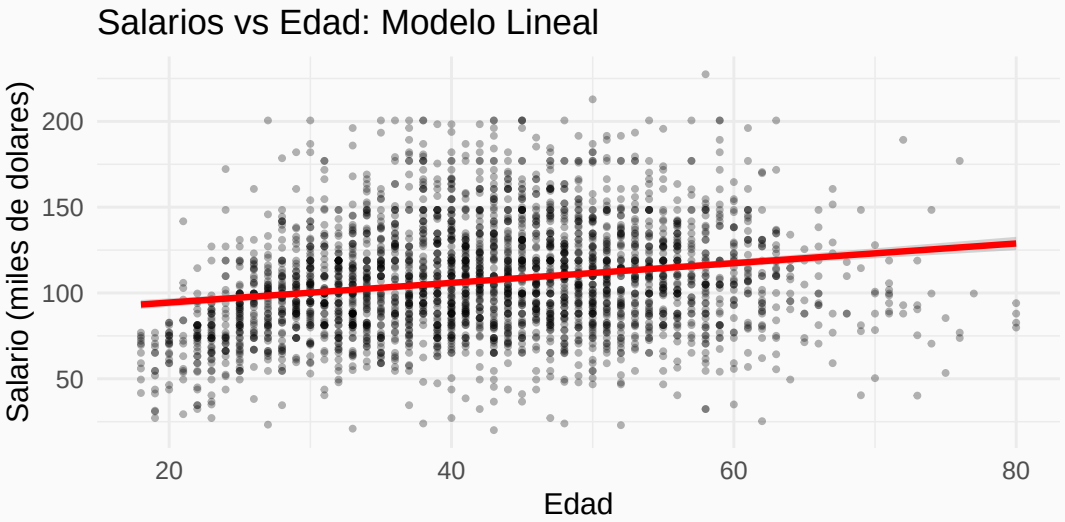
Pregunta economica: Como evolucionan los salarios durante la vida laboral?

Teoria de capital humano (Mincer, Becker):

- Los trabajadores acumulan experiencia
- Los salarios aumentan con la edad/experiencia
- Eventualmente alcanzan un pico (mediados de los 40s)
- Luego se estabilizan o declinan

Pregunta estadistica: Como modelamos esta relacion no lineal?

El Problema con Modelos Lineales



Problema obvio: La relacion no es lineal. El modelo lineal predice mal en los extremos.

Parte 1: El Trade-off Sesgo-Varianza Formalmente

Descomposicion del Error de Prediccion

Supongamos el modelo verdadero es $y = f(x) + \varepsilon$ donde $E[\varepsilon] = 0$ y $\text{Var}(\varepsilon) = \sigma^2$.

Para un estimador $\hat{f}(x)$, el **error cuadratico medio esperado** en un punto x es:

$$E[(y - \hat{f}(x))^2] = \text{Bias}^2[\hat{f}(x)] + \text{Var}[\hat{f}(x)] + \sigma^2$$

Componentes:

- $\text{Bias}^2[\hat{f}(x)] = (E[\hat{f}(x)] - f(x))^2$: Sesgo al cuadrado
- $\text{Var}[\hat{f}(x)]$: Varianza del estimador
- σ^2: Error irreducible

Demostracion de la Descomposicion

$$
\begin{aligned}
E[(y - \hat{f}(x))^2] &= E[(f(x) + \varepsilon - \hat{f}(x))^2] \\
&= E[(f(x) - \hat{f}(x))^2] + E[\varepsilon^2] + 2E[(f(x) - \hat{f}(x))\varepsilon]
\end{aligned}
$$

Dado que $E[\varepsilon] = 0$ y ε es independiente de $\hat{f}$:

$$
\begin{aligned}
&= E[(f(x) - \hat{f}(x))^2] + \sigma^2 \\
&= E[f(x)^2 - 2f(x)\hat{f}(x) + \hat{f}(x)^2] + \sigma^2 \\
&= f(x)^2 - 2f(x)E[\hat{f}(x)] + E[\hat{f}(x)^2] + \sigma^2
\end{aligned}
$$

Demostracion (continuacion)

Sumando y restando $E[\hat{f}(x)]^2$:

$$
\begin{aligned}
&= f(x)^2 - 2f(x)E[\hat{f}(x)] + E[\hat{f}(x)]^2 + E[\hat{f}(x)^2] - E[\hat{f}(x)]^2 + \sigma^2 \\
&= \underbrace{(f(x) - E[\hat{f}(x)])^2}_{\text{Bias}^2} + \underbrace{(E[\hat{f}(x)^2] - E[\hat{f}(x)]^2)}_{\text{Var}} + \sigma^2
\end{aligned}
$$

Por lo tanto:

$$
\boxed{E[(y - \hat{f}(x))^2] = \text{Bias}^2[\hat{f}(x)] + \text{Var}[\hat{f}(x)] + \sigma^2}
$$

Interpretacion de los Componentes

Sesgo $(E[\hat{f}(x)] - f(x))$:

- Error promedio del modelo
- Alta en modelos muy simples (e.g., lineal cuando la verdad es no lineal)
- Modelos simples no pueden capturar la complejidad de la relacion

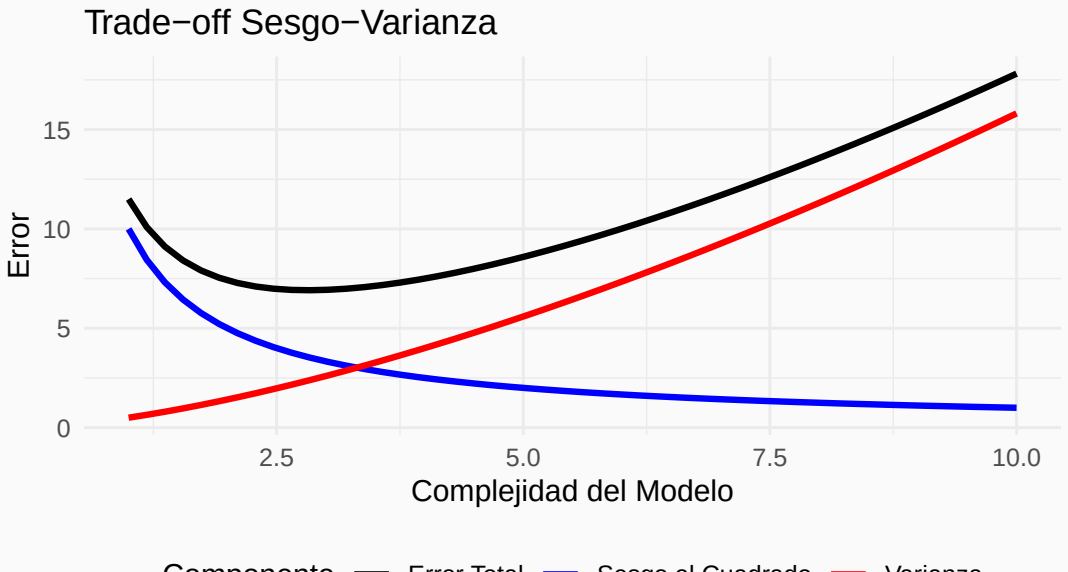
Varianza $\text{Var}[\hat{f}(x)]$:

- Cuanto cambia $\hat{f}(x)$ con diferentes muestras
- Alta en modelos muy flexibles (e.g., polinomios de alto grado)
- Modelos complejos "memorizan" el ruido en los datos

Error irreducible σ^2:

- Varianza intrinseca en y
- No podemos reducirlo con mejor modelado

El Trade-off


Trade−off Sesgo−Varianza
Error
15
10
5
0
2.5
5.0
7.5
10.0
Complejidad del Modelo
Componente
Error Total
Sesgo al Cuadrado
Varianza

Modelos Lineales vs Flexibles

Modelos Lineales ($y = \beta_0 + \beta_1 x$):

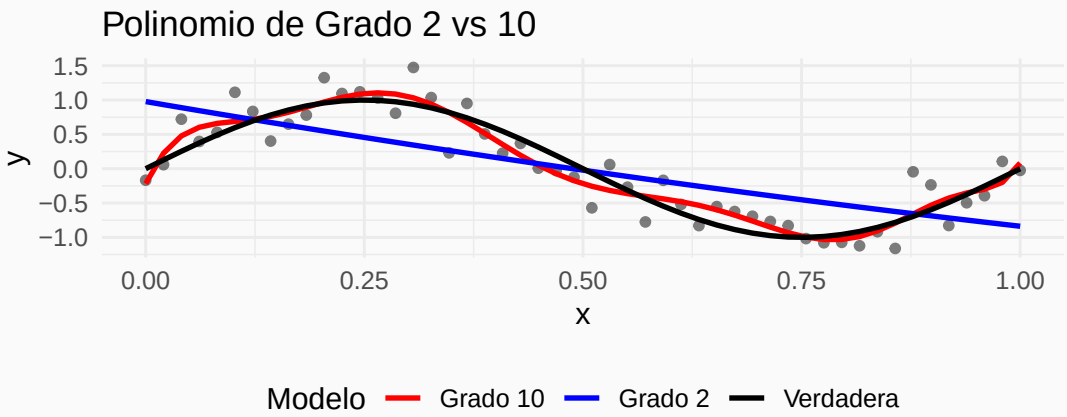
- **Sesgo**: Alto si $f(x)$ no es lineal
- **Varianza**: Baja (pocos parametros, estimacion estable)
- **Cuando usar**: Relacion aproximadamente lineal, pocos datos

Modelos Flexibles (polinomios de alto grado, splines complejos):

- **Sesgo**: Bajo (pueden aproximar cualquier funcion)
- **Varianza**: Alta (muchos parametros, sensibles al ruido)
- **Cuando usar**: Relacion claramente no lineal, muchos datos

Objetivo: Encontrar el punto optimo de complejidad.

Ejemplo: Modelo Cuadratico vs Decimo Grado



Grado 2: Mayor sesgo, menor varianza. Grado 10: Menor sesgo, mayor varianza (overfitting).

Parte 2: Regresion Polinomial y Splines

Regresion Polinomial

Idea: Ajustar un polinomio de grado d:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d + \varepsilon$$

Equivalente a regresion lineal con variables transformadas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \varepsilon$$

donde $x_1 = x$, $x_2 = x^2$, $\ldots$, $x_d = x^d$

Estimacion: MCO estandar

Problemas con Polinomios

1. **Comportamiento global**: Un cambio en una region afecta toda la funcion
2. **Oscilaciones en los bordes** (fenomeno de Runge):
   - Polinomios de alto grado oscilan salvajemente cerca de los extremos
   - Predicciones absurdas fuera del rango de los datos
3. **Mala extrapolacion**:
   - Polinomios divergen rapidamente fuera del rango observado
   - Suele no tener sentido economico (salarios negativos para jovenes/viejos)

La Solucion: Regresion por Tramos

Idea: Dividir el rango de x en regiones, ajustar funciones separadas en cada region.

Regresion polinomial por tramos:

1. Elegir puntos de corte (knots) $\xi_1, \xi_2, ..., \xi_K$
2. Ajustar polinomios separados en cada intervalo:
   - $[\min, \xi_1]$: $f_1(x) = \beta_{10} + \beta_{11}x + \cdots$
   - $[\xi_1, \xi_2]$: $f_2(x) = \beta_{20} + \beta_{21}x + \cdots$
   - $\ldots$
   - $[\xi_K, \max]$: $f_{K+1}(x) = \beta_{K+1,0} + \beta_{K+1,1}x + \cdots$

Ventaja: Flexibilidad local sin afectar regiones distantes.

Restricciones de Continuidad: Splines

Problema: Funciones por tramos pueden ser discontinuas en los knots.

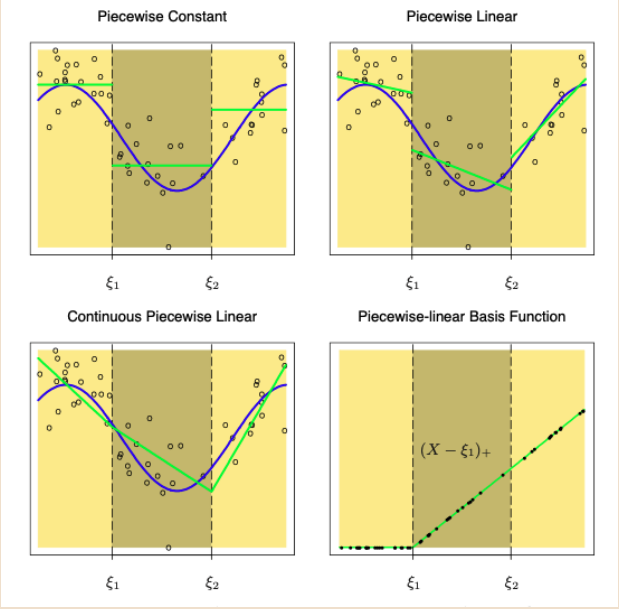
Solucion: Imponer **restricciones de suavidad**.

Definicion: Spline

Un spline de grado d con knots en $\xi_1, ..., \xi_K$ es un polinomio por tramos de grado d que es continuo y tiene derivadas continuas hasta orden $d - 1$ en cada knot.

Splines cubicos (grado 3):

- Continuos en los knots
- Primera derivada continua (sin esquinas)
- Segunda derivada continua (sin cambios abruptos en curvatura)


Piecewise Constant
Piecewise Linear
Continuous Piecewise Linear
Piecewise-linear Basis Function
ξ_1
ξ_2
$(X - \xi_1)_+$

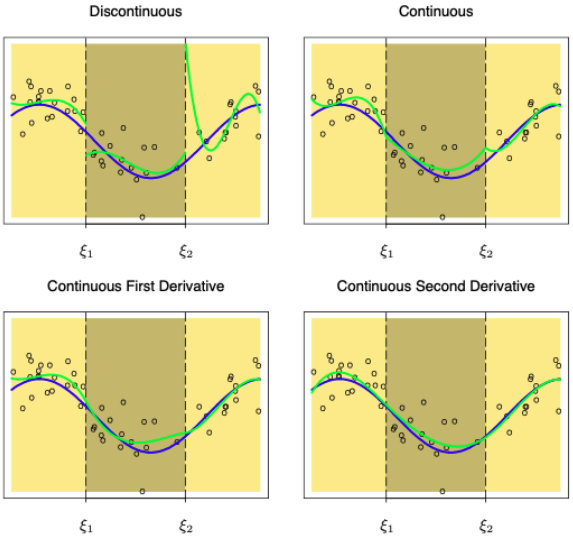


FIGURE 5.2. *A series of piecewise-cubic polynomials, with increasing orders of continuity.*

Representacion con Funciones Base

Un spline cubico con K knots se puede escribir como:

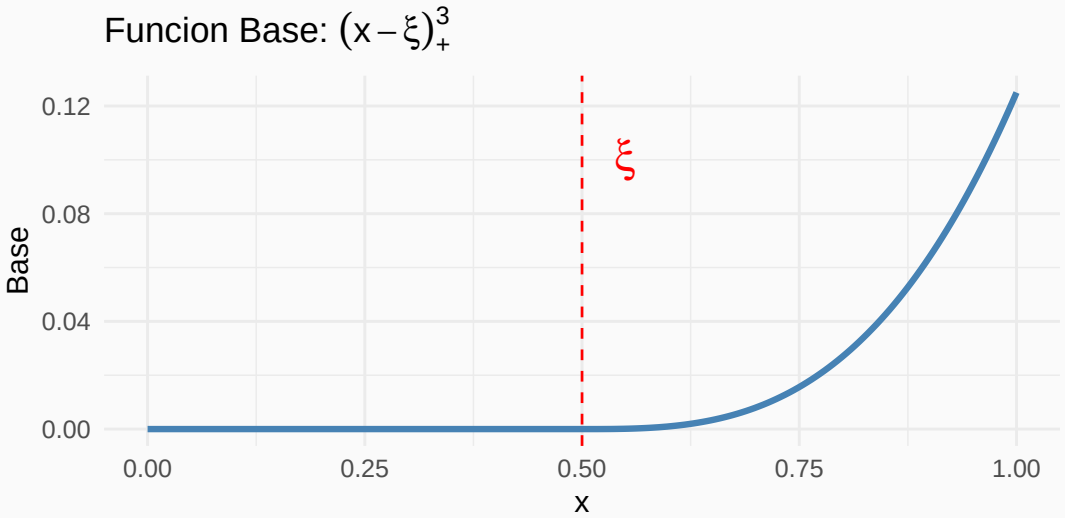
$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3$$

donde $(x - \xi_k)_+ = \max(0, x - \xi_k)$ es la **funcion de truncamiento**.

Parametros: $\beta_0, \beta_1, \beta_2, \beta_3, \theta_1, ..., \theta_K$ (total: $4 + K$)

Estimacion: Regression lineal estandar con $K + 4$ variables.

Visualizando la Base de Truncamiento



La funcion es cero antes del knot, cubica despues.

B-splines: Una Base Mas Eficiente

Problema con truncated power basis: Numericamente inestable para muchos knots.

Solucion: **B-splines** (basis splines)

- Funciones base con soporte local (no cero solo cerca de algunos knots)
- Numericamente estables
- Mismo espacio de funciones, mejor computacionalmente

En R: Usar `splines::bs()` en vez de construir la base manualmente.

```r
library(splines)
fit <- lm(wage ~ bs(age, knots = c(25, 40, 60)), data = Wage)
```

Eligiendo el Numero de Knots

Pregunta clave: Cuantos knots usar? Donde colocarlos?

Opciones comunes:

1. **Knots equiespaciados**: Dividir el rango en intervalos iguales
2. **Cuantiles**: Colocar knots en cuantiles de x (e.g., cuartiles)
   - Mas knots donde hay mas datos
3. **Cross-validation**: Probar diferentes numeros de knots, elegir el que minimiza error de prediccion

Trade-off: Mas knots = mas flexibilidad = menor sesgo pero mayor varianza.

Splines Penalizados (P-splines)

Problema: Incluso con splines, necesitamos elegir numero de knots.

Solucion: Usar muchos knots, pero **penalizar rugosidad**.

$$\min_{\beta} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$$

Componentes:

- Primer termino: Ajuste a los datos (RSS)
- Segundo termino: Penalizacion por rugosidad (integral de segunda derivada al cuadrado)
- $\lambda \geq 0$: Controla el trade-off

Conexion con Ridge: Penalizacion en espacio de funciones en vez de parametros.

Interpretacion del Parametro de Suavizado λ

$\lambda = 0$:

- Sin penalizacion
- Interpola todos los puntos (overfitting)
- Alta varianza, bajo sesgo

$\lambda \to \infty$:

- Penalizacion maxima
- Fuerza $f''(x) = 0$ (linea recta)
- Bajo varianza, alto sesgo

λ **intermedio**: Balance optimo entre sesgo y varianza.

Parte 3: Regularizacion en Espacio de Funciones

Problema general: Entre todas las funciones, encontrar f que minimice:

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$$

Teorema (Wahba, 1990)

La solucion es un **natural cubic spline** con knots en cada x_i observado.

Sorprendente: El problema infinito-dimensional se reduce a ajustar un spline con parametros finitos.

En practica: Usamos basis de B-splines y penalizacion matricial.

Forma Matricial de P-splines

$$f(x) = \sum_{j=1}^{K+d} \beta_j B_j(x)$$

El problema se convierte en:

$$\min_{\beta} ||y - B\beta||^2 + \lambda \beta' D \beta$$

- B: Matriz de bases evaluadas en los datos
- D: Matriz de penalizacion (aproxima $\int [f''(x)]^2 dx$)

Solucion cerrada:

$$\hat{\beta} = (B'B + \lambda D)^{-1} B'y$$

¡Es exactamente como Ridge regression con matriz de penalizacion D!

Conexion con Ridge Regression

Ridge (Clase 4):

$$\min_{\beta} ||y - X\beta||^2 + \lambda||\beta||^2$$

P-splines:

$$\min_{\beta} ||y - B\beta||^2 + \lambda\beta' D\beta$$

Similitudes:

- Ambos penalizan la norma de β (posiblemente ponderada)
- Ambos tienen solucion cerrada
- Ambos encogen coeficientes hacia cero

Diferencia clave: P-splines penalizan **diferencias entre coeficientes adyacentes** (rugosidad) en vez de magnitud absoluta.

Seleccion de λ por Cross-Validation

Procedimiento:

1. Dividir datos en K folds
2. Para cada valor de λ candidato:
   a. Para cada fold $k = 1, ..., K$:
      - Entrenar en todos menos fold k
      - Predecir en fold k
      - Calcular error: $\text{MSE}_k = \frac{1}{n_k} \sum_{i \in \text{fold } k} (y_i - \hat{f}_\lambda(x_i))^2$
   b. Promediar: $\text{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \text{MSE}_k$
3. Elegir $\lambda^* = \arg\min_\lambda \text{CV}(\lambda)$

Resultado: λ que balancea sesgo y varianza optimamente para prediccion fuera de muestra.

Grados de Libertad Efectivos

Para smoothing splines, podemos definir **grados de libertad efectivos**:

$$\text{df}(\lambda) = \text{tr}(S_\lambda)$$

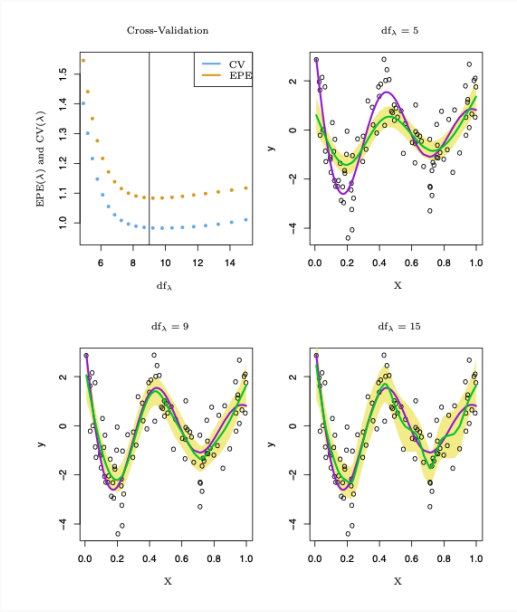
donde S_λ es la matriz smoother: $\hat{y} = S_\lambda y$

Interpretacion:

- df $= 2$: Linea recta
- df $= n$: Interpola todos los puntos
- Valores intermedios: Flexibilidad intermedia

Util: Podemos especificar df deseados en vez de λ directamente.

```
fit <- smooth.spline(age, wage, df = 10)
```


Cross-Validation
CV
EPE
EPE(λ) and CV(λ)
dfλ
dfλ = 5
y
x
dfλ = 9
dfλ = 15

Puntos Clave de Parte 3

1. **Smoothing splines**: Penalizan rugosidad ($\int [f''(x)]^2 dx$)
2. **Conexion con Ridge**: Penalizacion en espacio de parametros extendidos
3. λ **controla trade-off sesgo-varianza**:
   - λ pequeno: Flexible, bajo sesgo, alta varianza
   - λ grande: Rigido, alto sesgo, baja varianza
4. **Cross-validation** selecciona λ optimo
5. **Grados de libertad efectivos**: Medida intuitiva de complejidad

Cuando Usar Cada Metodo

Regresion lineal:

- Relacion aproximadamente lineal
- Interpretabilidad es clave

Polinomios de bajo grado (2-4):

- No linealidad moderada
- Interpretacion simple

Splines con knots fijos:

- Control explicito sobre flexibilidad
- Sabes donde quieres mas/menos flexibilidad

Smoothing splines:

- Maxima flexibilidad

Aplicacion: Perfil de Ingresos por Edad

Contexto economico:

- Ecuacion de Mincer: $\log(\text{wage}) = \beta_0 + \beta_1 \text{edu} + \beta_2 \text{exp} + \beta_3 \text{exp}^2 + \varepsilon$
- Asume forma cuadratica en experiencia
- Pero es esto flexible suficiente?

Hoy exploraremos:

1. Ajustar modelos lineales y polinomiales
2. Mostrar problemas de polinomios de alto grado
3. Ajustar splines con CV
4. Comparar in-sample vs out-of-sample performance

RStudio Demo

Vamos al codigo:

Archivo: `wage_splines_example.R`

Lo que haremos:

1. Cargar datos de ISLR::Wage
2. Ajustar modelos de complejidad creciente
3. Visualizar overfitting
4. Usar CV para seleccionar lambda optimo
5. Comparar errores de prediccion

Resumen de la Clase

Conceptos clave:

1. **Descomposicion sesgo-varianza**: $\text{MSE} = \text{Bias}^2 + \text{Variance} + \sigma^2$
2. **Regresion polinomial**: Simple pero problemas globales y en bordes
3. **Splines**: Flexibilidad local con suavidad
    - Piecewise polynomials
    - Cubic splines