# Introducción a la analítica

Profesores César Augusto Gómez, Mauricio Alejandro Mazo y Juan Carlos Salazar

# CLASIFICADORES COMUNES EN AE

Similar a lo que sucede en regresión, no hay una fuerte relación entre la tasa de error del conjunto de entrenamiento y la del conjunto de prueba. De hecho, a medida que se usan métodos de clasificación muy flexibles, la tasa de error de entrenamiento declinará pero la tasa de error de prueba podría no hacerlo. La siguiente figura ilustra esta afirmación:

# CLASIFICADORES COMUNES EN AE

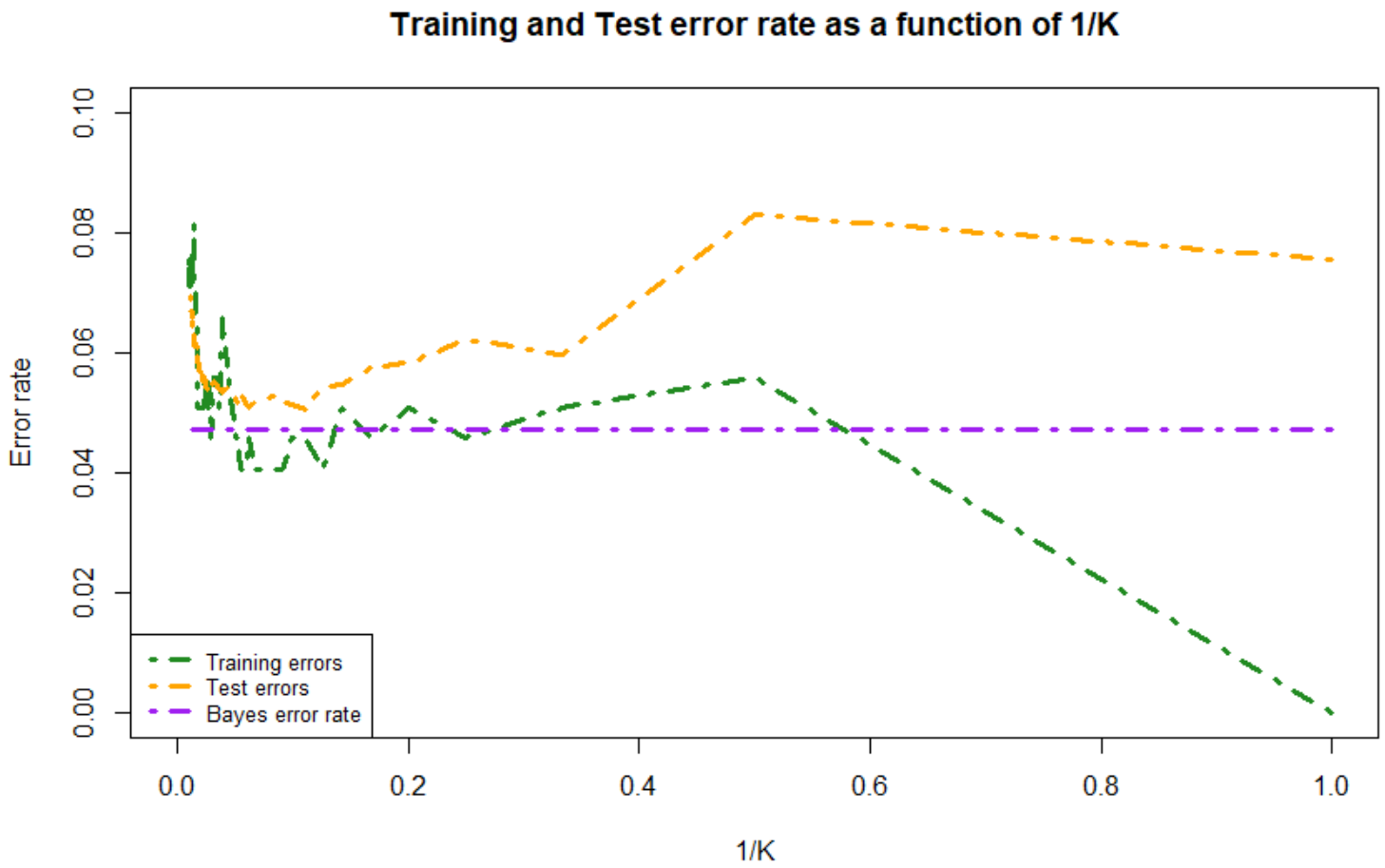



Figura 1: Training and Test Error rates using Knn

# ¿CÓMO USAR KNN CON VARIABLES CATEGÓRICAS COMO FEATURES?

Naive Bayes está diseñado para manejar automáticamente features continuas y categoricas simultáneamente. Knn solo trabaja con features numéricas; por lo tanto, si se quiere implementar Knn y hay features categóricas estas se deben convertir a variables dummy. Por ejemplo, si el sujeto $i$ tiene una variable status con tres niveles: Single, Married y Divorced y esa persona es Single, su variable dummy sera $(1, 0, 0)$. Afortunadamente, hay una librería en R que permite convertir variables categóricas a variables dummy, tal librería se llama *dummies* y se invoca con library(dummies) una vez instalada.

# ¿CÓMO USAR KNN CON VARIABLES CATEGÓRICAS COMO FEATURES?

**EJEMPLO: NAIVE BAYES VERSUS KNN**. Considere el siguiente conjunto de datos

| ID | Home_Owner | marital_Status | Job_Experience | Defaulted |
|---|---|---|---|---|
| 1 | Yes | Single | 3 | No |
| 2 | No | Married | 4 | No |
| 3 | No | Single | 5 | No |
| 4 | Yes | Married | 4 | No |
| 5 | No | Divorced | 2 | Yes |
| 6 | No | Married | 4 | No |
| 7 | Yes | Divorced | 2 | No |
| 8 | No | Married | 3 | Yes |
| 9 | No | Married | 3 | No |
| 10 | Yes | Single | 2 | Yes |
| 11 | No | Married | 3 | No |
| 12 | No | Married | 4 | Yes |
| 13 | Yes | Married | 3 | Yes |
| 14 | No | Single | 5 | No |
| 15 | Yes | Divorced | 2 | No |
| 16 | Yes | Divorced | 2 | No |
| 17 | Yes | Married | 1 | Yes |
| 18 | No | Single | 5 | No |
| 19 | Yes | Single | 1 | Yes |
| 20 | No | Married | 2 | Yes |

# ¿CÓMO USAR KNN CON VARIABLES CATEGÓRICAS COMO FEATURES?

Se implementará Naive Bayes y Knn y se calcularán sus errores:

```
library(class)
library(naivebayes)
library(dummies)
set.seed(123)
defaulted <- read.csv("Defaulted_3.csv",
                      stringsAsFactors=FALSE,sep=';')
#defaulted <- read.csv("Defaulted_3.csv",
#                       stringsAsFactors=FALSE,sep=';')
df=data.frame(defaulted)
## 100*P% of the sample size
smp_size <- floor(0.6 * nrow(df))
train_ind <- sample(seq_len(nrow(df)), size = smp_size)
train <- df[train_ind, ]
test <- df[-train_ind, ]
y_train=df[train_ind,5]
y_test=df[-train_ind,5]
normalize <- function(x) {
  norm <- ((x - min(x))/(max(x) - min(x)))
  return (norm)
}
```

# ¿CÓMO USAR KNN CON VARIABLES CATEGÓRICAS COMO FEATURES?

```
###########Naive Bayes#################################
defaulted_classifier <- naive_bayes(Defaulted~Home_Owner
                                    +marital_Status+Job_Experience,
                                    data=train[,2:5],laplace=0.128)
predict_train<-predict(defaulted_classifier,newdata=train[,2:4],type="class")
predict_test<-predict(defaulted_classifier,newdata=test[,2:4],type="class")
t<-table(predict_train,y_train)
t1<-table(predict_test,y_test)
Train_error_NB<-(t[1,2]+t[2,1])/(sum(t))
Test_error_NB<-(t1[1,2]+t1[2,1])/(sum(t1))
```

# ¿CÓMO USAR KNN CON VARIABLES CATEGÓRICAS COMO FEATURES?

Se implementará Naive Bayes y Knn y se calcularán sus errores:

```
#######Knn#########################
dummyHO<-dummy(df$Home_Owner ,sep="_")
dummyMS<-dummy(df$marital_Status, sep="_")
job<-normalize(df$Job_Experience)
Newdata<-cbind(df,dummyHO,dummyMS,job)
train1 <- Newdata[ train_ind,6:11 ]
test1  <- Newdata[-train_ind,6:11 ]
y_train1=df[train_ind,5]
y_test1=df[-train_ind,5]
fit.knn_train<-knn(train=train1, test=train1,cl=y_train1, k=1, prob=TRUE,use.all=TRUE)
fit.knn_Test<-knn(train=train1, test=test1,cl=y_train1, k=1, prob=TRUE)
Predicted_train<-factor(fit.knn_train)
Predicted_test<-factor(fit.knn_Test)
t<-table(Predicted_train,y_train)
t1<-table(Predicted_test,y_test1)
Train_error_Knn<-(sum(t[1,2],t[2,1]))/(sum(t))
Test_error_Knn<-(sum(t1[1,2],t1[2,1]))/(sum(t1))
list(Train_error_NB=Train_error_NB,Test_error_NB=Test_error_NB,
     Train_error_Knn=Train_error_Knn,Test_error_Knn=Test_error_Knn)
```

```
## $Train_error_NB
## [1] 0.08333333
##
## $Test_error_NB
## [1] 0.75
##
## $Train_error_Knn
## [1] 0
##
## $Test_error_Knn
## [1] 0.75
```

La Regresión Lineal es una herramienta útil y poderosa para predecir una respuesta cuantitativa y se considera una herramienta de Aprendizaje Estadístico. Adaptaciones recientes de regresión lineal permiten enfrentar problemas donde esta tiene un desempeño pobre, como es el caso de Regresión Ridge (RR) y Regresión LASSO (LASSO) que se discuten más adelante. Por lo tanto, comprender bien en qué consiste la RL antes de estudiar RR o LASSO es de vital importancia.

Se retomará la base de datos de Advertising (Gastos en publicidad: Tv, Radio, Newspaper). El gráfico de ventas versus Tv, Radio y Newspaper:

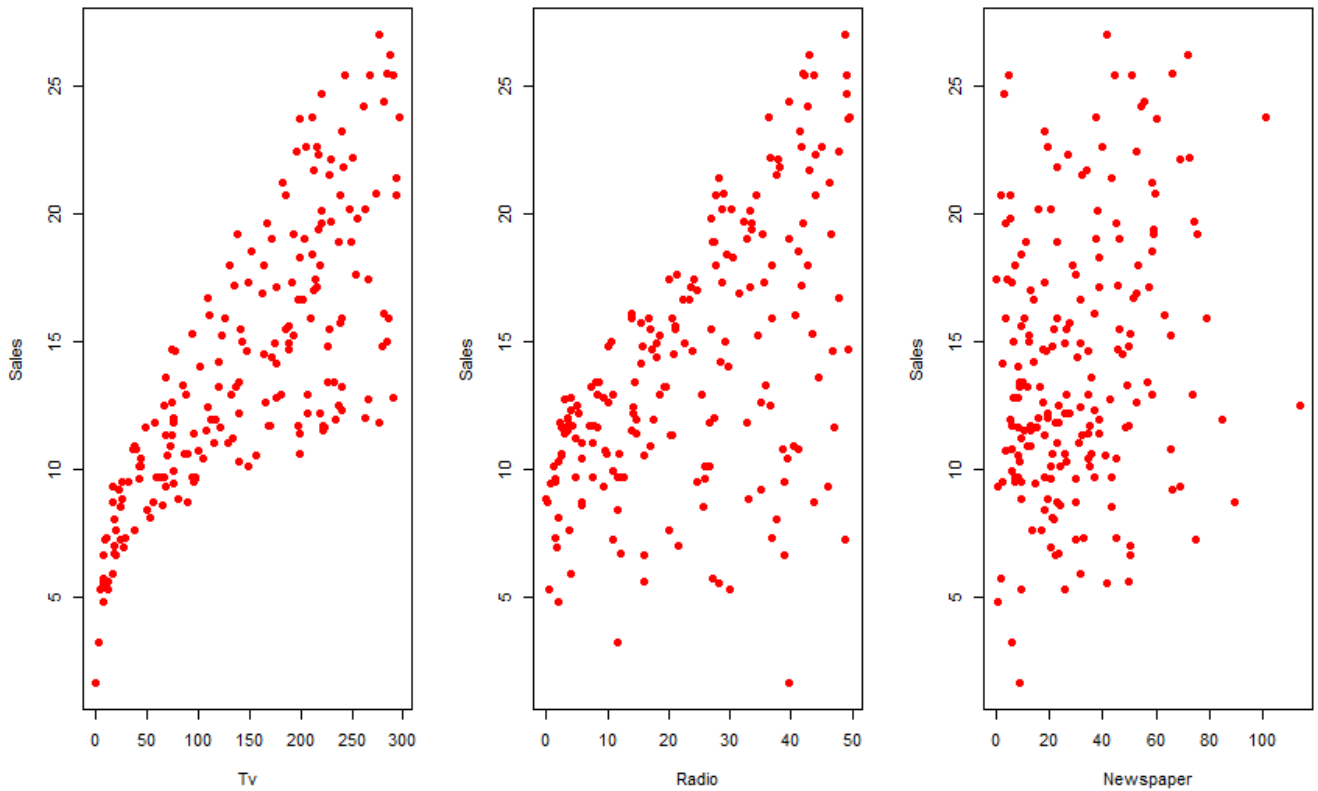



Figura 2: Sales versus advertising in Tv, Radio, and Newspaper

Si se nos contrata como estadísticos (o científico de datos o Bussiness Inteligence (BI) analyst) para recomendar, con base en estos datos, un plan de marketing para el próximo año que incremente las ventas ¿Qué información sería útil a fin de proporcionar tal recomendación? Estas serían algunas preguntas importantes a responder:

- ¿Hay alguna relación entre el presupuesto en publicidad y ventas?
- ¿Qué tan fuerte es la relación entre el presupuesto en publicidad y ventas?
- ¿Cuál medio contribuye más a las ventas?
- ¿Qué tan preciso se puede estimar el efecto de cada medio en las ventas?

- ¿Qué tan preciso podemos estimar ventas futuras?
- ¿Es la relación lineal?
- ¿Hay sinergia entre los medios publicitarios?

La RL se puede usar para tratar de responder cada una de estas interesantes preguntas.

## REGRESIÓN LINEAL SIMPLE.

La RL SIMPLE asume que hay una relación aproximadamente lineal entre una respuesta cuantitativa $Y$ y un predictor o feature $X$. Matemáticamente, esta relación lineal se puede escribir como:

$$Y \approx \beta_0 + \beta_1 X$$

Por ejemplo,

$$Sales \approx \beta_0 + \beta_1 TV$$

o

$$Sales \approx \beta_0 + \beta_1 Radio$$

$\beta_0$: Intercepto de la recta, $\beta_1$: Pendiente de la recta. Son dos constantes desconocidas y se conocen como parámetros o coeficientes del modelo.

**REGRESIÓN LINEAL SIMPLE**.

Una vez que se usan los datos de entrenamiento para obtener estimaciones $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de los coeficientes del modelo, respectivamente, se pueden predecir ventas futuras con base, por ejemplo, en un valor particular de publicidad por TV, al calcular:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Donde $\widehat{y}$ indica una predicción de $Y$ dado $X = x$.

**REGRESIÓN LINEAL SIMPLE**. Estimación de $\beta_0$ y $\beta_1$. Generalmente, $\beta_0$ y $\beta_1$ son desconocidos, por lo tanto se pueden usar los datos de entrenamiento $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ para obtener las estimaciones $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de estos coeficientes. En el ejemplo de Advertising, estos datos consisten de presupuesto en publicidad por TV y ventas del producto en $n = 200$ mercados. El objetivo, es obtener, usando estos datos, estimaciones $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de tal forma que el modelo lineal $Y \approx \beta_0 + \beta_1 X$ ajuste bien, es decir $y_i \approx \widehat{\beta}_0 + \widehat{\beta}_1 X$ para $i = 1, 2, \ldots, n$.

En otras palabras, se requiere encontrar $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de tal manera que la línea resultante sea tan cercana como sea posible a las $n = 200$ observaciones. Usualmente, el método más usado es el de Mínimos Cuadrados Ordinarios (OLS), pero hay otras aproximaciones que en algunos escenarios pueden tener un mejor desempeño que OLS (se discuten más adelante en el curso). Sea $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ una predicción para $Y = y_i$ con base en el $i$-ésimo valor para $X = x_i$. Entonces

$$e_i = y_i - \widehat{y}_i \quad \text{, representa el } i\text{-ésimo residual.}$$

La suma de cuadrados de los residuales (RSS) se define:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

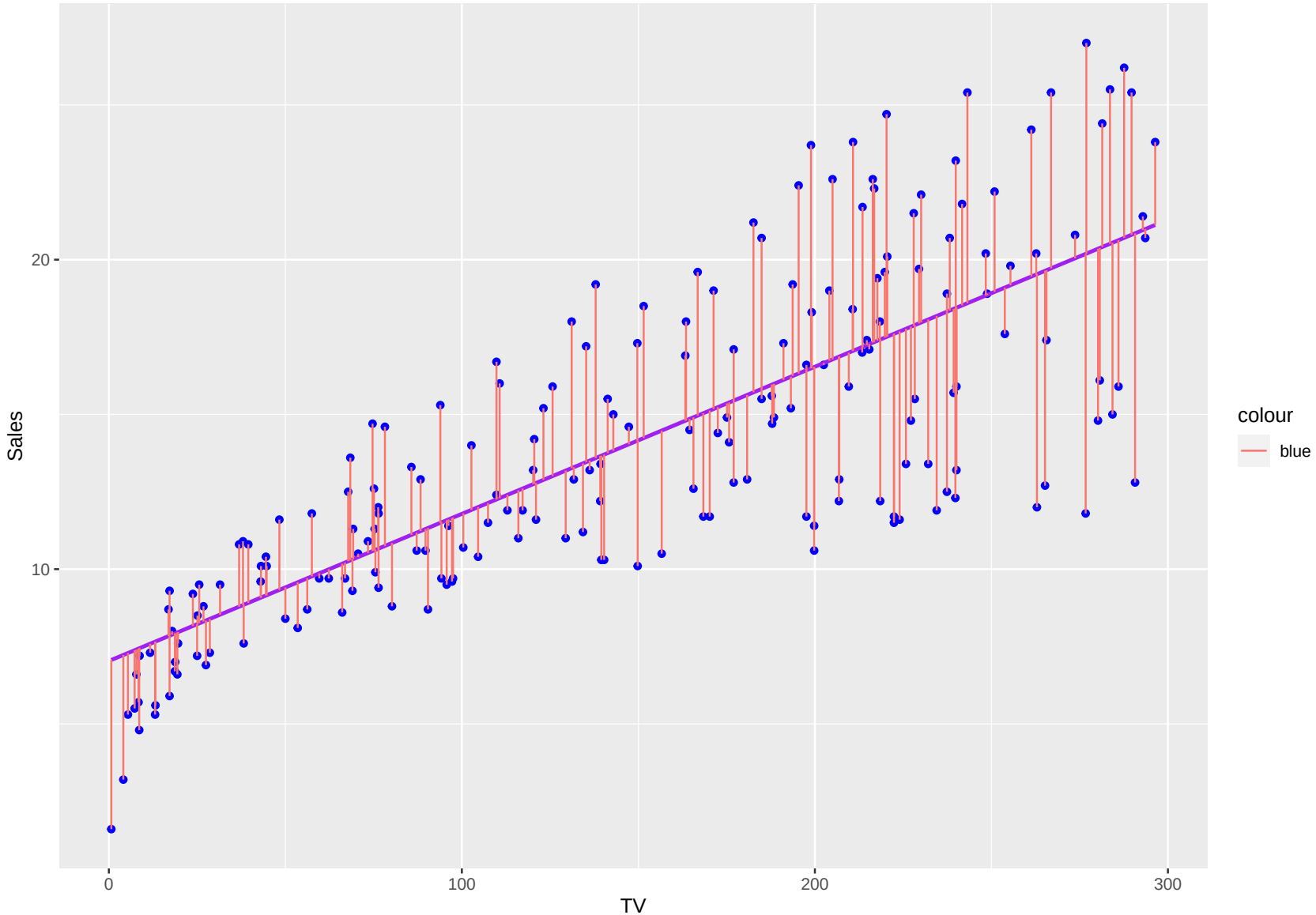
o equivalentemente

$$RSS = \sum_{i=1}^{n} \left( \widehat{y}_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right)^2$$

OLS genera $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que MINIMIZAN el RSS y producen la recta:

```
##
## Call:
## lm(formula = Sales ~ Tv, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## Tv          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

# REGRESIÓN LINEAL



Los estimadores de OLS se obtienen al minimizar la ecuación para el RSS

$$RSS = \sum_{i=1}^{n} \left( \widehat{y}_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right)^2$$

están dados por:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

y

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

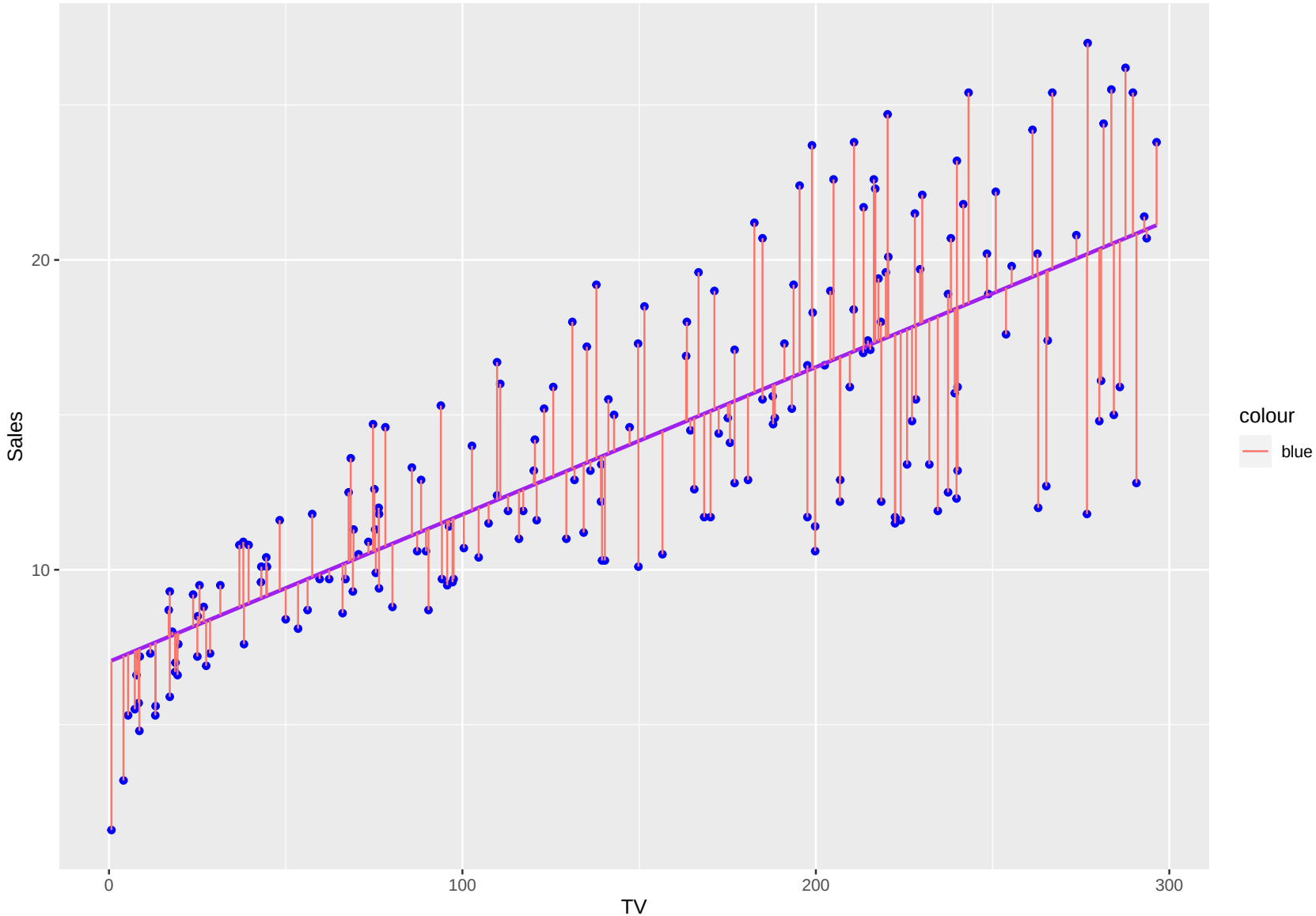
Donde $\bar{y}$ y $\bar{x}$ son las medias muestrales.

El siguiente programa en R genera las estimaciones de OLS para los datos de Advertising

```
library(ISLR)
library(ggplot2)
library(MASS)
 Advertising<-read.csv(file="Advertising.csv"
                       ,header=T,sep=',',dec='.')
#Advertising<-read.csv(file="Advertising.csv"
#                       ,header=T,sep=',',dec='.')
Sales=Advertising$sales
Tv=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
mod <- lm(Sales ~ Tv, data = Advertising)
summary(mod)
```

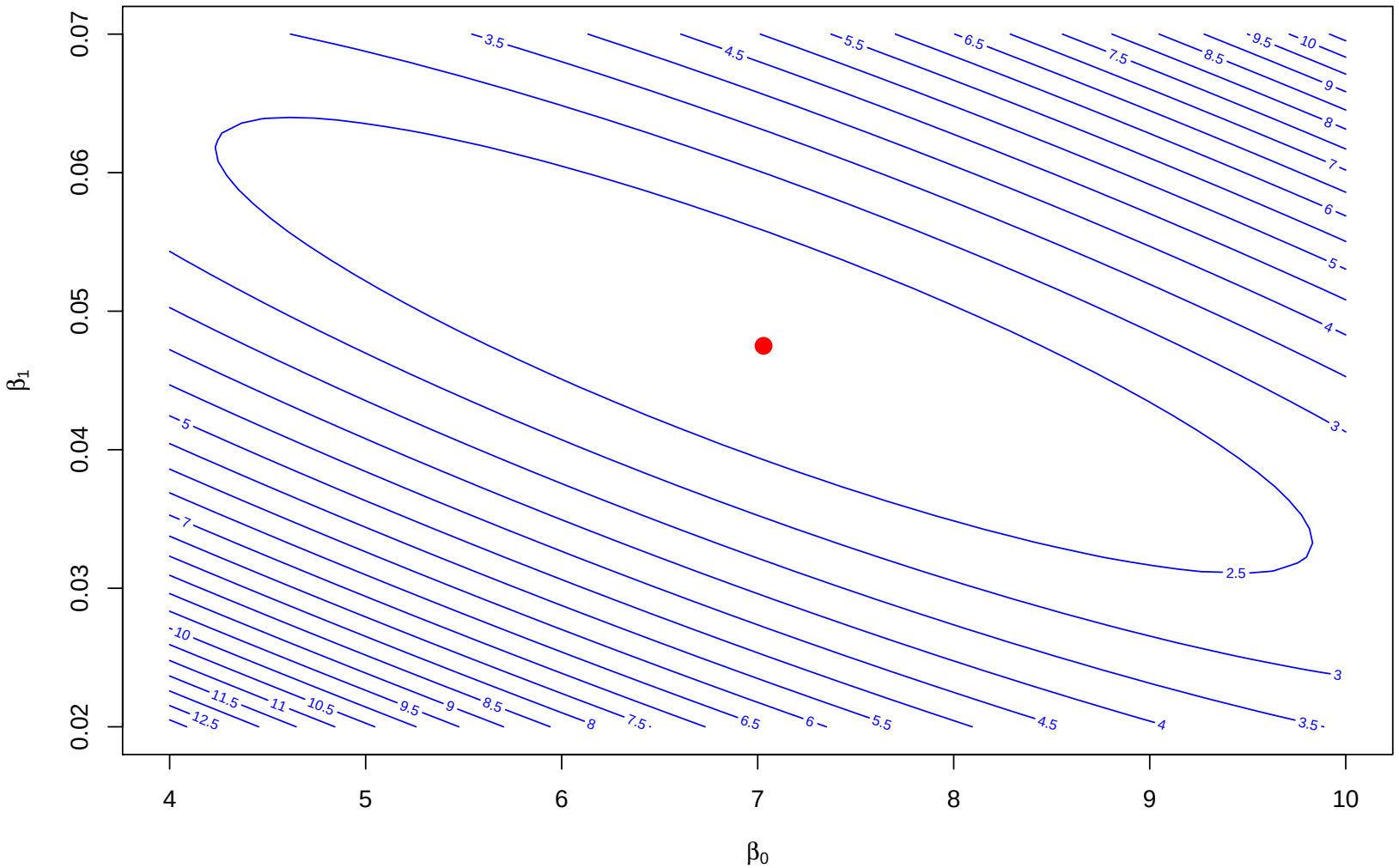
```
##
## Call:
## lm(formula = Sales ~ Tv, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## Tv          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

# REGRESIÓN LINEAL

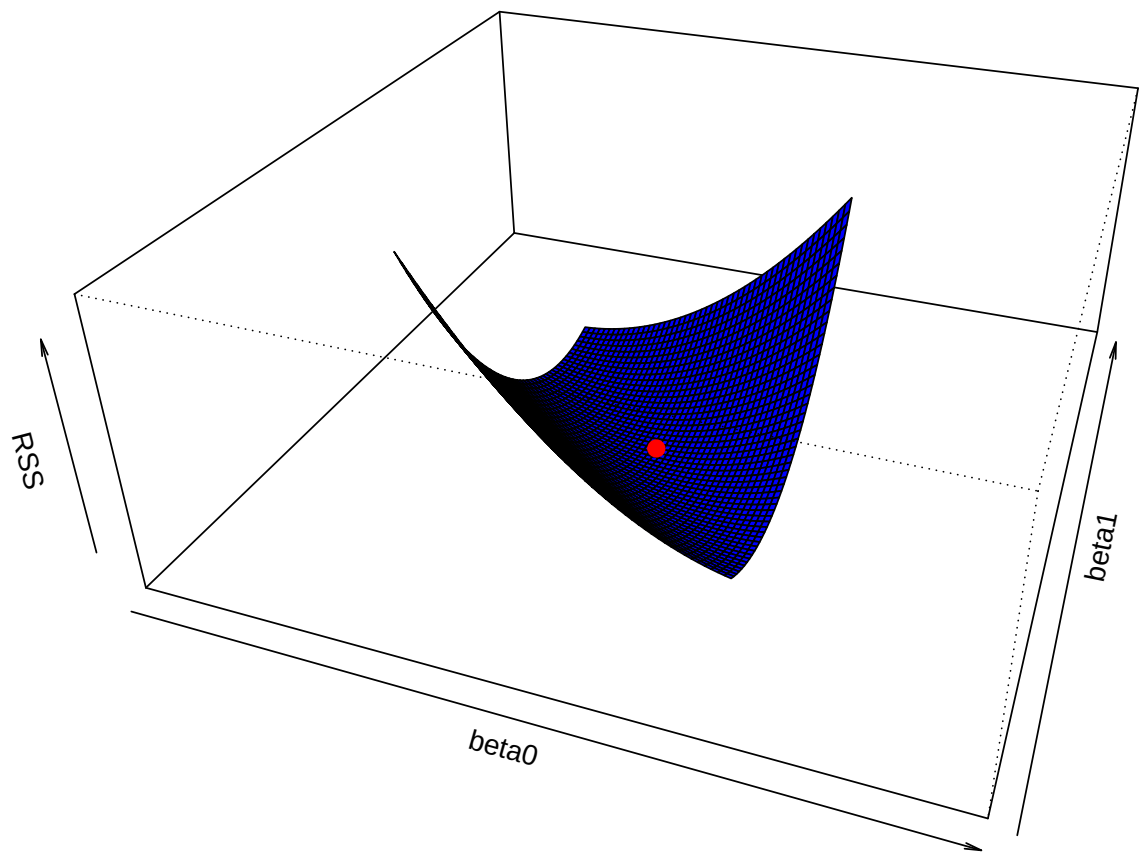


Note que $\widehat{\beta}_0 = 7.03$ y $\widehat{\beta}_1 = 0.0475$. La estimación para la pendiente $\beta_1$ significa que un gasto adicional de 1000 dolares en publicidad por televisión, se asocia con un incremento en ventas de aproximadamente 47.5 unidades del producto. El significado geométrico del vector $\widehat{\theta} = \left(\widehat{\beta}_0 = 7.03, \widehat{\beta}_1 = 0.0475\right)$, como aquel donde se minimiza el RSS, se ilustra en los siguientes slides, donde se grafica el RSS para distintos valores de $\beta_0$ y $\beta_1$. Observe que al RSS lo minimiza el vector $\widehat{\theta}$:

Considerando todos los $n = 200$ datos. Gráfico de contornos, el punto rojo es $\widehat{\theta} = \left(\widehat{\beta}_0 = 7.03, \widehat{\beta}_1 = 0.0475\right)$:



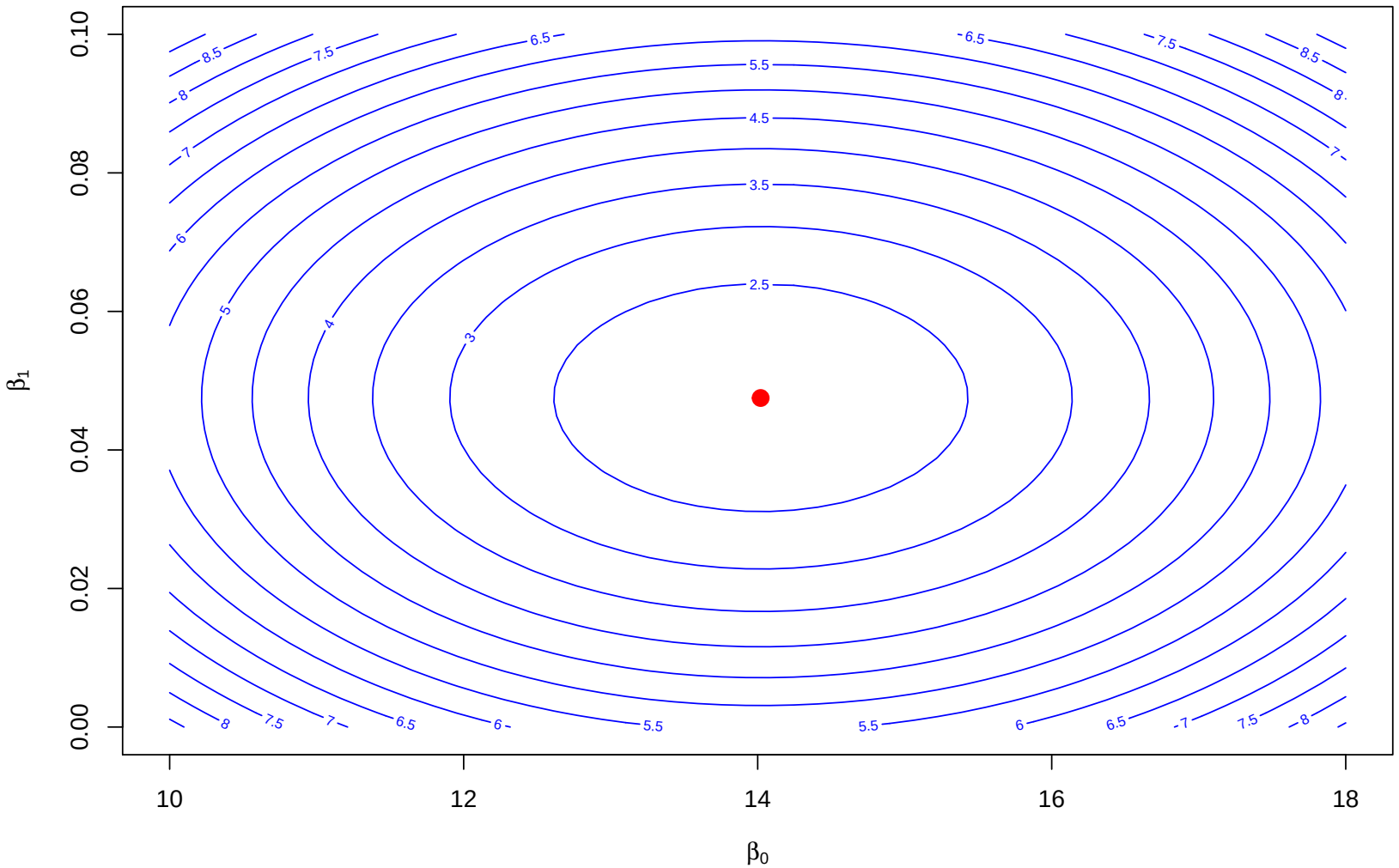
Considerando todos los $n = 200$ datos. Gráfico 3D, el punto rojo es $\widehat{\theta} = \left(\widehat{\beta}_0 = 7.03, \widehat{\beta}_1 = 0.0475\right)$:



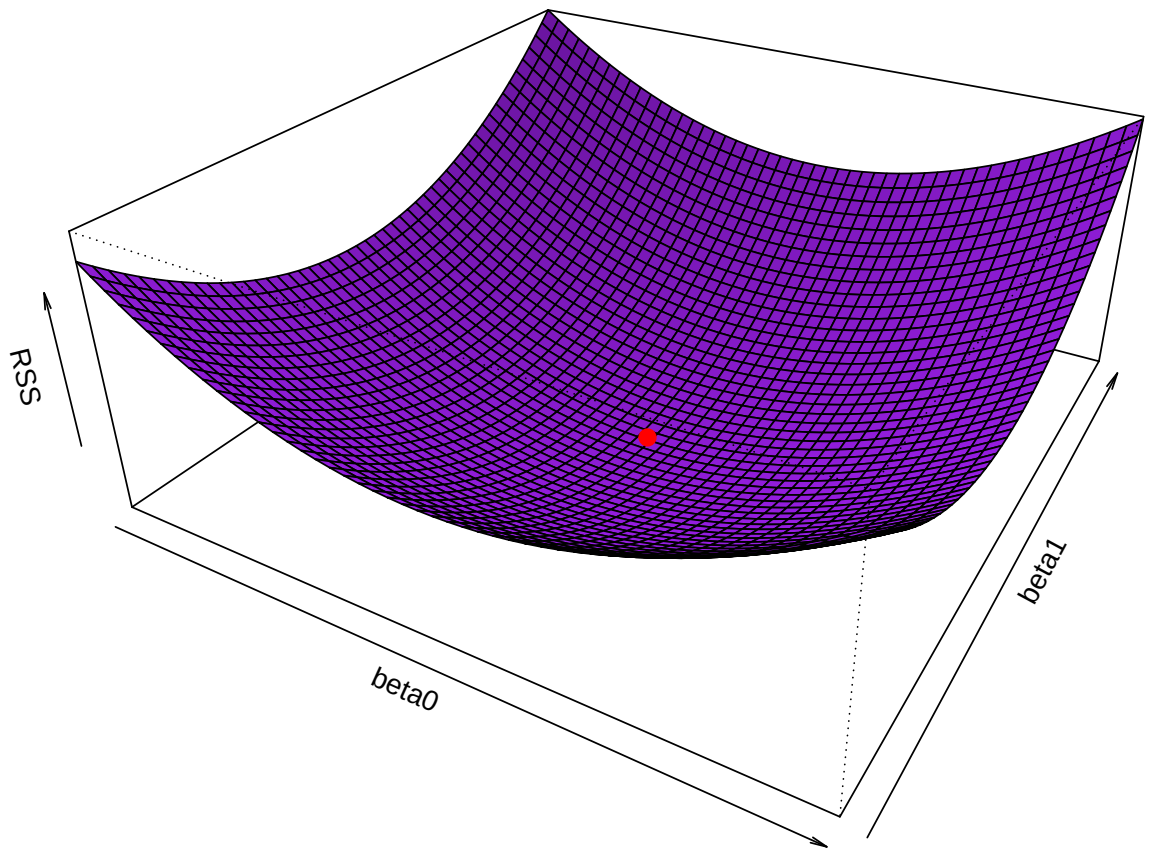
Considerando la feature gastos en TV, centrada respecto su promedio, el punto rojo es $\widehat{\theta} = \left(\widehat{\beta}_0 = 14.02, \widehat{\beta}_1 = 0.0475\right)$:

```
##
## Call:
## lm(formula = Sales ~ Tv, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.022500   0.230422   60.86   <2e-16 ***
## Tv           0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

El punto rojo es $\widehat{\theta} = \left(\widehat{\beta}_0 = 14.02, \widehat{\beta}_1 = 0.0475\right)$:



El punto rojo es $\widehat{\theta} = \left(\widehat{\beta}_0 = 14.02, \widehat{\beta}_1 = 0.0475\right)$:



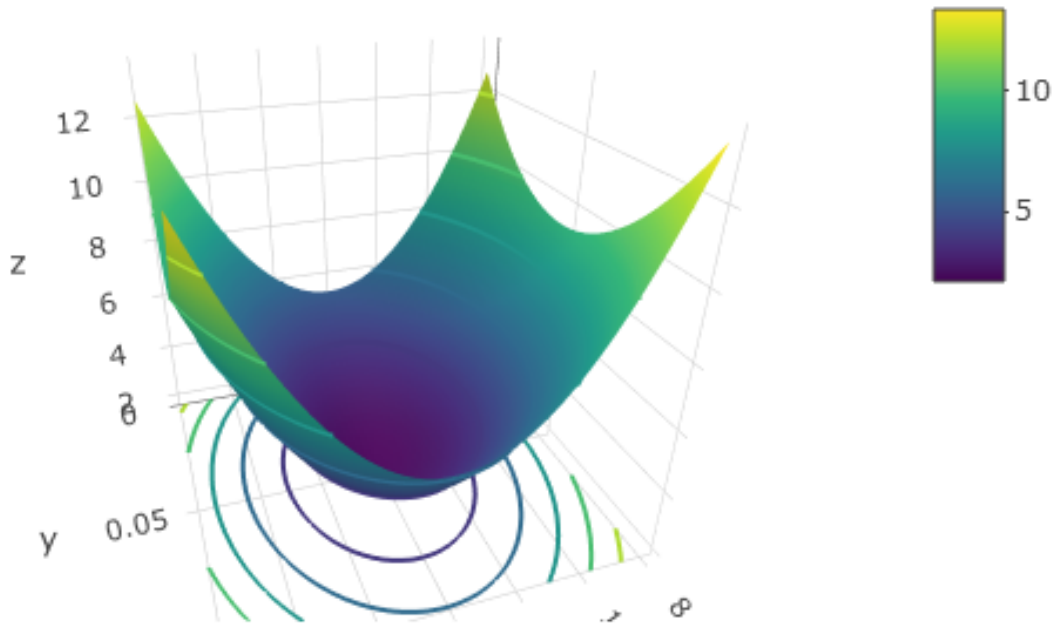
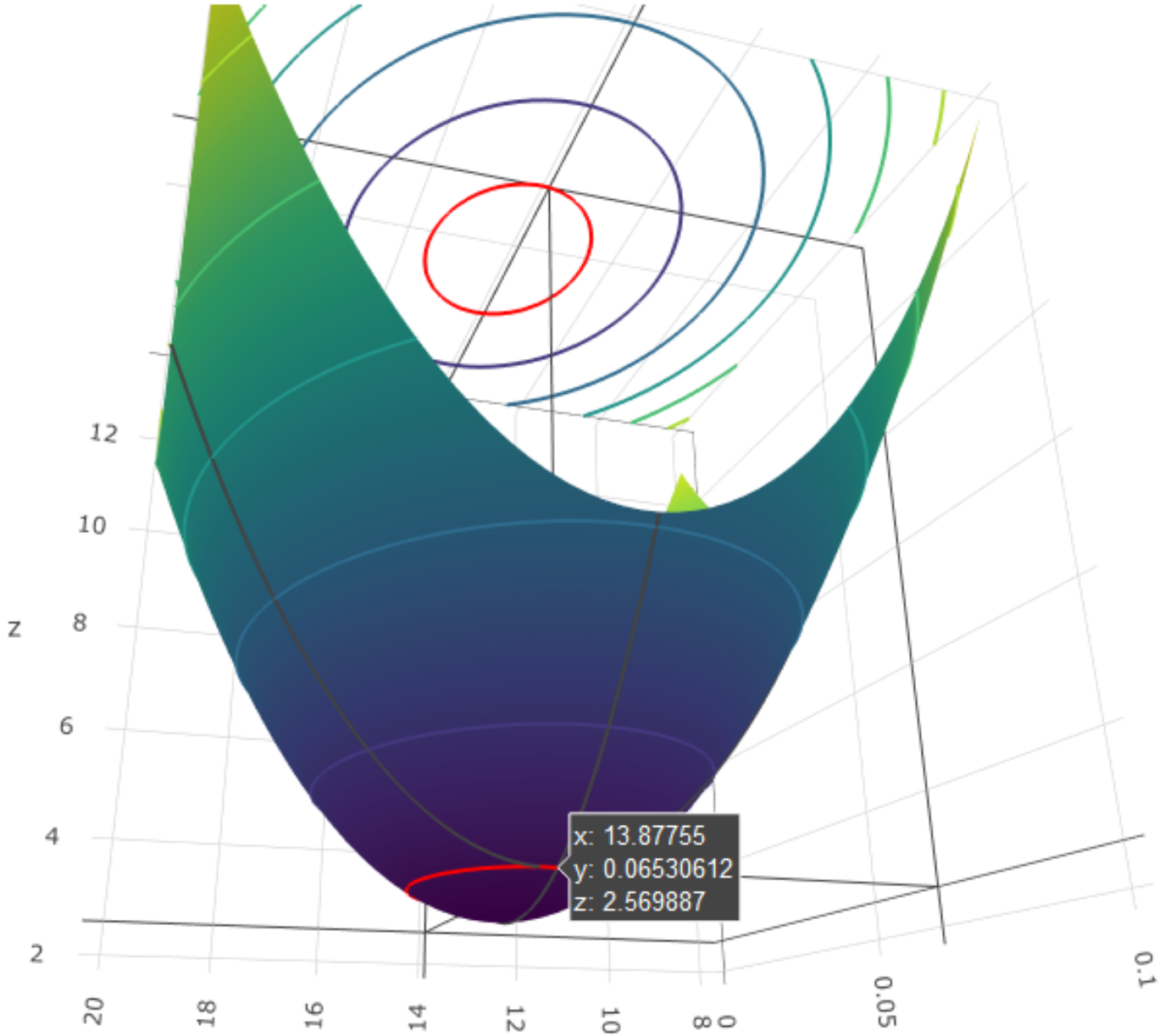
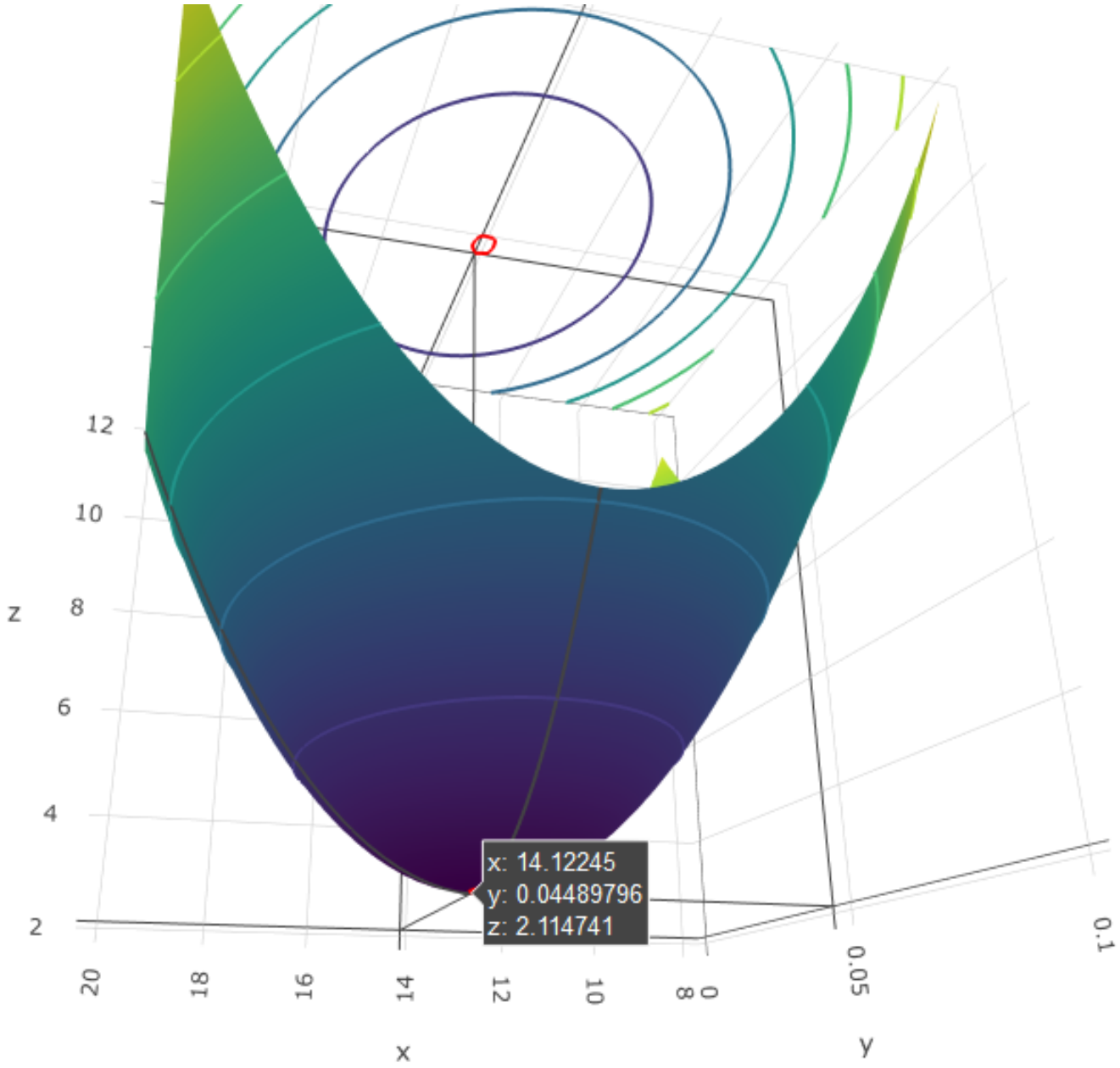



Figura 3: Surface and Contours RSS advertising dataset

Interactivamente, se puede ir buscando el mínimo:



Interactivamente, se puede ir buscando el mínimo:



**¿Cómo evaluar la precisión de los coeficientes estimados?**. Generalmente, se asume que la verdadera relación entre $X$ y $Y$ es de la forma $Y = f(X) + \varepsilon$ para alguna función desconocida $f$, donde $\varepsilon$ es un término de error aleatorio de media cero. Si $f$ se aproxima con una función lineal, entonces $f(X) = \beta_0 + \beta_1 X$ y así

$$\begin{aligned} Y &= f(X) + \varepsilon \\ &= \beta_0 + \beta_1 X + \varepsilon \end{aligned}$$

$\beta_0$ es el intercepto que corresponde al valor esperado de $Y$ cuando $X = 0$ y $\beta_1$ es el incremento promedio en $Y$ asociado con un cambio de una unidad en $X$. El término de error, por su parte, captura todo lo que este modelo simple no puede capturar: La verdadera relación podría no ser lineal, podría haber otras variables que causan variación en $Y$ y podría haber errores de medición. Típicamente, se asume que el término de error $\varepsilon$ es independiente de $X$

El modelo

$$\begin{aligned} Y &= f(X) + \varepsilon \\ &= \beta_0 + \beta_1 X + \varepsilon \end{aligned}$$

define la recta de regresión poblacional, la cual es la mejor aproximación lineal de la verdadera relación entre $X$ y $Y$. Las estimaciones de OLS, $\widehat{\beta}_0$ y $\widehat{\beta}_1$ caracterizan la recta de mínimos cuadrados (OLS line)

$$\begin{aligned} \widehat{y} &= \widehat{f(X)} \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X \end{aligned}$$

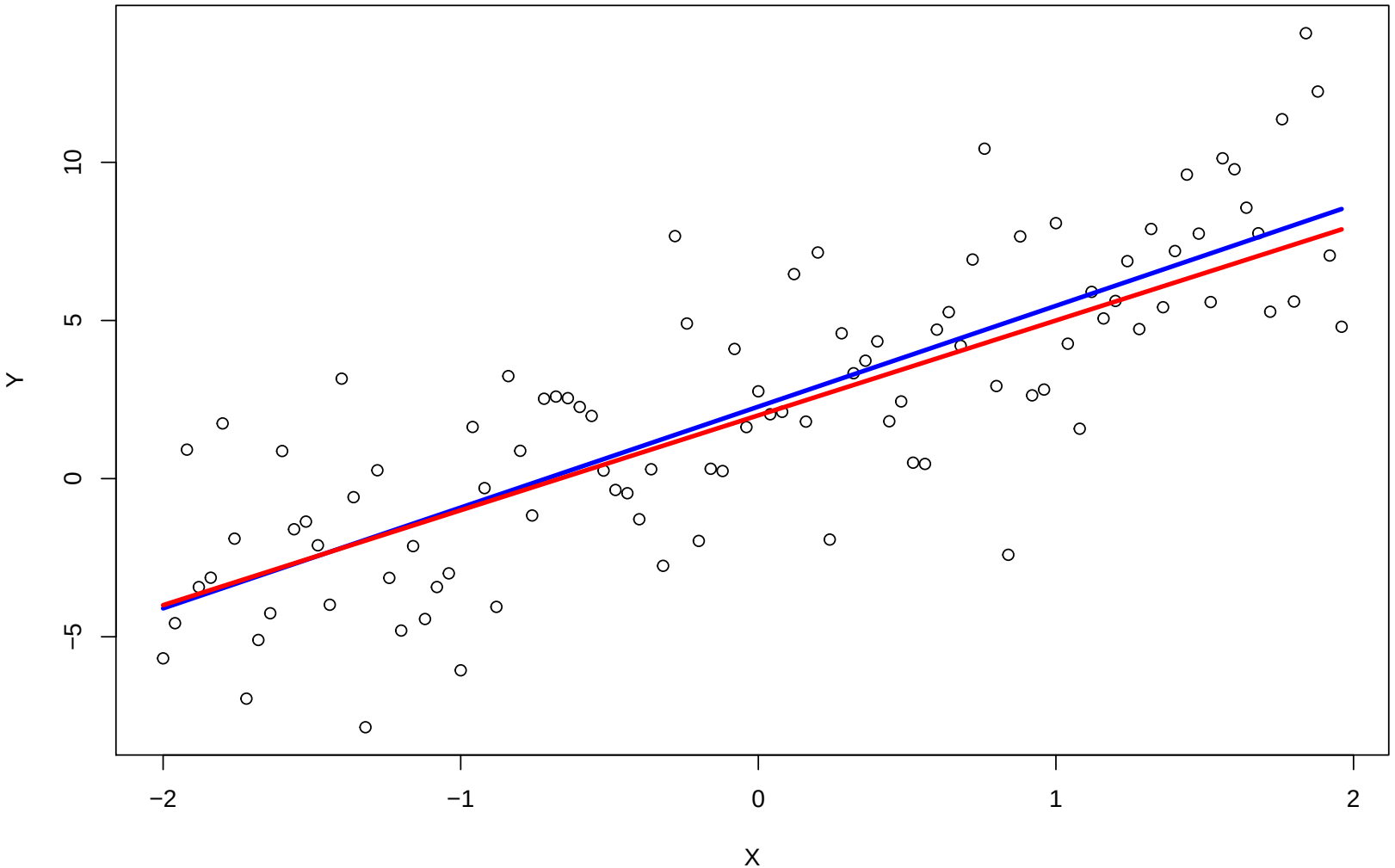
En el siguiente gráfico, se muestran dos rectas obtenidas a partir de datos simulados del modelo:

$$Y = 2 + 3X + \varepsilon \quad \text{, donde } \varepsilon \sim N(0,\ \sigma = 3)$$

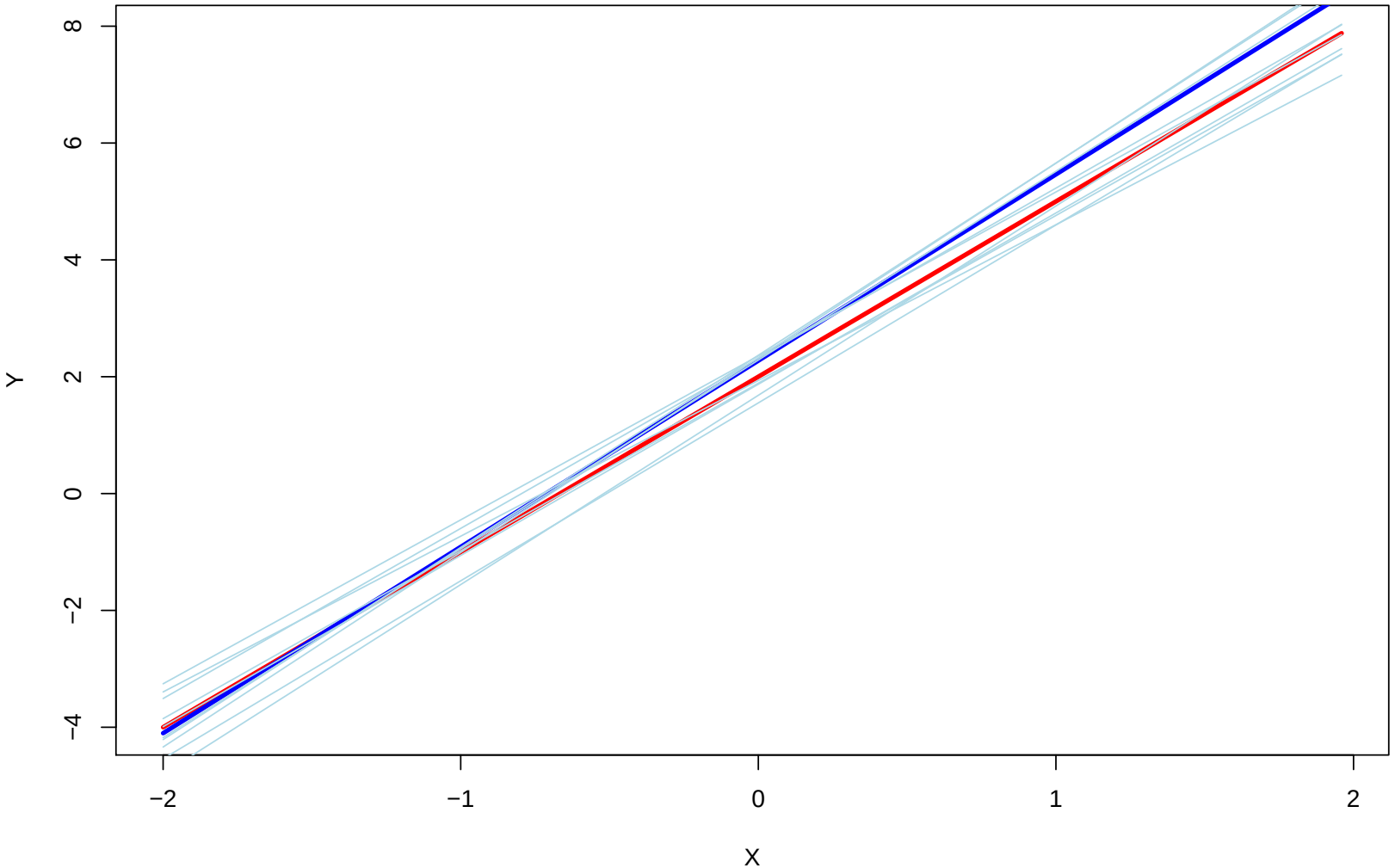
La linea de color rojo es el modelo verdadero $Y = 2 + 3X$ y la línea azul es el modelo de OLS $\widehat{y} = 2.275 + 3.1883 \cdot x$:

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3607 -1.6571 -0.1039  1.9455  6.2846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2750     0.2744   8.291 6.06e-13 ***
## X             3.1883     0.2376  13.418  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.744 on 98 degrees of freedom
## Multiple R-squared:  0.6475, Adjusted R-squared:  0.6439
## F-statistic:   180 on 1 and 98 DF,  p-value: < 2.2e-16
```

# REGRESIÓN LINEAL



# REGRESIÓN LINEAL



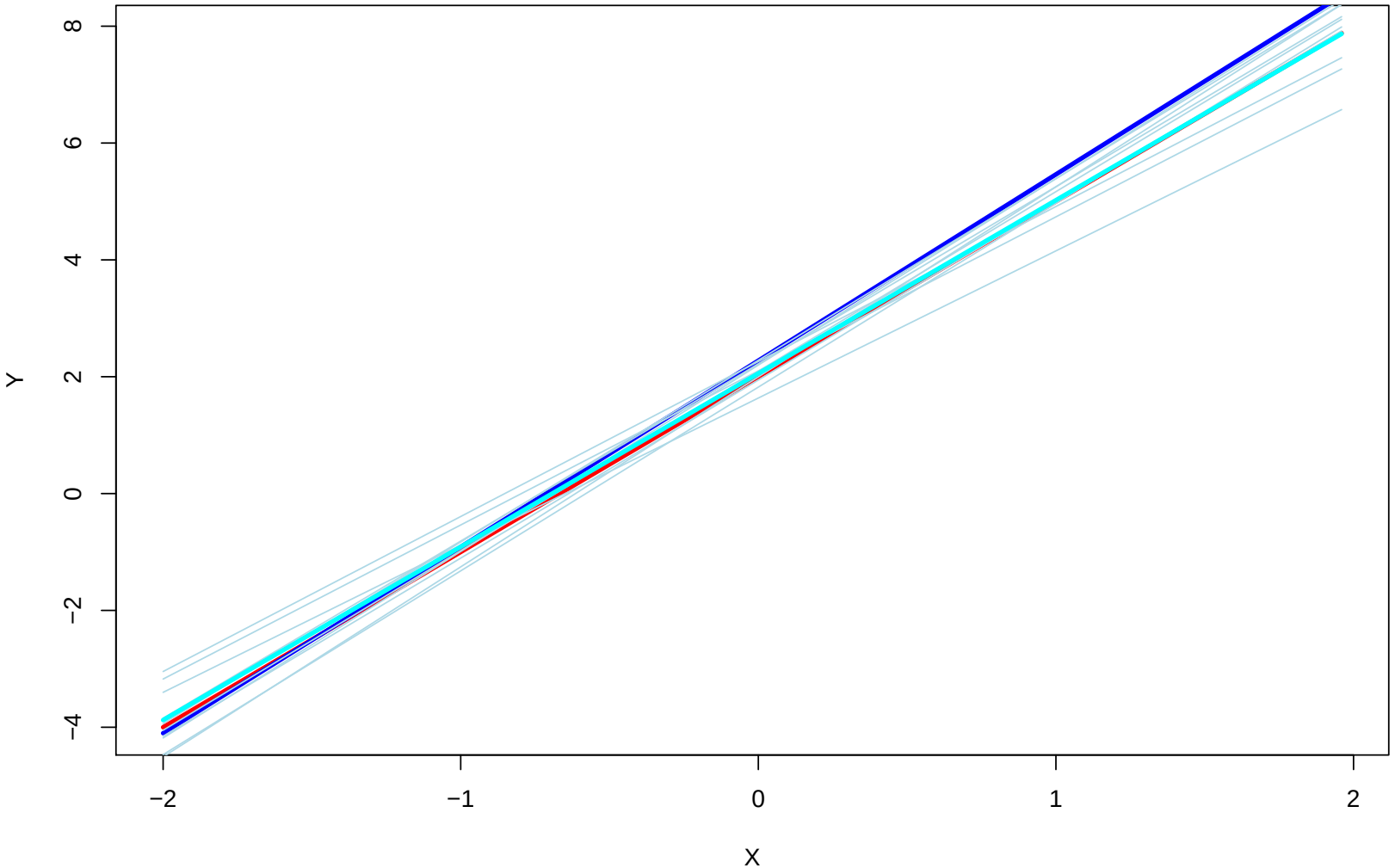
El gráfico anterior ilustra, en azul claro, 10 rectas de mínimos cuadrados, cada una calculada con base en conjuntos aleatorios de observaciones diferentes. Cada recta de OLS es distinta, pero en promedio, están cercanas a la verdadera recta (la línea roja).

A primera vista, la diferencia entre la recta poblacional y la recta OLS puede parecer sutil y confusa. Solo se cuenta con un conjunto de datos, por lo tanto ¿Qué significa que dos rectas distintas describan la relación entre el predictor y la respuesta? El concepto de estas dos líneas, es una extensión natural de la aproximación estándar típica en estadística, de usar información de una muestra para estimar características de una población

(el paradigma estadístico: Nivel muestral: $\widehat{\beta}_0 \approx \beta_0$ Nivel poblacional y Nivel muestral: $\widehat{\beta}_1 \approx \beta_1$ Nivel poblacional). Los coeficientes desconocidos $\beta_0$ y $\beta_1$ definen la recta poblacional, mientras que $\widehat{\beta}_0$ y $\widehat{\beta}_1$ definen la recta de OLS. Un estimador insesgado no sobrestima o subestima sistemáticamente el verdadero parámetro. Los estimadores de OLS de $\beta_0$ y $\beta_1$, $\widehat{\beta}_0$ y $\widehat{\beta}_1$, son insesgados.

¿Qué significa que sean insesgados? Si se estima a $\beta_0$ y $\beta_1$ usando un conjunto de datos particular, lo más seguro es que las estimaciones $\widehat{\beta}_0$ y $\widehat{\beta}_1$ no sean exactamente iguales a $\beta_0$ y $\beta_1$. Pero si se pudiera promediar muchas de estas estimaciones con base en un número grande de muestras, entonces este promedio estaría muy cercano a $\beta_0$ y $\beta_1$. Esto se ilustra en el siguiente gráfico, donde la recta color cyan, es el promedio de las 10 rectas OLS. Se observa que esta recta coincide casi con la verdadera recta (la roja). Esto ilustra la propiedad de insesgamiento de $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

# REGRESIÓN LINEAL



¿Qué tan lejos estará una sola estimación $\widehat{\beta}_0$ del verdadero $\beta_0$? y ¿Qué tan lejos estará una sola estimación $\widehat{\beta}_1$ del verdadero $\beta_1$? En general, estas preguntas se responden calculando los errores estándar de $\widehat{\beta}_0$ y de $\widehat{\beta}_1$, respectivamente. Este error estándar dice la cantidad promedio que esta estimación difiere del verdadero valor poblacional ($\beta_0$ o $\beta_1$)