Tarea 1

Mateo Perez

Octubre 2024

Ejercicio 1

Demostrar:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2\right] \leq E\left[\frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{\beta}^T x_i')^2\right]$$

Primer paso:

$$E\left[\frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{\beta}^T x_i')^2\right] = \frac{1}{m}\sum_{i=1}^{m} E\left[(y_i' - \hat{\beta}^T x_i')^2\right]$$

Como:

$$E\left[(y_i' - \hat{\beta}^T x_i')^2\right] = E_t\left[E_v\left[(y_i' - \hat{\beta}^T x_i')^2|t\right]\right]$$

Entonces:

$$E\left[\frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{\beta}^T x_i')^2\right] = \frac{1}{m}\sum_{i=1}^{m} E_t\left[E_v\left[(y_i' - \hat{\beta}^T x_i')^2|t\right]\right]$$

Ahora tomo que:

$$E_v\left[(y_i' - \hat{\beta}^T x_i')^2|t\right] = E_v\left[(y_1' - \hat{\beta}^T x_1')^2|t\right]$$

Por lo tanto:

$$E\left[\frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{\beta}^T x_i')^2\right] = \frac{1}{m}\sum_{i=1}^{m} E_t\left[E_v\left[(y_1' - \hat{\beta}^T x_1')^2|t\right]\right]$$

Entonces:

$$E\left[\frac{1}{m}\sum_{i=1}^{m}(y_i' - \hat{\beta}^T x_i')^2\right] = E_t\left[E_v\left[(y_1' - \hat{\beta}^T x_1')^2|t\right]\right] = E\left[(y_1' - \hat{\beta}^T x_1')^2\right]$$

Segundo paso:

Se consideran las siguientes VA:

$$A = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2$$

$$B = \frac{1}{n}\sum_{i=1}^{n}(y'_i - \tilde{\beta}^T x'_i)^2$$

Dado que y_i y y'_i, y x_i y x'_i son iid, entonces se puede afirmar que A y B tienen la misma distribución y por lo tanto $E[A] = E[B]$.

Tercera parte:

Como:

$$y'_i - \tilde{\beta}^T x'_i = (y'_i - \hat{\beta}^T x'_i) + (\hat{\beta}^T x'_i - \tilde{\beta}^T x'_i)$$

y $E[(a+b)^2] \leq E[a^2] + E[b^2]$, entonces:

$$(y'_i - \tilde{\beta}^T x'_i)^2 \leq (y'_i - \hat{\beta}^T x'_i)^2 + (\hat{\beta}^T x'_i - \tilde{\beta}^T x'_i)^2$$

Y por lo tanto:

$$\frac{1}{n}\sum_{i=1}^{n}(y'_i - \tilde{\beta}^T x'_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y'_i - \hat{\beta}^T x'_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}^T x'_i - \tilde{\beta}^T x'_i)^2$$

Como el último término es cercano a 0, entonces:

$$B \leq \frac{1}{n}\sum_{i=1}^{n}(y'_i - \hat{\beta}^T x'_i)^2$$

Cuarta parte:

Siguiendo el paso anterior:

$$E[B] \leq E\left[\frac{1}{n}\sum_{i=1}^{n}(y'_i - \hat{\beta}^T x'_i)^2\right]$$

Por lo que queda demostrado que:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2\right] \leq E\left[\frac{1}{n}\sum_{i=1}^{n}(y'_i - \hat{\beta}^T x'_i)^2\right]$$

Ejercicio 2

a) El tamaño muestral n es extremadamente grande y el número de pedictores p es pequeño.

Se espera que un método flexible funcione mejor ya que se puede ajustar mejor a la cantidad de datos.

b) El númeto de predictores p es extremadamente grande y el número de observaciones n es pequeño.

Se espera que un método inflexible funcione mejor ya que los flexibles seguramente generen sobreajuste.

c) La relación entre los predictores y la respuesta es marcadamente no lineal.

Se espera que un método flexible funcione mejor ya que este podría ajustar relaciones no lineales sin problemas.

d) La varianza del término de error $sigma^2 = V(\epsilon)$ es extremadamente alta.

Se espera que un método inflexible funcione mejor ya que los otos tenderán a sobreajustar.

Ejercicio 3

Modelos más flexibls tienen las ventajas de que pueden adaptarse mejor al tener mucha cantidad de variables y tienen una capacidad bastante alta para capturar relaciones complejas, por ejemplo no lineales, pero tienen las desventajas de que existe un mayor riesgo de producir sobreajustes y generalmente se precisa de una mayor cantidad de datos para su buen funcionamiento.

Por otro lado los modelos menos flexibles suelen usarse cuando existe una baja cantidad de datos, o cuando el interés principal está en poder interpretar los resultados, su desventaja principal es la baja capacidad para capturar relaciones complejas en los datos, además de que generalmente producen predictores con mayor sesgo a los flexibles.

Ejercicio 4

En clasificación se busca los “k” vecinos más cercanos al dato y se lo clasifica según la mayoría de etiquetas de esos vecinos, es decir si tiene 5 vecinos y 3 tienen una misma etiqueta y 2 otra, ese dato se clsifica como los 3 mencionados.
En regresión lo que sucede es que se toma el promedio de los “k” vecinos más cercanos para hacer una predicción sobre una clasificación.
La principal diferencia entonces es que la clasificación se usa para valores discretos, mientras que la regresión se usa para valores continuos.

Ejercicio 5

A)

Es probable que la SCR de la regresión cúbica sea menor o igual que la SCR de la regresión lineal ya que la regresión cúbica al ser más flexible, tiene la capacidad de ajustarse mejor a las variaciones en los datos de entrenamiento, aunque esas variaciones no representen la verdadera relación lineal.

B)

Se espera que la SCR sea menor para la regresión lineal que para la regresión cúbica ya que la regresión lineal va a captar mejor la verdadera relación entre X e Y, debido a que efectivamente esta relación es lineal.

C)

Se espera que la SCR de entrenamiento para la regresión cúbica sea menor que la SCR de entrenamiento para la regresión lineal, debido a la mayor flexibilidad de la regresión cúbica para ajustarse a un modelo no lineal.

D)

No existe suficiente información para asegurar que la SCR en el conjunto de test sea menor o mayor para la regresión cúbica que para la regresión lineal, debido a que aunque la regresión cúbica sea más flexible para captar esta relación no lineal, exite otro problema que viene por el lado del sobreajuste en el que muy seguramente caiga la regresión cúbica.

Ejercicio 6

```
library(ggplot2)
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.4.1
```

```
## Warning: package 'dials' was built under R version 4.4.1
```

```
## Warning: package 'infer' was built under R version 4.4.1
```

```
## Warning: package 'modeldata' was built under R version 4.4.1
```

```
## Warning: package 'parsnip' was built under R version 4.4.1
```

```
## Warning: package 'recipes' was built under R version 4.4.1
```

```
## Warning: package 'rsample' was built under R version 4.4.1
```

```
## Warning: package 'tune' was built under R version 4.4.1
```

```
## Warning: package 'workflows' was built under R version 4.4.1
```

```
## Warning: package 'workflowsets' was built under R version 4.4.1
```

```
## Warning: package 'yardstick' was built under R version 4.4.1
```

```
library(MASS)
library(dplyr)
library(parsnip)
library(splines)
library(mgcv)
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.4.1
```

```
library(rsample)
library(yardstick)
```

```
datos <- data.frame(
  ID = 1:20,
  Y = c(8, 9, 14, 10, 10, 15, 11, 6, 7, 8, 13, 11, 11, 10, 8, 15, 11, 4, 12, 8),
  X = c(6, 8, 12, 9, 9, 13, 11, 6, 5, 9, 13, 10, 11, 10, 8, 15, 11, 3, 11, 7)
)

pliegues <- list(
  pliegue_1 = c(4, 3, 19, 16),
  pliegue_2 = c(2, 15, 7, 18),
```

```r
  pliegue_3 = c(9, 14, 12, 20),
  pliegue_4 = c(17, 6, 8, 10),
  pliegue_5 = c(1, 5, 13, 11)
)

# Función para calcular el MSE
calcular_mse <- function(train, test) {
  mod <-
    parsnip::linear_reg() %>%
    parsnip::set_engine("lm")

  modelo <-
    mod |>
    parsnip::fit(Y ~ X, data = train)

  mse <- augment(modelo, new_data = test) %>%
    rmse(Y, .pred)^2

}

mse_list <- c()

# Paso a paso del procedimiento,
# en cada paso se deja afuera un pliegue distinto,
# se ajusta el modelos para los otros 4 pliegues
# y se testea con el que se deja afuera, calculando el MSE,
# en total cada pliegue se usa 4 veces para training y 1 vez para test

for (i in 1:length(pliegues)) {
  # Test
  test_ids <- pliegues[[i]]

  # Training y test
  test_set <- datos[datos$ID %in% test_ids, ]
  train_set <- datos[!datos$ID %in% test_ids, ]

  # MSE para el pliegue
  mse <- calcular_mse(train_set, test_set)
  mse_list <- c(mse_list, mse)
}

# Estimación del MSE por validación cruzada usando 5 pliegues (promedio de los 5 MSE's)
mse_cv <- mean(mse_list)
mse_cv
```

Ejercicios ISLR

Capítulo 5 Ejercicio 8

a)

```
set.seed(1)
x <- rnorm(100)
y <- x - 2 * x^2 + rnorm(100)
```

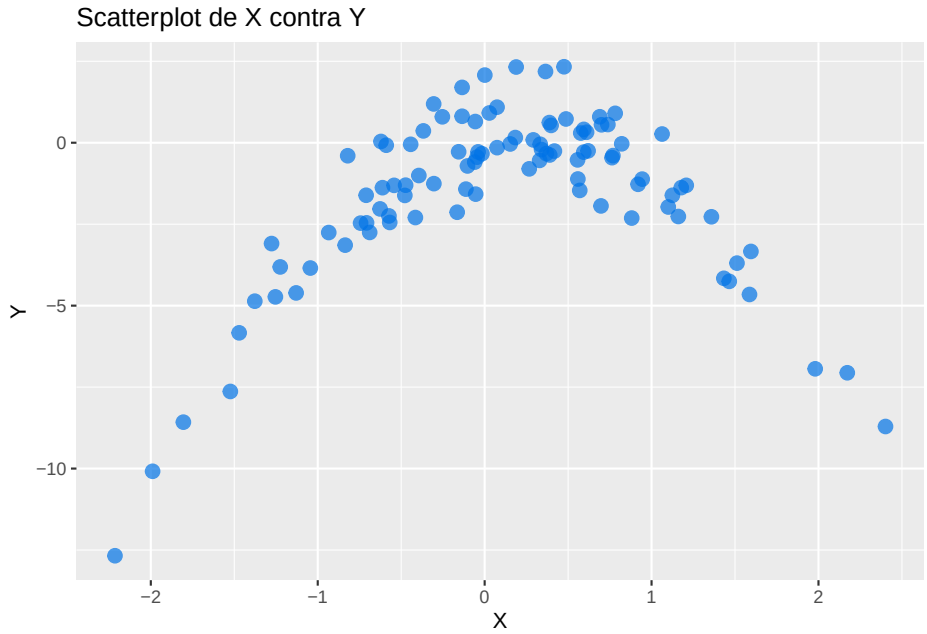
En este caso n = 100 y p = 1, y el modelo es:

$$Y = \beta_0 + \beta_1.X + \beta_2.X^2 + \epsilon$$

donde $\beta_0 = 0, \beta_1 = 1, \beta_2 = -2$ y ϵ es el error.

b)

```
ggplot(mapping = aes(x, y)) +
  geom_point(color = "#0073e6", size = 3, alpha = 0.7) +
  labs(title = "Scatterplot de X contra Y",
       x = "X",
       y = "Y")
```



Se observa una clara relación entre las variables, no lineal, se nota muy marcada una subida de los valores de Y en función de un aumento en X hasta un punto máximo donde la situación cambia y una vez pasado ese punto a medida que aumenta la X baja la Y. Destacar cerca del centro de los valores de X hay muchos puntos y no se aprecian tan claras estas relaciones.

c)

```
set.seed(1899)

data <- data.frame(x, y)

rmses <- c()
for (i in 1:4) {
  mod <- linear_reg() %>%
         set_engine("lm")

  fit <- mod %>%
         parsnip::fit(y ~ poly(x, degree = i, raw = TRUE), data = data)

  rmses[[i]] <- augment(fit,
                        new_data = data) %>%
                rmse(y, .pred)
}

bind_rows(as.data.frame(rmses[[1]]), as.data.frame(rmses[[2]]), as.data.frame(rmses[[3]]), as.data.frame
```

```
##   .metric .estimator .estimate
## 1    rmse   standard 2.5740055
## 2    rmse   standard 0.9435524
## 3    rmse   standard 0.9431827
## 4    rmse   standard 0.9347677
```

d)

```
set.seed(2002)

rmses <- c()
for (i in 1:4) {
  mod <- linear_reg() %>%
         set_engine("lm")

  fit <- mod %>%
         parsnip::fit(y ~ poly(x, degree = i, raw = TRUE), data = data)

  rmses[[i]] <- augment(fit,
                        new_data = data) %>%
                rmse(y, .pred)
}

bind_rows(as.data.frame(rmses[[1]]), as.data.frame(rmses[[2]]), as.data.frame(rmses[[3]]), as.data.frame
```

```
##   .metric .estimator .estimate
## 1    rmse   standard 2.5740055
## 2    rmse   standard 0.9435524
## 3    rmse   standard 0.9431827
## 4    rmse   standard 0.9347677
```

Son los mismos resultados, no hay aleatoriedad en esta parte.

e)

El modelo 2, el cuadrático, lo que tiene todo el sentido debido a lo qu habíamos observado en el scatterplot, donde se veía una especie de parábola.

f)

```
modelo1 <- y ~ x
modelo2 <- y ~ x + I(x^2)
modelo3 <- y ~ x + I(x^2) + I(x^3)
modelo4 <- y ~ x + I(x^2) + I(x^3) + I(x^4)

summary(glm(modelo4, data = data))
```

Viendo los resumenos se concuerda con la información vista en cv, los coeficientes extras asociados a los modelos 3 y 4 no son significativos, mientras que el modelo 2 ajusta mejor que el 1 siendo significativo en todos los parámetros.

Capítulo 7 Ejercicio 1

a)

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$

Como $x \leq \xi$, entonces:

$$f_1(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

donde $a_1 = \beta_0, b_1 = \beta_1, c_1 = \beta_2, d_1 = \beta_3$

b)

Como $x > \xi$, entonces:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$

Ahora reescribimos $(x - \xi)^3$ como:

$$(x - \xi)^3 = x^3 - 3\xi x^2 + 3\xi^2 x - \xi^3$$

Entonces:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x^3 - 3\xi x^2 + 3\xi^2 x - \xi^3)$$

Por lo tanto:

$$f(x) = (\beta_0 - \beta_4 \xi^3) + (\beta_1 + 3\beta_4 \xi^2) x + (\beta_2 + 3\beta_4 \xi) x^2 + (\beta_3 + \beta_4) x^3$$

Donde $a_2 = (\beta_0 - \beta_4 \xi^3), b_2 = (\beta_1 + 3\beta_4 \xi^2), c_2 = (\beta_2 + 3\beta_4 \xi), d_2 = (\beta_3 + \beta_4)$

c)

$$f_1(\xi) = \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3$$
$$f_2(\xi) = (\beta_0 - \beta_4 \xi^3) + (\beta_1 + 3\beta_4 \xi^2)\xi + (\beta_2 + 3\beta_4 \xi)\xi^2 + (\beta_3 + \beta_4)\xi^3$$
$$f_2(\xi) = \beta_0 - \beta_4 \xi^3 + \beta_1 \xi + 3\beta_4 \xi^3 + \beta_2 \xi^2 + 3\beta_4 \xi^3 + \beta_3 \xi^3$$
$$f_2(\xi) = \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3 + 2\beta_4 \xi^3$$
$$f_1(\xi) = f_2(\xi)$$

d)

$$f_1'(\xi) = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2$$
$$f_2'(\xi) = (\beta_1 + 3\beta_4 \xi^2) + 2(\beta_2 + 3\beta_4 \xi)\xi + 3(\beta_3 + \beta_4)\xi^2 = \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2 + 12\beta_4 \xi^2$$
$$f_1'(\xi) = f_2'(\xi)$$

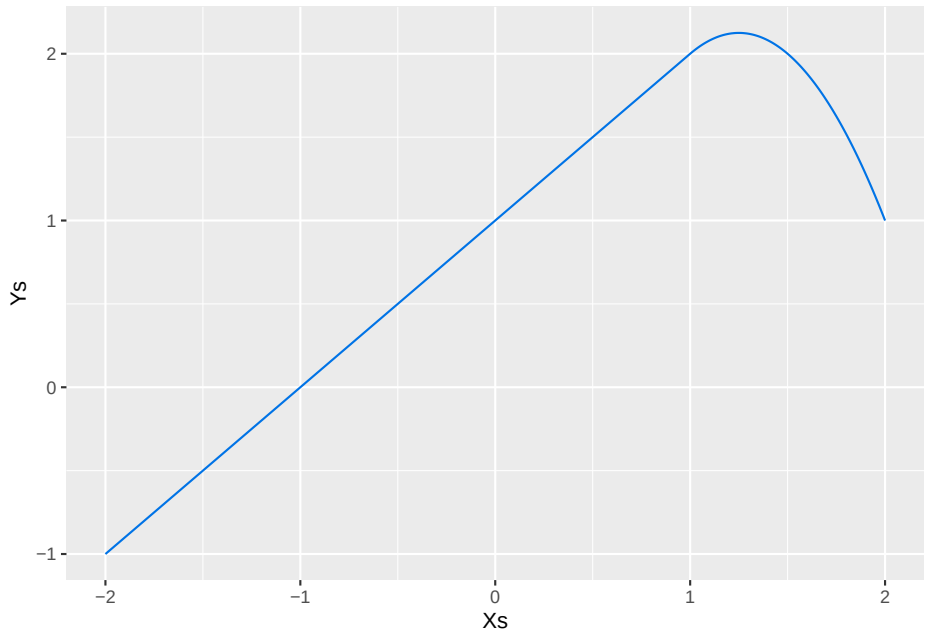
e)

$$f_1''(\xi) = 2\beta_2 + 6\beta_3 \xi$$
$$f_2''(\xi) = 2(\beta_2 + 3\beta_4 \xi) + 6(\beta_3 + \beta_4)\xi = 2\beta_2 + 6\beta_3 \xi + 6\beta_4 \xi$$
$$f_1''(\xi) = f_2''(\xi)$$

Capítulo 7 Ejercicio 2

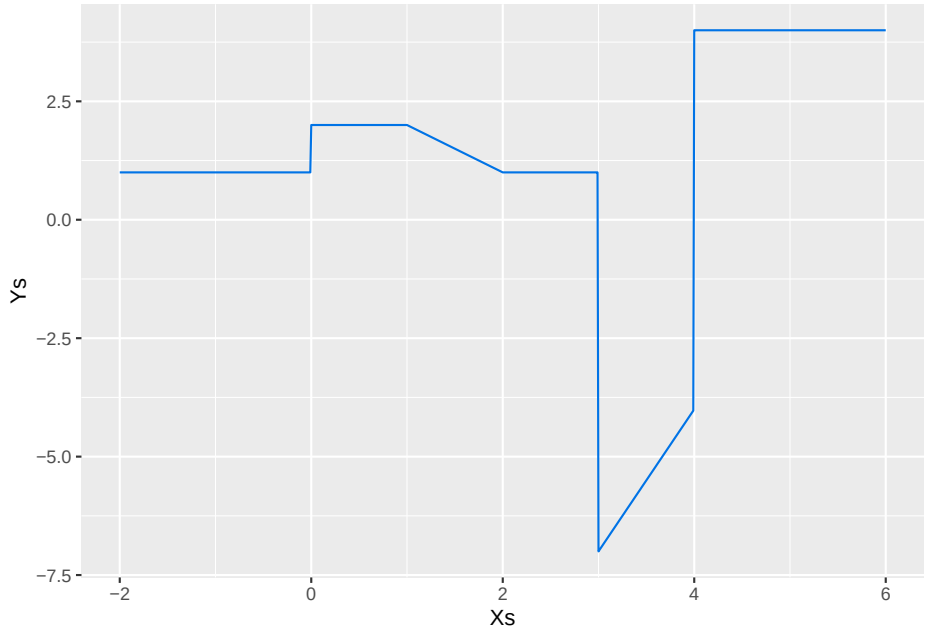
Capítulo 7 Ejercicio 3

```
b1 <- function(X) { X }
b2 <- function(X) { (X - 1)^2 * (X >= 1) }

Y_hat <- function(X) {
  ifelse(X < 1,
         1 + b1(X),
         1 + b1(X) - 2 * b2(X))
}

Xs <- seq(-2, 2, by = 0.01)
Ys <- Y_hat(Xs)

ggplot(mapping = aes(Xs, Ys)) +
  geom_line(col = "#0073e6")
```



Capítulo 7 Ejercicio 4

```
b1 <- function(X) { (0 <= X & X <= 2) - (X - 1) * (1 <= X & X <= 2) }
b2 <- function(X) { (X - 3) * (3 <= X & X <= 4) + (4 < X & X <= 5) }

Y_hat <- function(X) {
  ifelse(X < 0, 1,
         ifelse(X < 1, 2,
                ifelse(X <= 2,3 - X,
                       ifelse(X < 3, 1,
                              ifelse(X < 4, 3 * (X - 3) + 1 - 8,4))))
         )
}

Xs <- seq(-2, 6, by = 0.01)
Ys <- Y_hat(Xs)

ggplot(mapping = aes(Xs, Ys)) +
  geom_line(col = "#0073e6")
```



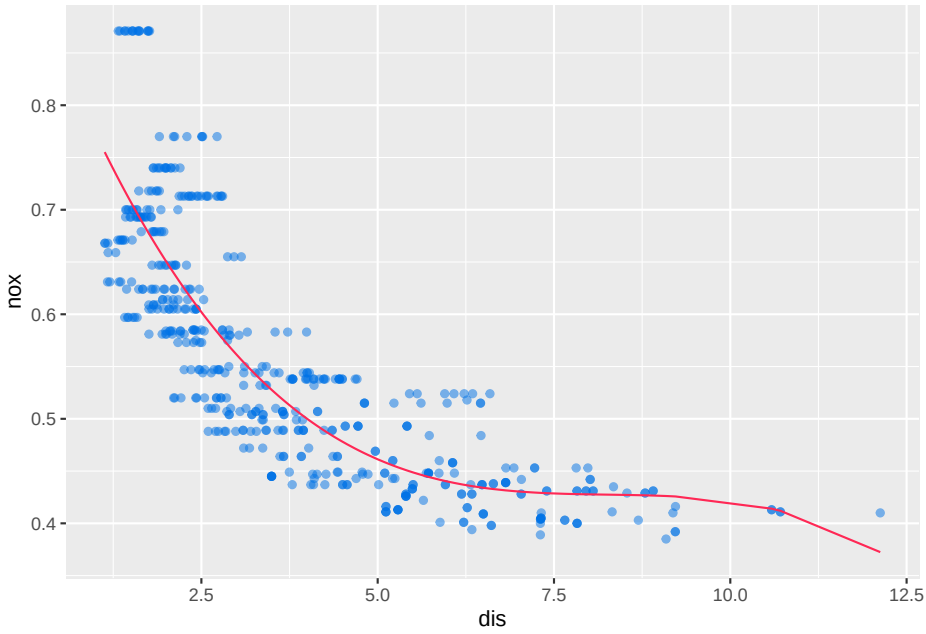
Capítulo 7 Ejercicio 9

a)

```
boston_data <- Boston

lm_fit <-
  parsnip::linear_reg() %>%
  parsnip::set_engine("lm") %>%
  parsnip::fit(nox ~ poly(dis, degree = 3, raw = TRUE),
               data = boston_data)

lm_fit %>%
  extract_fit_engine() %>%
  summary()
```

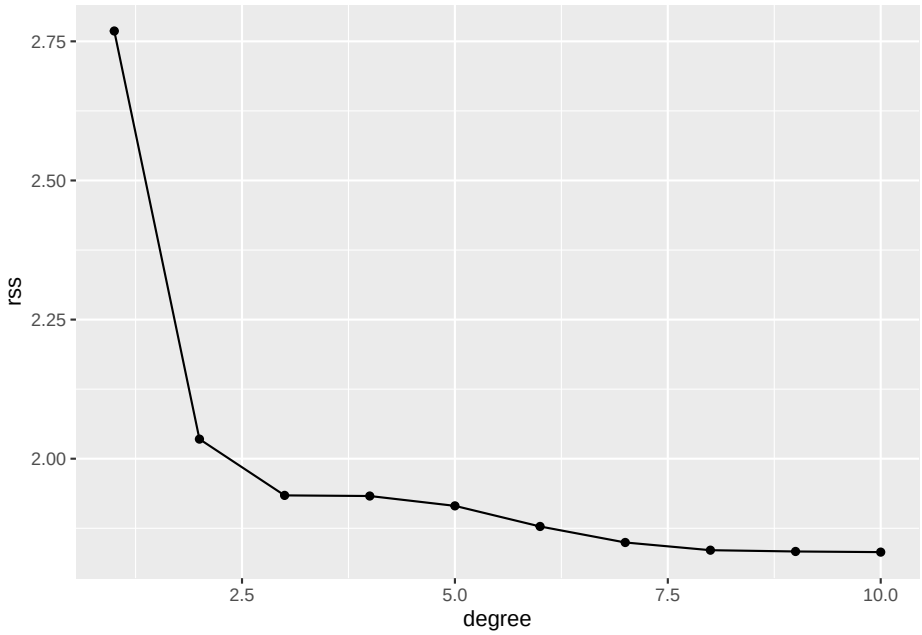
```
##
## Call:
## stats::lm(formula = nox ~ poly(dis, degree = 3, raw = TRUE),
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121130 -0.040619 -0.009738  0.023385  0.194904
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         0.9341281  0.0207076  45.110  < 2e-16 ***
## poly(dis, degree = 3, raw = TRUE)1 -0.1820817  0.0146973 -12.389  < 2e-16 ***
## poly(dis, degree = 3, raw = TRUE)2  0.0219277  0.0029329   7.476 3.43e-13 ***
## poly(dis, degree = 3, raw = TRUE)3 -0.0008850  0.0001727  -5.124 4.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06207 on 502 degrees of freedom
## Multiple R-squared:  0.7148, Adjusted R-squared:  0.7131
## F-statistic: 419.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
predictions <- predict(lm_fit, new_data = boston_data)
boston_data$predicts <- predictions$.pred

ggplot(boston_data, aes(x = dis, y = nox)) +
  geom_point(color = "#0073e6", alpha = 0.5) +
  geom_line(aes(y = predicts), color = "#FF2855")
```



b)

```
results <- list()

for (degree in 1:10) {
  lm_fit <-
    parsnip::linear_reg() %>%
    parsnip::set_engine("lm") %>%
    parsnip::fit(nox ~ poly(dis, degree = degree, raw = TRUE),
                 data = boston_data)

  boston_data[[paste0("predicts_", degree)]] <- predict(lm_fit, new_data = boston_data)$`.pred`

  rss <- sum((boston_data$nox - boston_data[[paste0("predicts_", degree)]])^2)

  results[[degree]] <- list(degree = degree, rss = rss)
}

rss_df <- bind_rows(lapply(results, function(x) data.frame(degree = x$degree, rss = x$rss)))

ggplot(rss_df, aes(x = degree, y = rss)) +
  geom_line() +
  geom_point()
```



c)

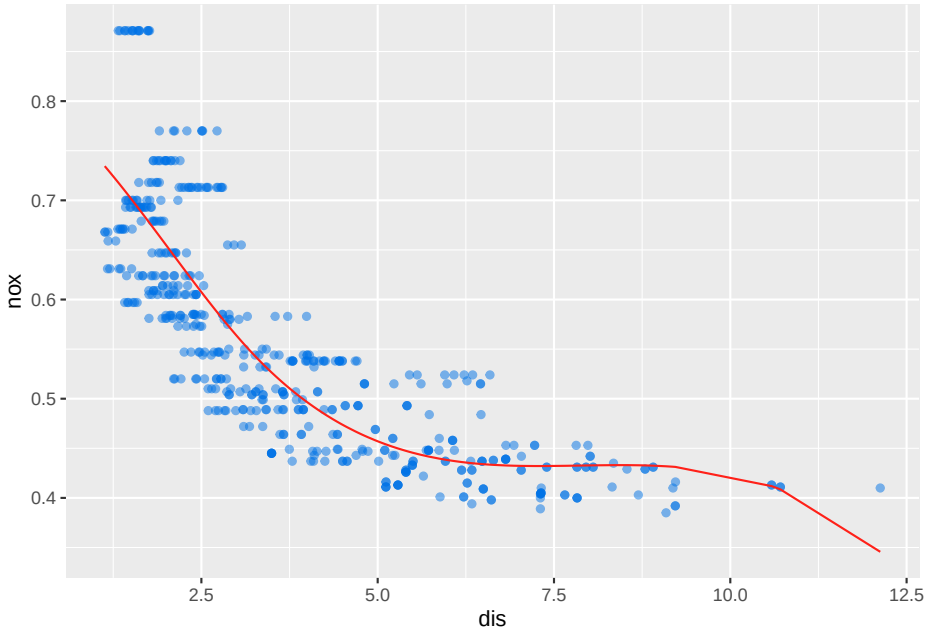
d)

```
spline_fit <-
  parsnip::linear_reg() %>%
  parsnip::set_engine("lm") %>%
  parsnip::fit(nox ~ bs(dis, df = 4), data = boston_data)

spline_fit %>%
  extract_fit_engine() %>%
  summary()
```

```
##
## Call:
## stats::lm(formula = nox ~ bs(dis, df = 4), data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.124622 -0.039259 -0.008514  0.020850  0.193891
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.73447    0.01460  50.306  < 2e-16 ***
## bs(dis, df = 4)1 -0.05810    0.02186  -2.658  0.00812 **
## bs(dis, df = 4)2 -0.46356    0.02366 -19.596  < 2e-16 ***
## bs(dis, df = 4)3 -0.19979    0.04311  -4.634 4.58e-06 ***
## bs(dis, df = 4)4 -0.38881    0.04551  -8.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.06195 on 501 degrees of freedom
## Multiple R-squared:  0.7164, Adjusted R-squared:  0.7142
## F-statistic: 316.5 on 4 and 501 DF,  p-value: < 2.2e-16
```

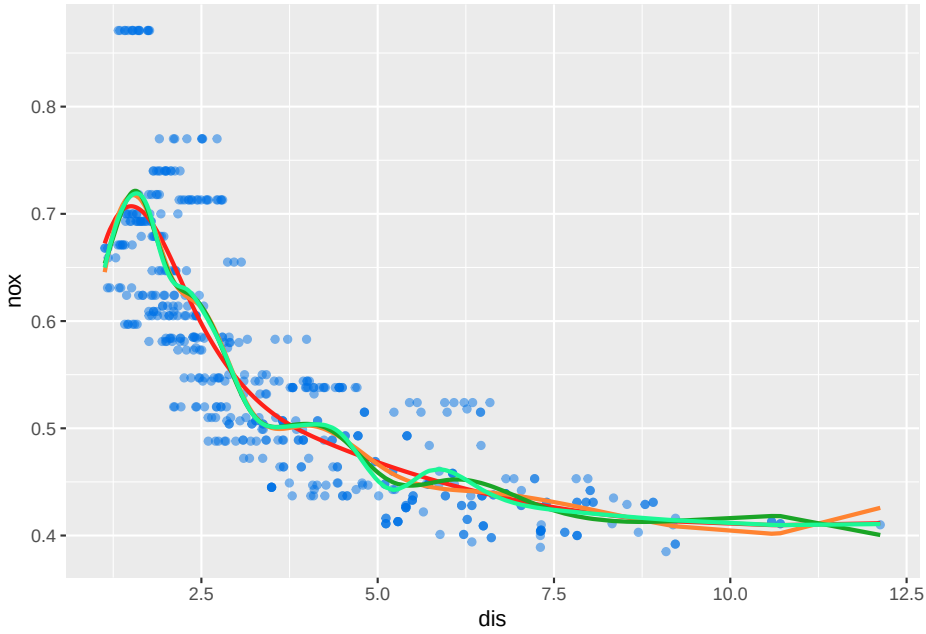
```r
boston_data$predicts_sp <- predict(spline_fit, new_data = boston_data)$`.pred`

ggplot(boston_data, aes(x = dis, y = nox)) +
  geom_point(color = "#0073e6", alpha = 0.5) +
  geom_line(aes(y = predicts_sp), color = "#FF2119")
```



e)

```r
rss_results <- data.frame(df = integer(), rss = numeric())

for (df in 5:20) {
  spline_fit <-
    parsnip::linear_reg() %>%
    parsnip::set_engine("lm") %>%
    parsnip::fit(nox ~ bs(dis, df = df), data = boston_data)

  predictions <- predict(spline_fit, new_data = boston_data)$`.pred`

  rss <- sum((boston_data$nox - predictions)^2)

  rss_results <- rbind(rss_results, data.frame(df = df, rss = rss))

  boston_data[[paste0("predicts__", df)]] <- predictions
}
```

```
ggplot(boston_data, aes(x = dis, y = nox)) +
  geom_point(color = "#0073e6", alpha = 0.5) +
  geom_line(aes(y = predicts__5), color = "#FF2119", size = 1) +
  geom_line(aes(y = predicts__10), color = "#FF8331", size = 1) +
  geom_line(aes(y = predicts__14), color = "#1AA526", size = 1) +
  geom_line(aes(y = predicts__18), color = "#1BF895", size = 1)
```




```
rss_results
```

```
##    df      rss
## 1   5 1.840173
## 2   6 1.833966
## 3   7 1.829884
## 4   8 1.816995
## 5   9 1.825653
## 6  10 1.792535
## 7  11 1.796992
## 8  12 1.788999
## 9  13 1.782350
## 10 14 1.781838
## 11 15 1.782798
## 12 16 1.783546
## 13 17 1.779789
## 14 18 1.775838
## 15 19 1.774487
## 16 20 1.776727
```

f)

Capítulo 7 Ejercicio 10

a)

```
college_data <- College

set.seed(1899)

college_split <- initial_split(college_data,
                               prop = 4/5)

college_train <- training(college_split)
college_test <- testing(college_split)

mod <-
  parsnip::linear_reg() %>%
  parsnip::set_engine("lm")

completo <-
  mod |>
  parsnip::fit(Outstate ~ ., data = college_train)

completo$fit %>%
  summary()
```

```
##
## Call:
## stats::lm(formula = Outstate ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6322.3 -1281.1   -98.8  1324.6  5399.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.031e+03  8.602e+02  -2.361   0.0185 *
## PrivateYes   2.309e+03  2.812e+02   8.210 1.34e-15 ***
## Apps        -3.072e-01  7.346e-02  -4.182 3.32e-05 ***
## Accept       8.231e-01  1.448e-01   5.685 2.04e-08 ***
## Enroll      -5.052e-01  3.854e-01  -1.311   0.1904
## Top10perc    3.009e+01  1.257e+01   2.394   0.0170 *
## Top25perc   -5.897e+00  9.644e+00  -0.611   0.5412
## F.Undergrad -1.147e-01  6.581e-02  -1.742   0.0819 .
## P.Undergrad  1.367e-02  6.459e-02   0.212   0.8324
## Room.Board   9.126e-01  9.636e-02   9.470  < 2e-16 ***
## Books       -2.783e-01  4.876e-01  -0.571   0.5684
## Personal    -2.125e-01  1.296e-01  -1.640   0.1015
## PhD          1.273e+01  1.028e+01   1.239   0.2159
## Terminal     2.685e+01  1.071e+01   2.507   0.0125 *
## S.F.Ratio   -3.827e+01  2.701e+01  -1.417   0.1570
## perc.alumni  3.951e+01  8.628e+00   4.579 5.68e-06 ***
## Expend       1.855e-01  2.421e-02   7.661 7.37e-14 ***
## Grad.Rate    2.514e+01  6.137e+00   4.097 4.75e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1983 on 603 degrees of freedom
## Multiple R-squared:  0.7718, Adjusted R-squared:  0.7654
## F-statistic:   120 on 17 and 603 DF,  p-value: < 2.2e-16
```

```
fit <-
  mod |>
  parsnip::fit(Outstate ~ Private + Apps + Accept + Top10perc + Room.Board +
                       Terminal + perc.alumni + Expend + Grad.Rate, data = college_train)

fit$fit %>%
  summary()
```

```
##
## Call:
## stats::lm(formula = Outstate ~ Private + Apps + Accept + Top10perc +
##     Room.Board + Terminal + perc.alumni + Expend + Grad.Rate,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8027.5 -1289.8   -92.1  1360.4  5940.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.204e+03  5.476e+02  -7.678 6.45e-14 ***
## PrivateYes   2.842e+03  2.607e+02  10.901  < 2e-16 ***
## Apps        -2.715e-01  7.221e-02  -3.760 0.000186 ***
## Accept       4.364e-01  1.084e-01   4.024 6.43e-05 ***
## Top10perc    1.680e+01  7.341e+00   2.289 0.022420 *
## Room.Board   9.787e-01  9.583e-02  10.212  < 2e-16 ***
## Terminal     3.596e+01  7.147e+00   5.032 6.38e-07 ***
## perc.alumni  4.210e+01  8.637e+00   4.874 1.40e-06 ***
## Expend       2.061e-01  2.181e-02   9.448  < 2e-16 ***
## Grad.Rate    3.069e+01  6.061e+00   5.064 5.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2025 on 611 degrees of freedom
## Multiple R-squared:  0.7589, Adjusted R-squared:  0.7553
## F-statistic: 213.7 on 9 and 611 DF,  p-value: < 2.2e-16
```

```
rmse_linear <- augment(fit,
                       new_data = college_train) %>%
               rmse(Outstate, .pred)

rmse_linear
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       2008.
```

b)

```
rec_gam <- recipes::recipe(Outstate ~ Private + Apps + Accept + Top10perc + Room.Board +
                                      Terminal + perc.alumni + Expend + Grad.Rate,
                           data = college_train)

mgcv_spec <- parsnip::gen_additive_mod() %>%
  parsnip::set_engine("mgcv") %>%
  parsnip::set_mode("regression")

gam_wf <- workflows::workflow() %>%
  workflows::add_recipe(rec_gam) %>%
  workflows::add_model(mgcv_spec, formula = Outstate ~ Private + Apps + Accept + Top10perc + Room.Board
                                      Terminal + perc.alumni + Expend + Grad.Rate)

gam_fit <- parsnip::fit(gam_wf, data = college_data)
gam <- extract_fit_engine(gam_fit)

# gratia::draw(gam, residuals = T)
```

c)

```
rmse_gam_test <- augment(gam_fit,
                    new_data = college_test) %>%
                 rmse(Outstate, .pred)

rmse_gam_test
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       1896.
```

d)

```
gam_fit %>%
  extract_fit_engine() %>%
  summary()
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Outstate ~ Private + Apps + Accept + Top10perc + Room.Board +
##     Terminal + perc.alumni + Expend + Grad.Rate
##
## Parametric coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.954e+03  4.969e+02  -7.957 6.33e-15 ***
## PrivateYes   2.759e+03  2.318e+02  11.901  < 2e-16 ***
## Apps        -2.790e-01  6.701e-02  -4.164 3.48e-05 ***
## Accept       4.450e-01  1.003e-01   4.438 1.04e-05 ***
## Top10perc    1.875e+01  6.452e+00   2.906  0.00377 **
## Room.Board   9.565e-01  8.509e-02  11.241  < 2e-16 ***
## Terminal     3.246e+01  6.427e+00   5.051 5.50e-07 ***
## perc.alumni  4.474e+01  7.587e+00   5.897 5.55e-09 ***
## Expend       2.205e-01  2.063e-02  10.689  < 2e-16 ***
## Grad.Rate    2.914e+01  5.453e+00   5.343 1.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.753   Deviance explained = 75.5%
## GCV = 4.0566e+06  Scale est. = 4.0044e+06  n = 777
```