Modelos de Regresión I

Modelos Probabilísticos y Análisis Estadístico

Carlos Ricardo Bojacá

Departamento de Ciencias Básicas y Modelado
Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano

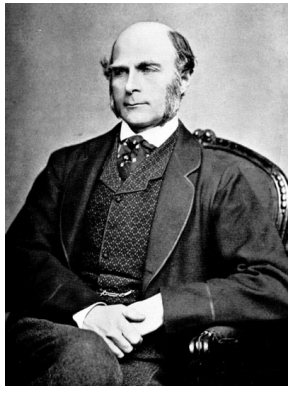
Introducción

Los modelos de regresión son el "caballo de batalla" de la ciencia de los datos (data science). Son los modelos mejor descritos, prácticos y entendidos teóricamente en estadística

La clave de los modelos de regresión es que producen ajustes con un alto grado de interpretación. Esto no sucede con algoritmos de aprendizaje de máquina, en los que se sacrifica la interpretabilidad por el rendimiento en la predicción

Los beneficios de los modelos de regresión (y sus generalizaciones), tales como simplicidad, parsimonia e interpretabilidad, los convierten en la primera opción para analizar cualquier problema práctico

Motivación



El modelo sigue siendo relevante

European Journal of Human Genetics (2009) 17, 1070–1075

www.nature.com/ejhg

ARTICLE

Predicting human height by Victorian and genomic methods

Yurii S Aulchenko*[,1,2,7], Maksim V Struchalin[1,3,7], Nadezhda M Belonogova[2,4], Tatiana I Axenovich[2], Michael N Weedon[5], Albert Hofman[1], Andre G Uitterlinden[6], Manfred Kayser[3], Ben A Oostra[1], Cornelia M van Duijn[1], A Cecile JW Janssens[1] and Pavel M Borodin[2,4]

[1]*Department of Epidemiology and Biostatistics and Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands;* [2]*Laboratory of Recombination and Segregation Analysis, Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia;* [3]*Department of Forensic Molecular Biology, Erasmus MC, Rotterdam, The Netherlands;* [4]*Department of Cytology and Genetics, Novosibirsk State University, Novosibirsk, Russia;* [5]*Department of Genetics of Complex Traits and Diabetes Genetics, Peninsula College of Medicine and Dentistry, Exeter, UK;* [6]*Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands*

In the Victorian era, Sir Francis Galton showed that 'when dealing with the transmission of stature from parents to children, the average height of the two parents, … is all we need care to know about them' (1886). One hundred and twenty-two years after Galton's work was published, 54 loci showing strong statistical evidence for association to human height were described, providing us with potential genomic means of human height prediction. In a population-based study of 5748 people, we find that a 54-loci genomic profile explained 4–6% of the sex- and age-adjusted height variance, and had limited ability to discriminate tall/short people, as characterized by the area under the receiver-operating characteristic curve (AUC). In a family-based study of 550 people, with both parents having height measurements, we find that the Galtonian mid-parental prediction method explained 40% of the sex- and age-adjusted height variance, and showed high discriminative accuracy. We have also explored how much variance a genomic profile should explain to reach certain AUC values. For highly heritable traits such as height, we conclude that in applications in which parental phenotypic information is available (eg, medicine), the Victorian Galton's method will long stay unsurpassed, in terms of both discriminative accuracy and costs. For less heritable traits, and in situations in which parental information is not available (eg, forensics), genomic methods may provide an alternative, given that the variants determining an essential proportion of the trait's variation can be identified.

European Journal of Human Genetics (2009) **17**, 1070–1075; doi:10.1038/ejhg.2009.5; published online 18 February 2009

Keywords: height; heritability; prediction; genomic profiling; discriminative accuracy; area under the receiver-operating characteristic curve (AUC)

Los datos de Galton

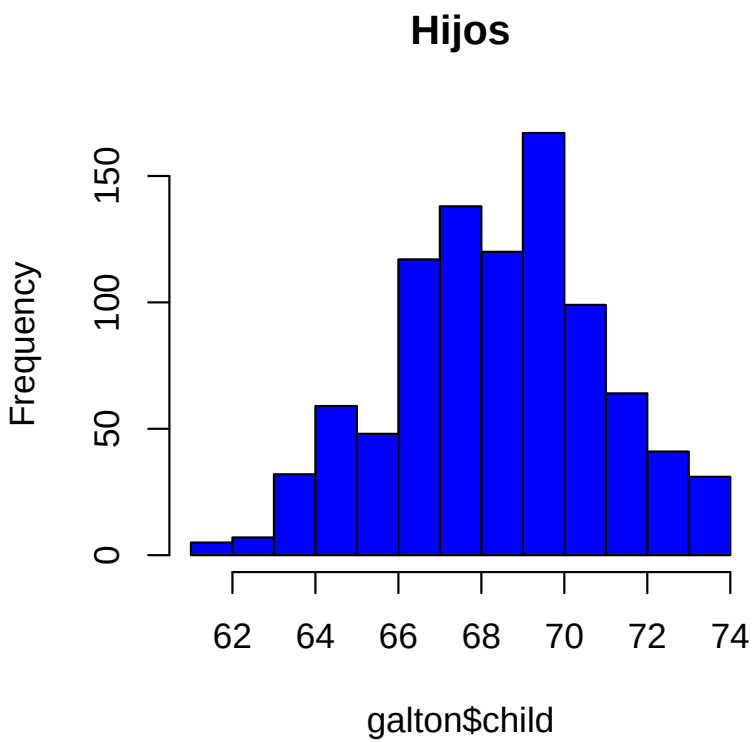
- Francis Galton fue un estadístico que inventó el término y los conceptos de regresión y correlación, fundó la revista Biometrika, fue el primo de Charles Darwin
- Instalar el paquete *UsingR* si no está instalado en su computador
- Examine las distribuciones marginales de las alturas de padres e hijos
  - La distribución de los padres proviene de parejas heterosexuales
  - Se hizo corrección por género multiplicando la altura de las mujeres por 1.08

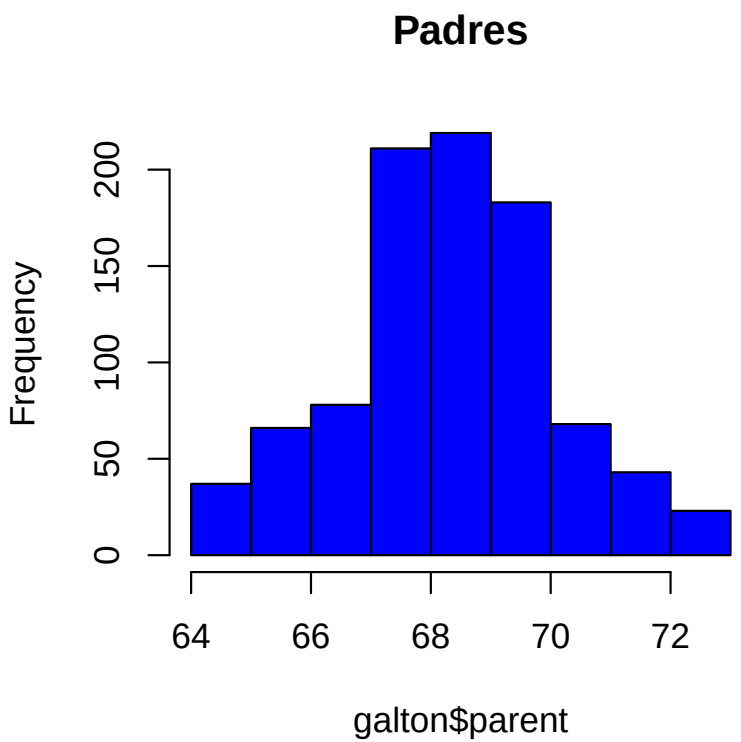
Los datos de Galton

```
library(UsingR); data(galton)
head(galton)

##   child parent
## 1  61.7   70.5
## 2  61.7   68.5
## 3  61.7   65.5
## 4  61.7   64.5
## 5  61.7   64.0
## 6  62.2   67.5
```

Los datos de Galton





Encontrando el punto medio vía cuadrados mínimos

- Considere solamente las alturas de los hijos
  - Cómo puede uno describir el punto "medio"?
  - Definición: sea Y_i la altura del i-ésimo hijo con $i = 1, \ldots, n = 928$, entonces el punto medio se define como el valor de μ que minimiza:

$$\sum_{n}^{i=1} (Y_i - \mu)^2$$

- Este es el centro de masa físico del histograma
- Mejor solución posible: $\mu = \bar{Y}$

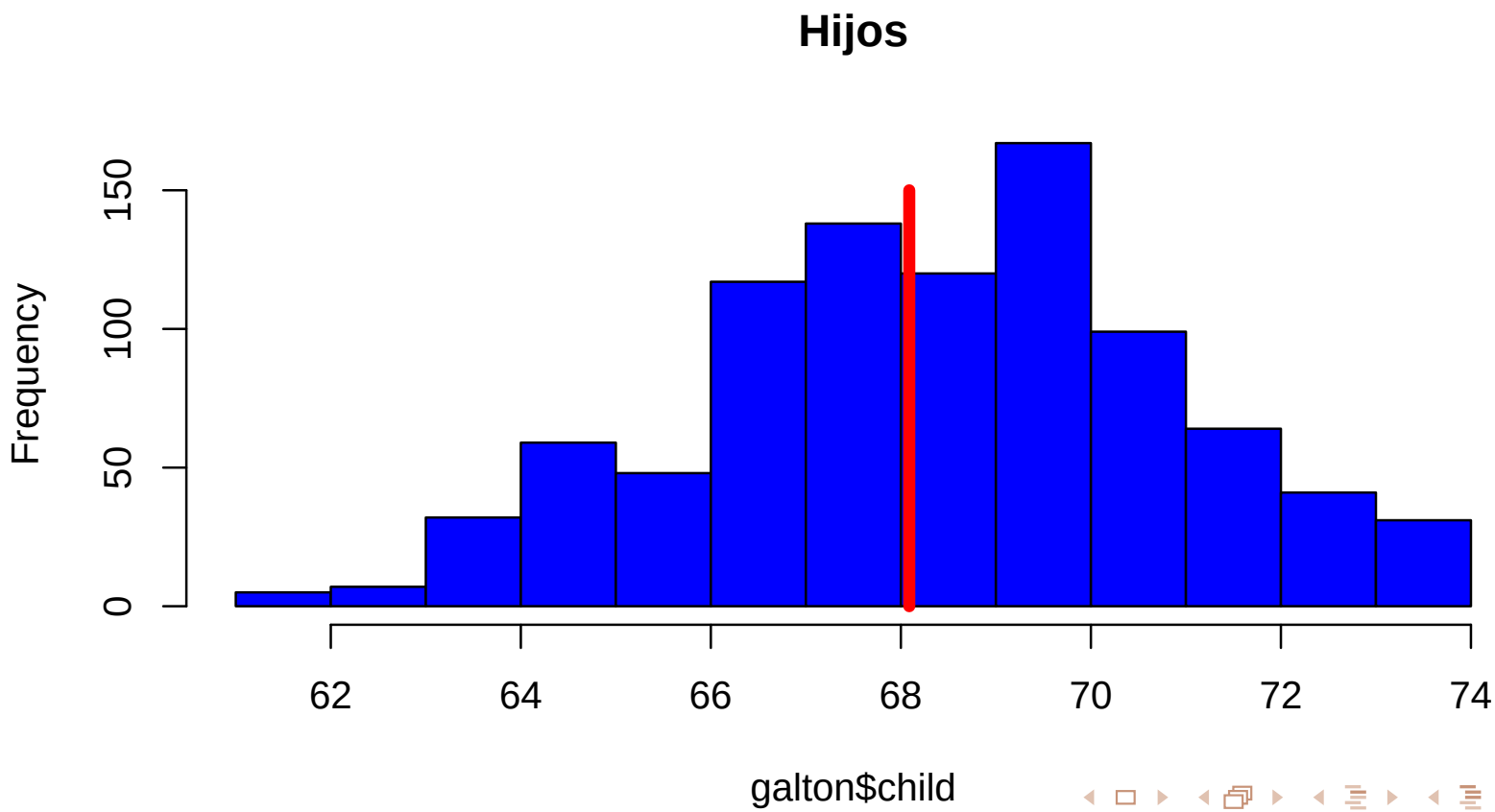
Encontrando el punto medio vía cuadrados mínimos

Utilice la función *manipulate* de RStudio para hallar el valor de μ que minimiza la suma de las desviaciones al cuadrado

```
library(manipulate)
myHist <- function(mu){
  mse <- mean((galton$child - mu)^2)
  ggplot(galton, aes(x = child)) + geom_histogram(fill =
    "salmon", colour = "black", binwidth = 1) +
    geom_vline(xintercept = mu, size = 3) +
    ggtitle(paste0("mu = ", mu, ", MSE = ", round(mse, 2)))
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

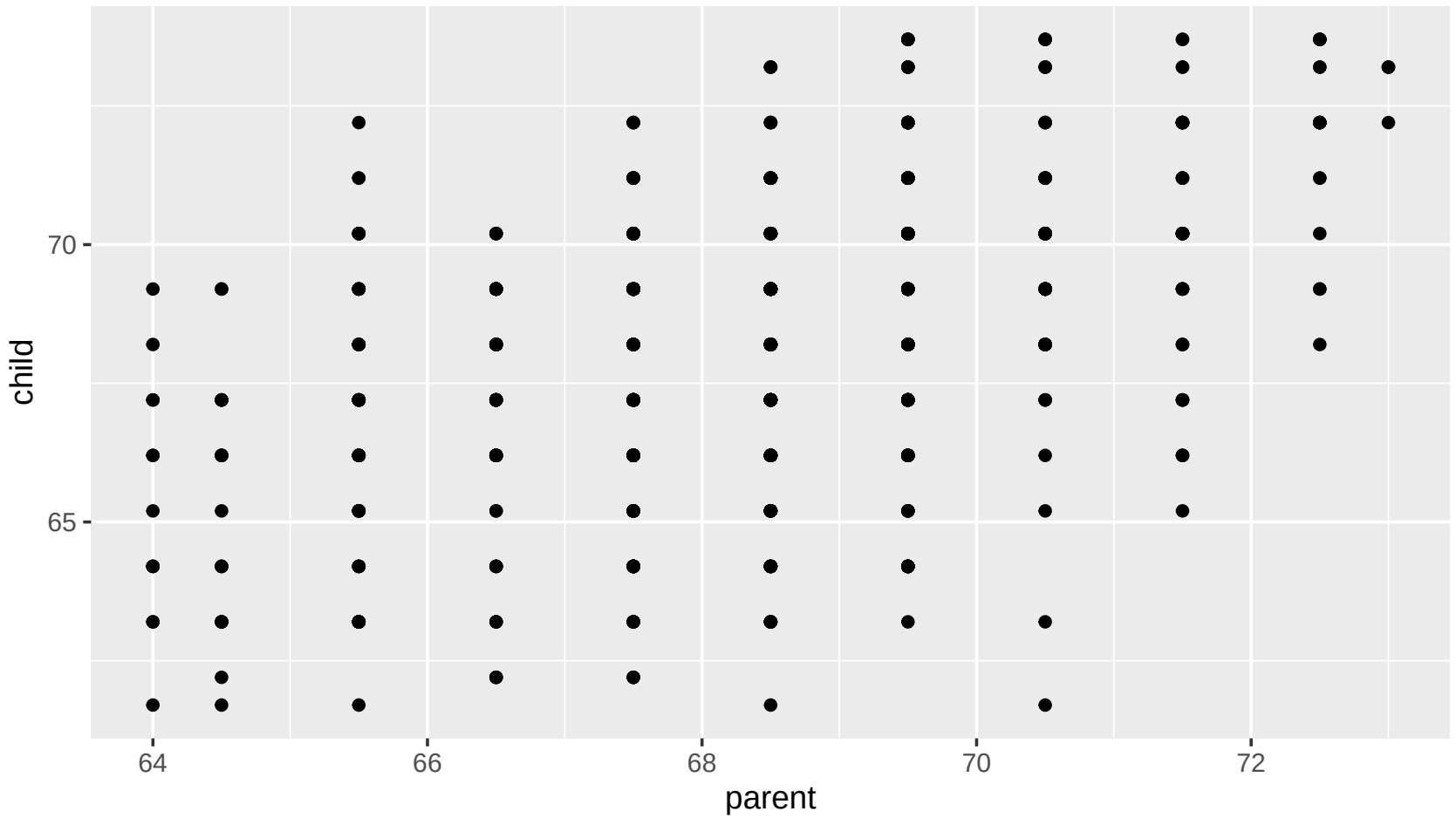
Encontrando el punto medio vía cuadrados mínimos

```
hist(galton$child, col="blue", main = "Hijos")
meanChild <- mean(galton$child)
lines(rep(meanChild, 100), seq(0, 150, length = 100), col = "red",
      lwd = 5)
```

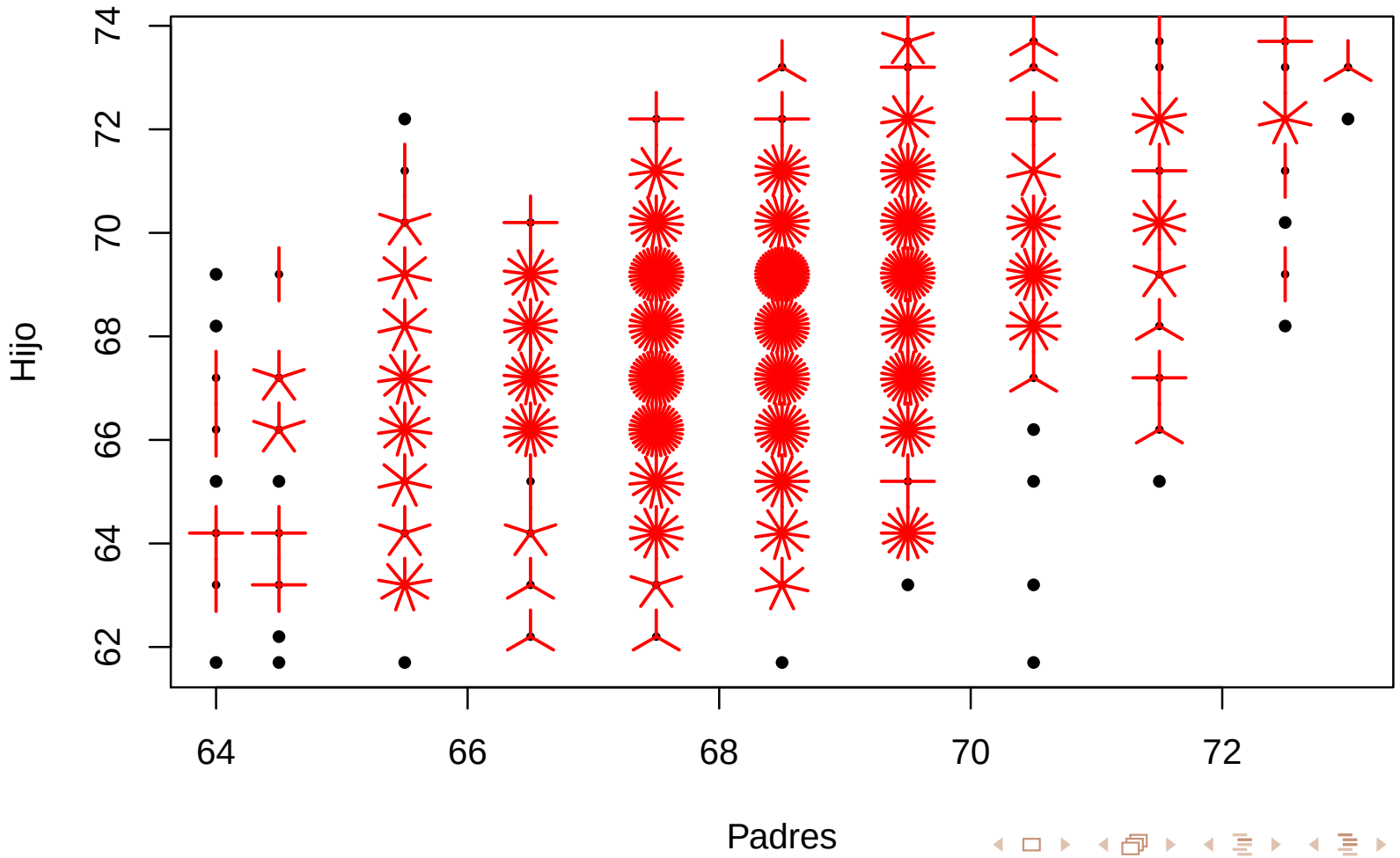


Comparación entre las alturas de padres e hijos

```
library(ggplot2)
ggplot(galton, aes(x = parent, y = child)) + geom_point()
```



Comparación entre las alturas de padres e hijos



Regresión a través del origen

- Suponga que X_i son las alturas de los padres
- Considere seleccionar una pendiente β que minimice

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

- Esto es utilizar el origen como punto de giro para seleccionar la línea que minimiza la suma de las distancias verticales elevadas al cuadrado entre los puntos y la línea
- Reste el promedio a ambas variables de manera que el origen represente el promedio de las alturas de los padres y los hijos

Regresión a través del origen

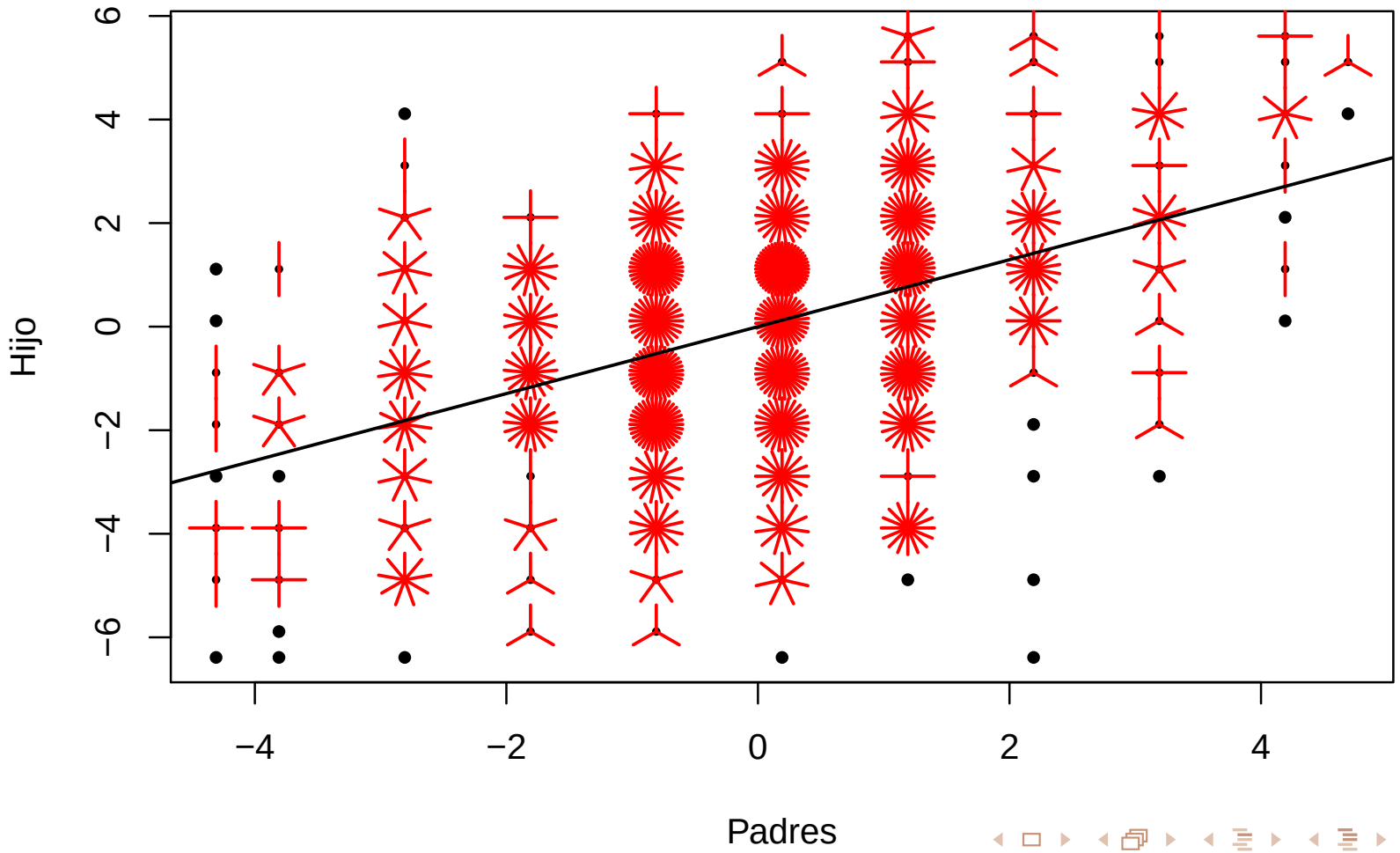
```
myPlot <- function(beta){
  y <- galton$child - mean(galton$child)
  x <- galton$parent - mean(galton$parent)
  sunflowerplot(x, y)
  abline(0, beta, lwd = 3)
  points(0, 0, cex = 2, pch = 19)
  mse <- mean( (y - beta * x)^2 )
  title(paste("beta = ", beta, "mse = ", round(mse, 3)))
}
manipulate(myPlot(beta), beta = slider(0.3, 1.2, step = 0.005))
```

Solución

```
lmgalton <- lm(I(child - mean(child)) ~ I(parent - mean(parent)) - 1,
   data = galton)
lmgalton

##
## Call:
## lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
##     1, data = galton)
##
## Coefficients:
## I(parent - mean(parent))
##                   0.6463
```

Solución



Ajuste de la mejor línea

Sea Y_i la i-ésima altura de un hijo y X_i la i-ésima altura de los padres

Considere encontrar la mejor línea

$$\text{Altura Hijo} = \beta_0 + \text{Altura Padres} \times \beta_1$$

Utilizando el método de mínimos cuadrados

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

Resultados

El ajuste por mínimos cuadrados a la línea $Y = \beta_0 + \beta_1 X$ a través de las parejas de datos (X_i, Y_i) siendo Y_i la respuesta obtiene la línea $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ donde:

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}; \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $\hat{\beta}_1$ tiene las unidades de Y/X, $\hat{\beta}_0$ tiene las unidades de Y
- La línea pasa a través del punto $(\bar{X}, \bar{Y})$
- La pendiente de la línea de regresión con X como la respuesta y Y como la variable predictora es $Cor(X, Y)Sd(X)/Sd(Y)$

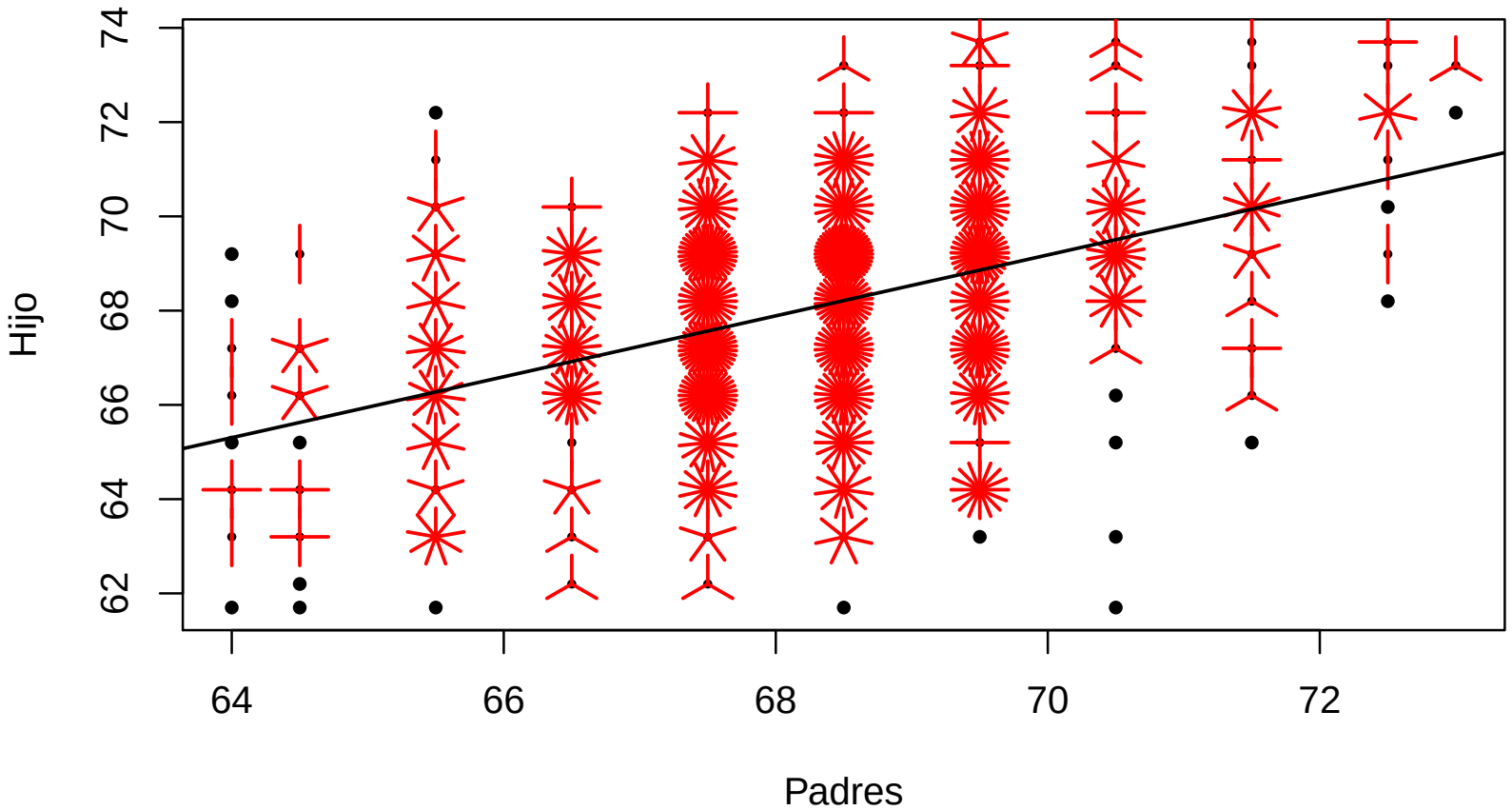
Resultados

- La pendiente es la misma si se centran o no los datos $(X_i - \bar{X}, Y_i - \bar{Y})$
- La solución por mínimos cuadrados para la regresión a través del origen, asumiendo $\beta_0 = 0$, es igual a:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

- Si se normalizan los datos, $\left\{ \frac{X_i - \bar{X}}{Sd(X)}, \frac{Y_i - \bar{Y}}{Sd(Y)} \right\}$, la pendiente será $Cor(Y, X)$

Resultados



Resultados

```
lmgalton2 <- lm(child ~ parent, data = galton)
sunflowerplot(galton$parent, galton$child, xlab = 'Padres', ylab = 'Hijo')
abline(reg = lmgalton2, lwd = 1.5)
```

Modelos estadísticos de regresión lineal

La estimación de parámetros a través del método de mínimos cuadrados es una operación matemática

Cómo extender las estimaciones al conjunto de toda una población?

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Asumiendo:

- ϵ_i son iid $N(0, \sigma^2)$
- $E[Y_i|X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$
- $Var(Y_i|X_i = x_i) = \sigma^2$

Interpretación de los coeficientes de regresión - β_0

β_0 representa el valor esperado de la respuesta cuando el valor del predictor es 0

$$E[Y|X = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

No siempre esta situación es de interés, por ejemplo cuando $X = 0$ es imposible o se encuentra muy lejos del rango de los datos (X es presión sanguínea, altura, etc.)
Considere:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + a\beta_1 + \beta_1(X_i - a) + \epsilon_i = \tilde{\beta}_0 + \beta_1(X_i - a) + \epsilon_i$$

Usualmente el valor de a es $\bar{X}$

Interpretación de los coeficientes de regresión - β_1

β_1 es el cambio esperado en la respuesta por cambio de una unidad en el predictor

$$E[Y|X = x+1] - E[Y|X = x] = \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) = \beta_1$$

Considere el impacto de cambiar las unidades

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + \frac{\beta_1}{a}(X_i a) + \epsilon_i = \beta_0 + \tilde{\beta}_1(X_i a) + \epsilon_i$$

La multiplicación de X por un factor a resulta en dividir el coeficiente β_1 por el factor a

Interpretación de los coeficientes de regresión - β_1

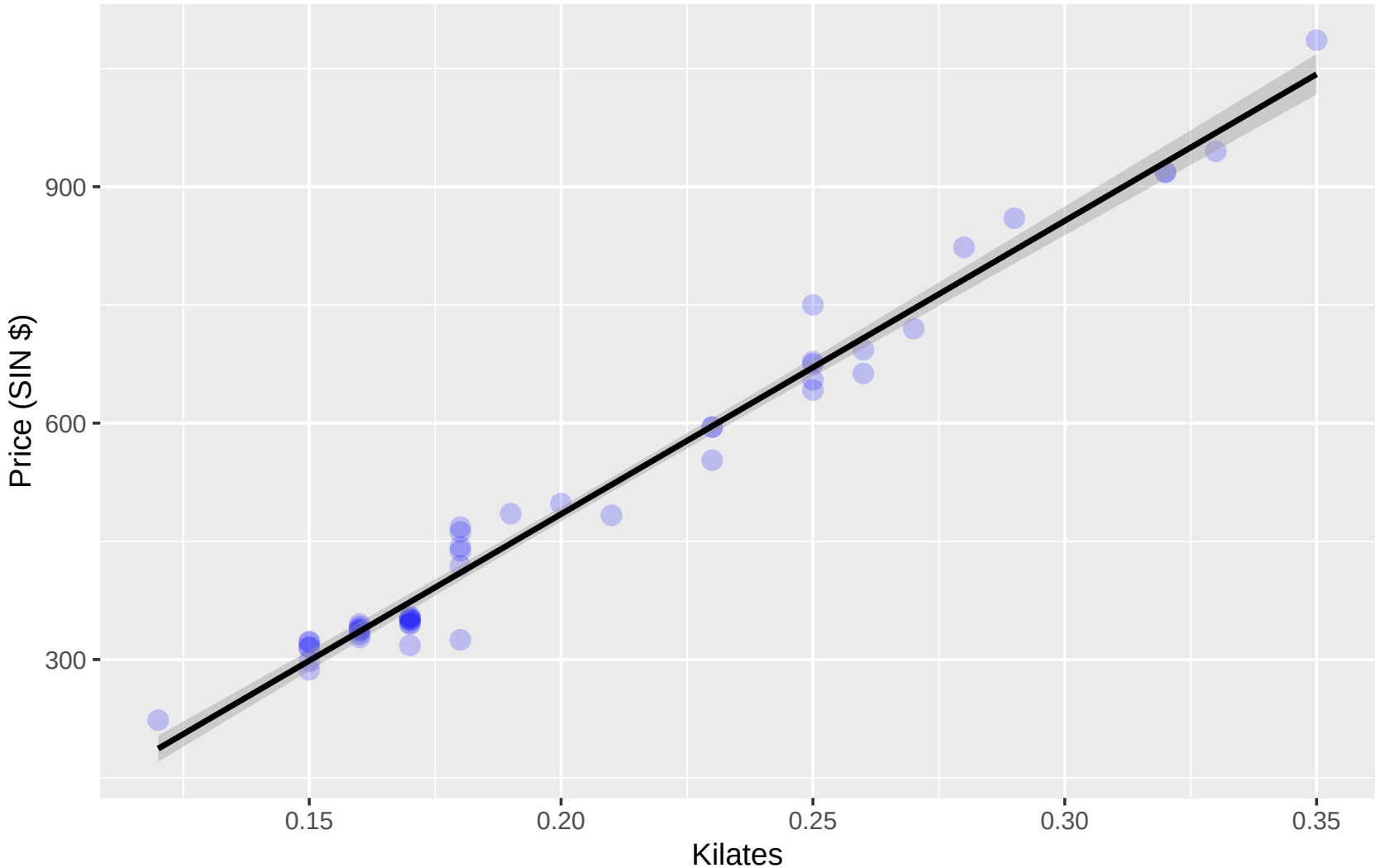
Suponga que X es la altura en metros (m) y Y el peso en kilogramos (kg). Entonces β_1 estará expresado en kg/m.

Qué pasa si se convierten las alturas a centímetros (cm)?

$$Xm \times \frac{100cm}{m} = (100X)cm$$

$$\beta_1 \frac{kg}{m} \times \frac{1m}{100cm} = \frac{\beta_1}{100} \frac{kg}{cm}$$

Interpretación de coeficientes



Interpretación de coeficientes

- Ajuste un modelo de regresión lineal donde $Y =$ precio y $X =$ es la masa
- Obtenga un valor interpretable para β_0 utilizando $\bar{X}$
- Obtenga un valor interpretable para β_1 de manera que su valor represente el cambio en el precio del diamante debido al cambio en 0.1 kilates
- Prediga el precio de tres diamantes que tienen pesos de 0.16, 0.27 y 0.34 kilates

Residuales

Los residuales representan la variación no explicada por el modelo. Los residuales e_i son valores estimados de los errores ϵ_i

Un residual se define como la diferencia entre el valor observado y el valor predicho por el modelo:

$$e_i = Y_i - \hat{Y}_i$$

Los residuales representan las distancias verticales entre los datos observados y su correspondiente predicción sobre la línea de regresión

Propiedades de los residuales

- $E[e_i] = 0$
- Si el modelo incluye intercepto, $\sum_{i=1}^{n} e_i = 0$
- Si se incluye una variable regresora, X_i, se incluye en el modelo $\sum_{i=1}^{n} e_i X_i = 0$
- Los residuales son útiles para determinar el grado de ajuste del modelo
- Los residuales positivos se van a ubicar por encima de la línea de regresión, los residuales negativos se ubicarán por debajo
- La variación total de la respuesta se descompone en variación residual (variación luego de remover los predictores) y variación sistemática (variación explicada por el modelo de regresión)

Residuales

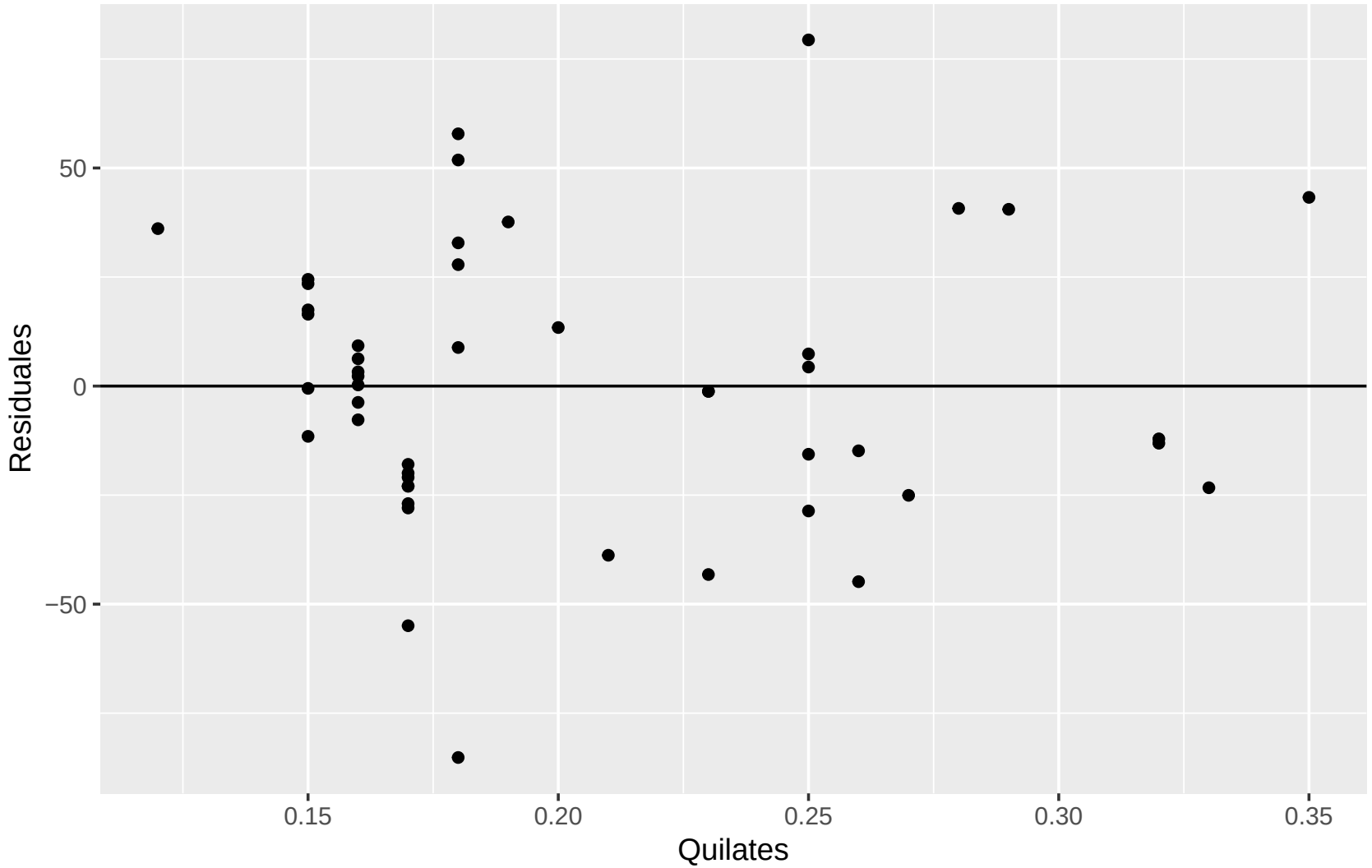
```
data(diamond)
(fit <- lm(price ~ carat, data = diamond))
```

```
##
## Call:
## lm(formula = price ~ carat, data = diamond)
##
## Coefficients:
## (Intercept)        carat
##      -259.6       3721.0
```

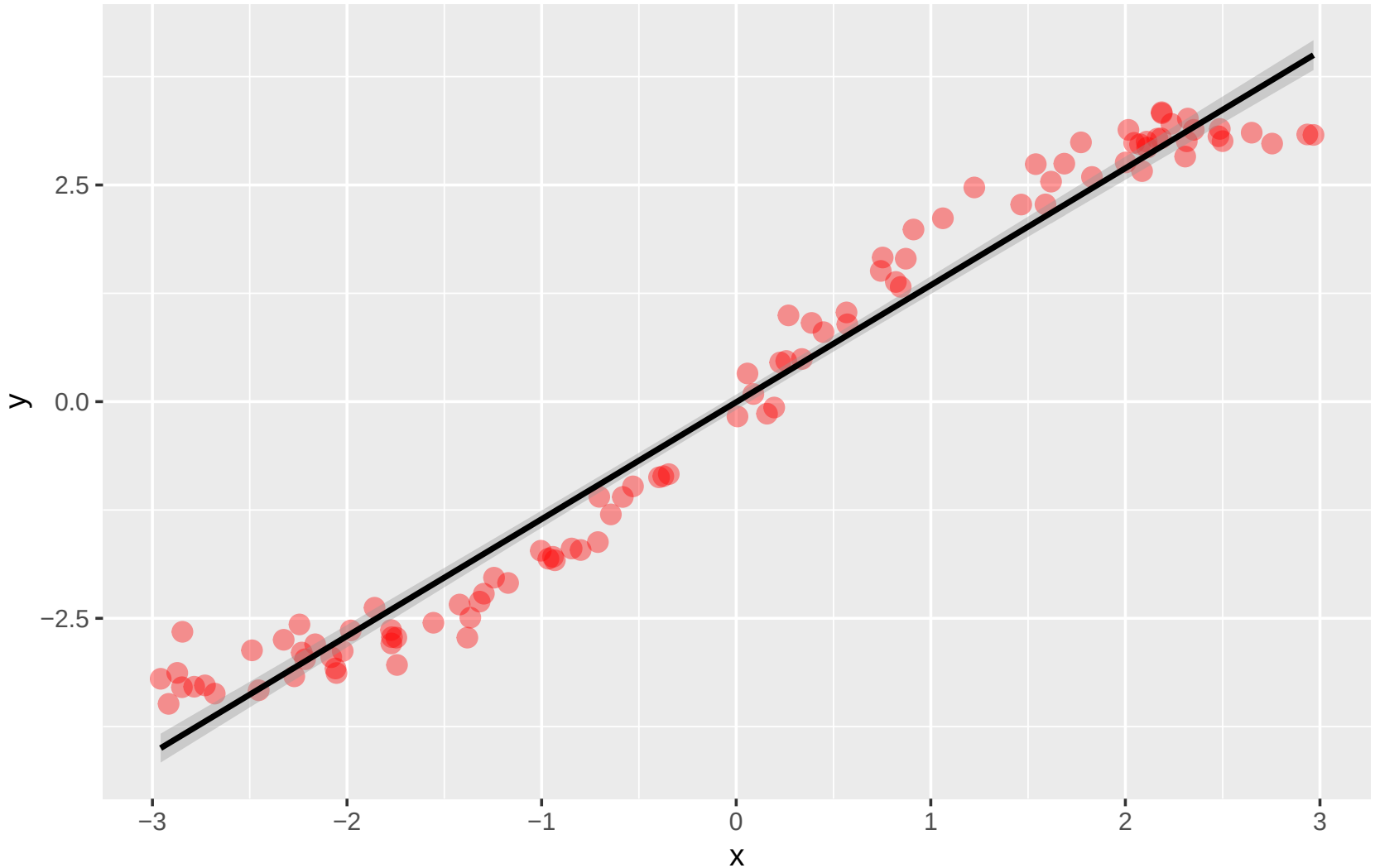
```
e <- resid(fit)
yhat <- predict(fit)
max(abs(e - (diamond$price - yhat)))
```

```
## [1] 9.485746e-13
```

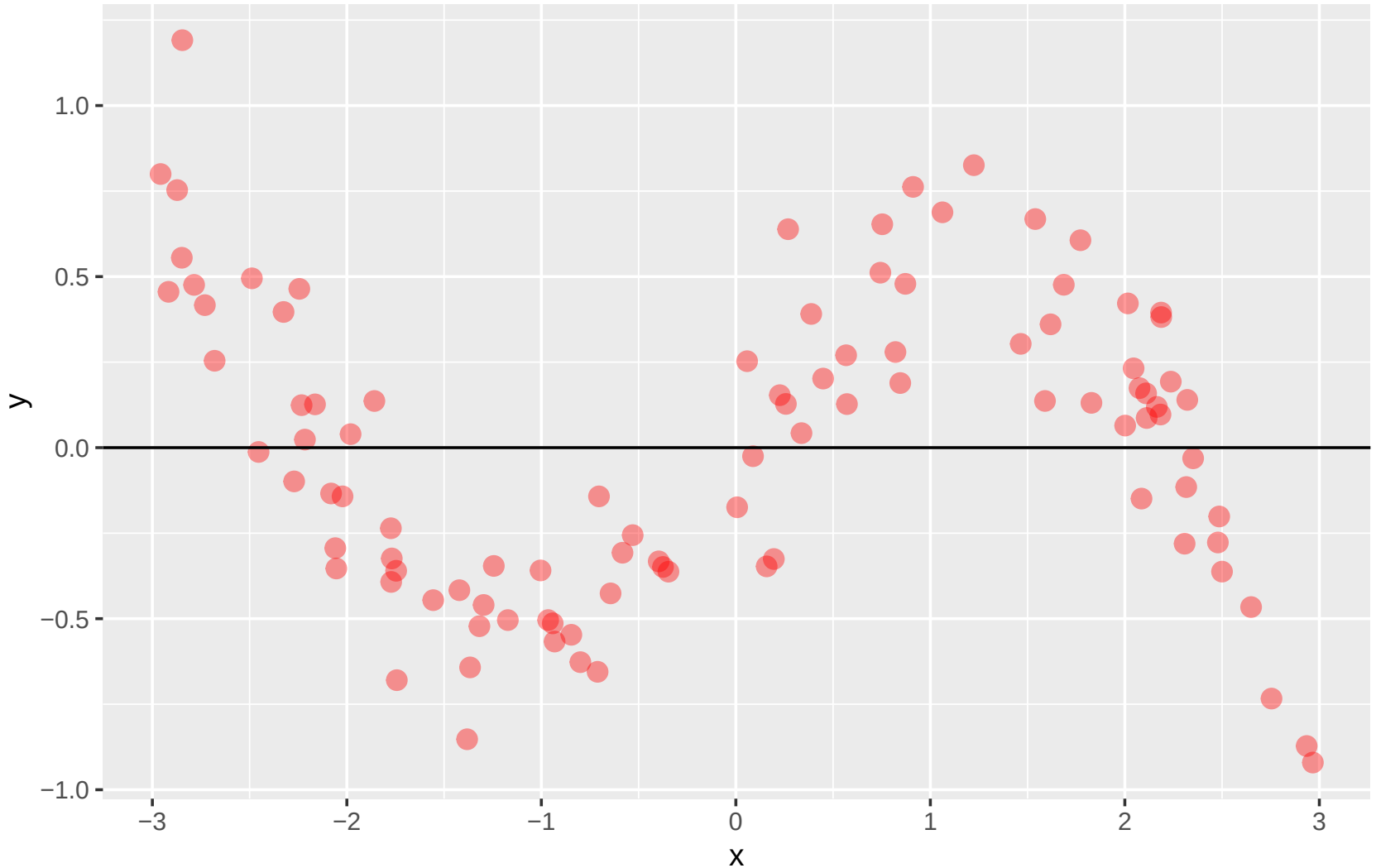
Residuales



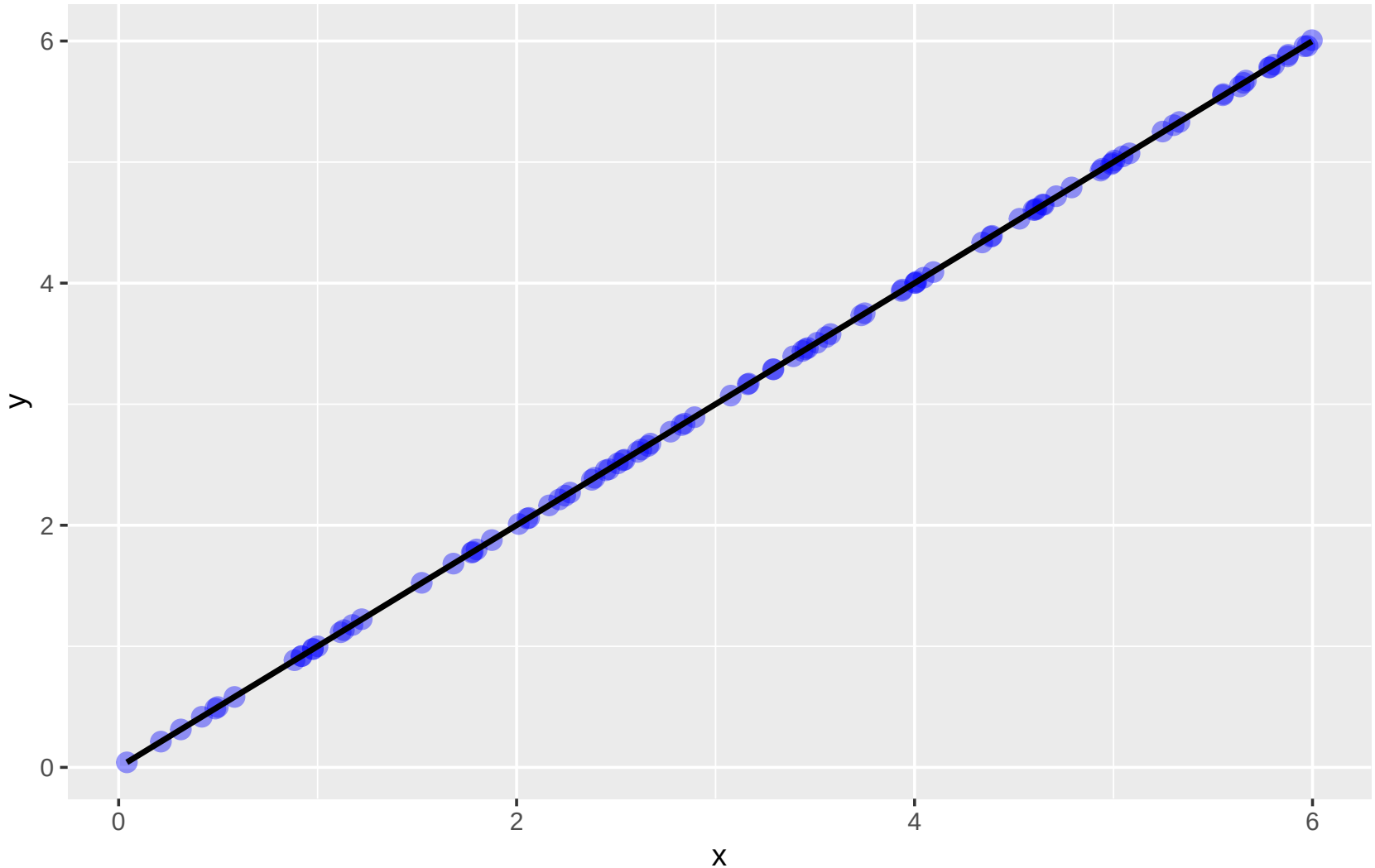
Residuales



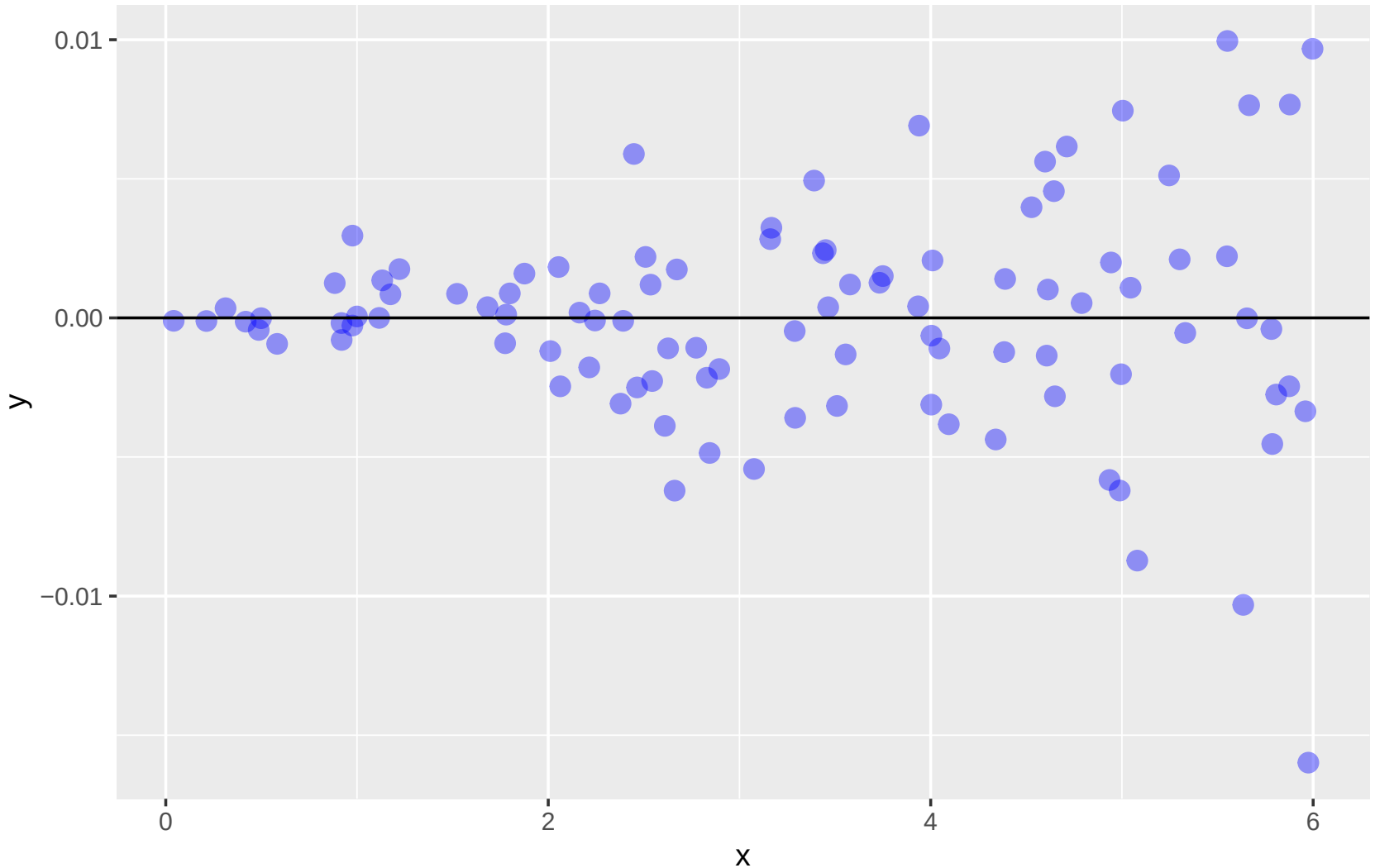
Residuales



Residuales



Residuales



Estimación de la variación residual

Modelo: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ donde $\epsilon_i \sim N(0, \sigma^2)$

El estimador de σ^2 es $\frac{1}{n}\sum_{i=1}^{n} e_i^2$, el promedio de los residuales al cuadrado

Aunque es común utilizar:

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

Lo que indica la pérdida de grados de libertad debida a la estimación de los parámetros del modelo

Estimación de la variación residual

```
x <- diamond$carat
y <- diamond$price
n <- length(y)
fit <- lm(y ~ x)
summary(fit)$sigma

## [1] 31.84052

sqrt(sum(resid(fit)^2) / (n - 2))

## [1] 31.84052
```

Resumen variación residual

$$\text{Variabilidad total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

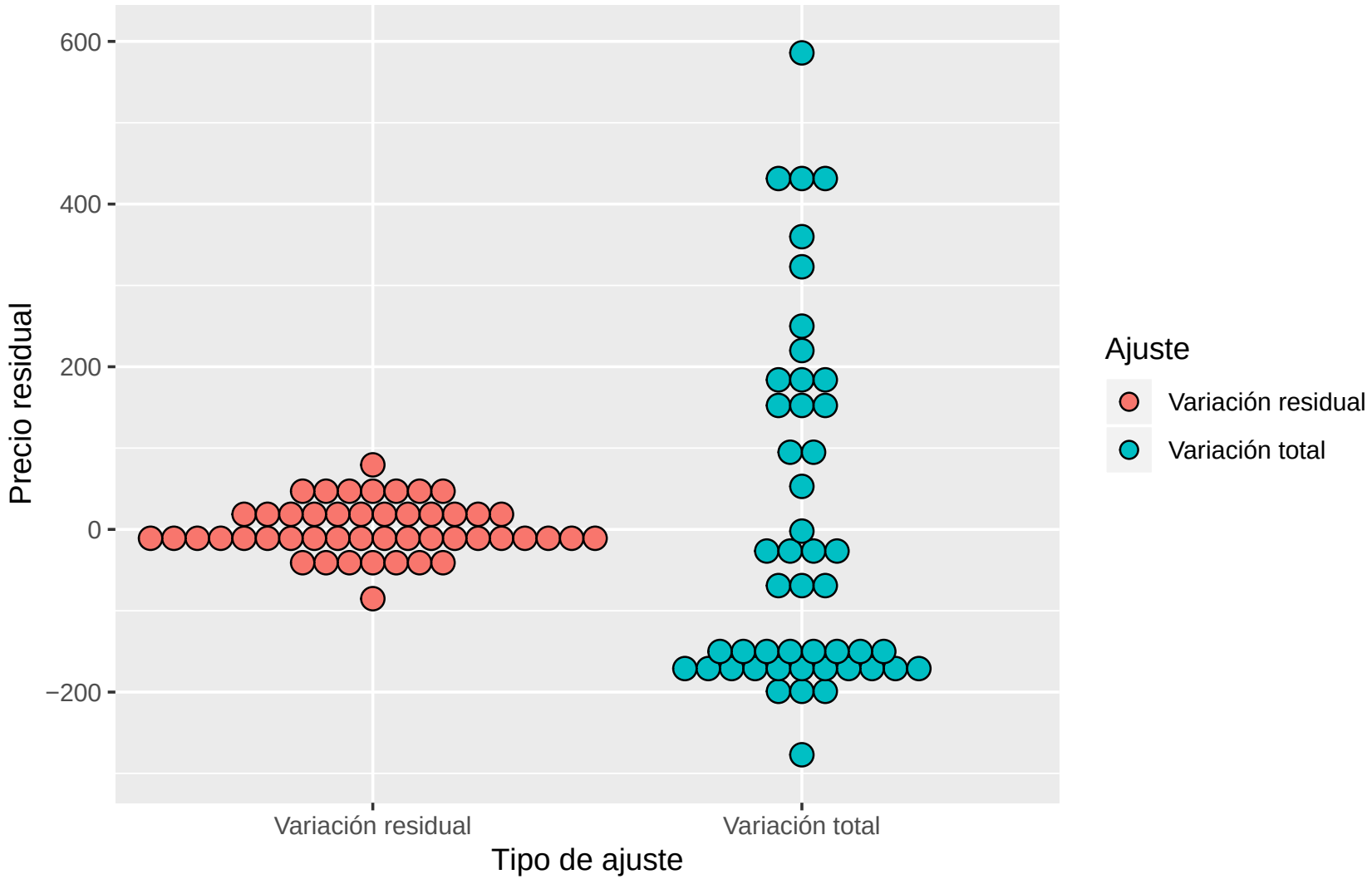
$$\text{Variabilidad regresión} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$\text{Variabilidad residual} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

La suma del error y la variabilidad de la regresión es igual a la variabilidad total:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$$

Residuales


600
400
200
0
−200
Precio residual
Variación residual
Variación total
Tipo de ajuste
Ajuste
Variación residual
Variación total

Coeficiente de determinación - R^2

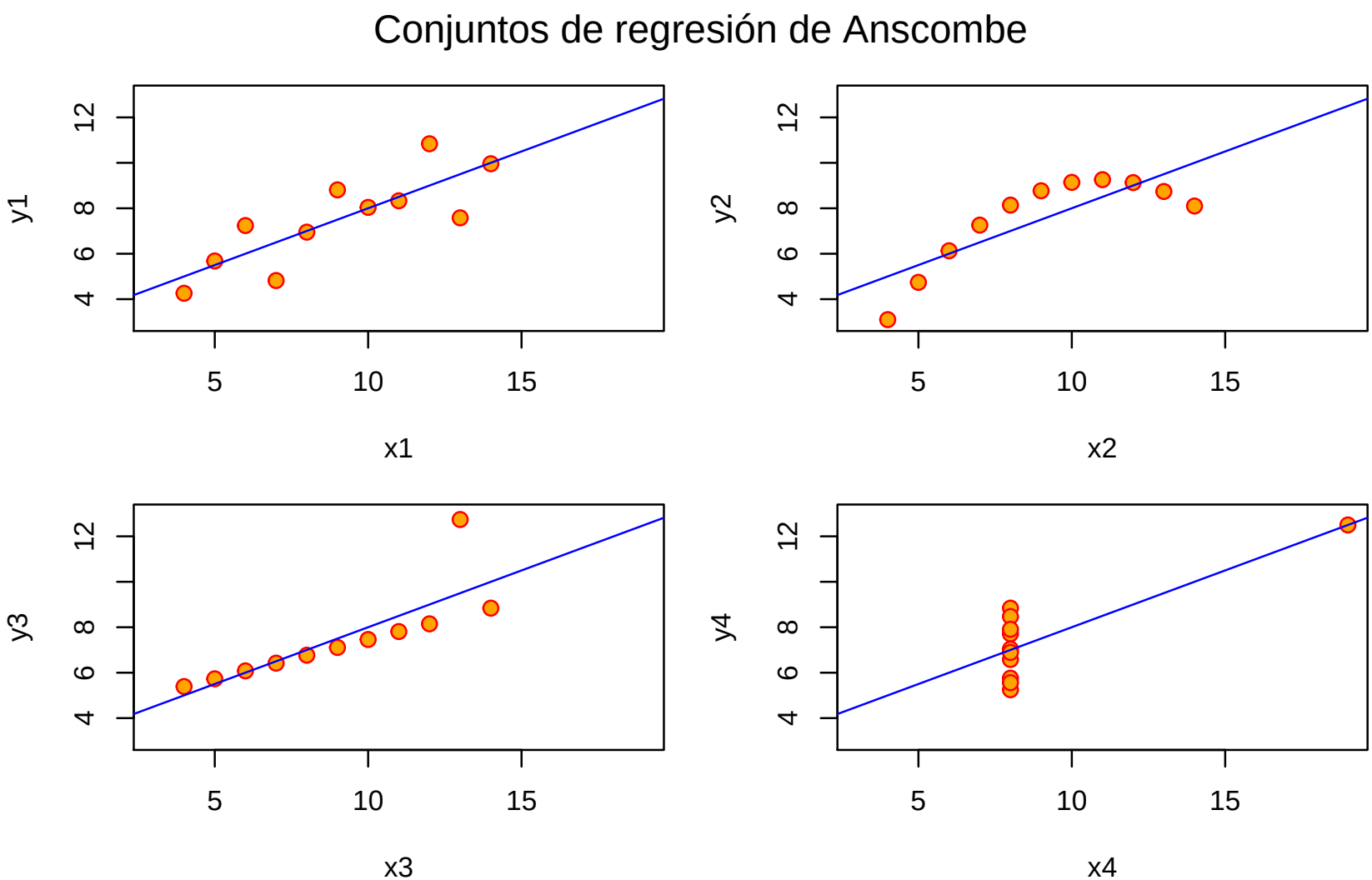
El coeficiente de determinación es el porcentaje de variabilidad total explicada por la relación lineal con el predictor

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Características:

- $0 \leq R^2 \leq 1$
- R^2 es el coeficiente de correlación al cuadrado
- R^2 puede ser una medida engañosa de ajuste del modelo
- La adición de variables regresoras al modelo siempre va a incrementar el valor de R^2

Coeficiente de determinación - R^2



Modelo lineal compuesto

Coeficientes de Determinación

El coeficiente de determinación ajustado representa el porcentaje de variabilidad de la variable de respuesta que explicaría el modelo si se hubiese obtenido de la población de donde proviene la muestra

$$R^2_{aj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

```
(r2a <- 1 - (1 - summary(lm(price ~ carat, diamond))$r.squared)
   ((nrow(diamond) - 1) / (nrow(diamond) - ncol(diamond))))

## [1] 0.9777882
```

Si los valores de los dos coeficientes de determinación se encuentran cercanos es indicativo que el modelo de regresión es bueno

Inferencia en la regresión

Considere estadísticos como los siguientes:

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}}$$

donde $\hat{\theta}$ es el estadístico estimado de interés, θ es el estimando de interés y $\hat{\sigma}_{\hat{\theta}}$ es el error estándar de $\hat{\theta}$. En muchos casos estos estadísticos tienen las siguientes propiedades:

- Se distribuyen de manera normal y presentan una distribución t de Student para muestras finitas bajo los supuestos de la normalidad
- Ellos pueden ser utilizados para probar $H_0 : \theta = \theta_0$ versus $H_a : \theta >, <, \neq \theta_0$
- Ellos pueden ser utilizados para crear intervalos de confianza para θ via $\hat{\theta} \pm Q_{1-\alpha/2}\hat{\sigma}_{\hat{\theta}}$ donde $Q_{1-\alpha/2}$ es el cuantil relevante ya sea de una distribución normal o t

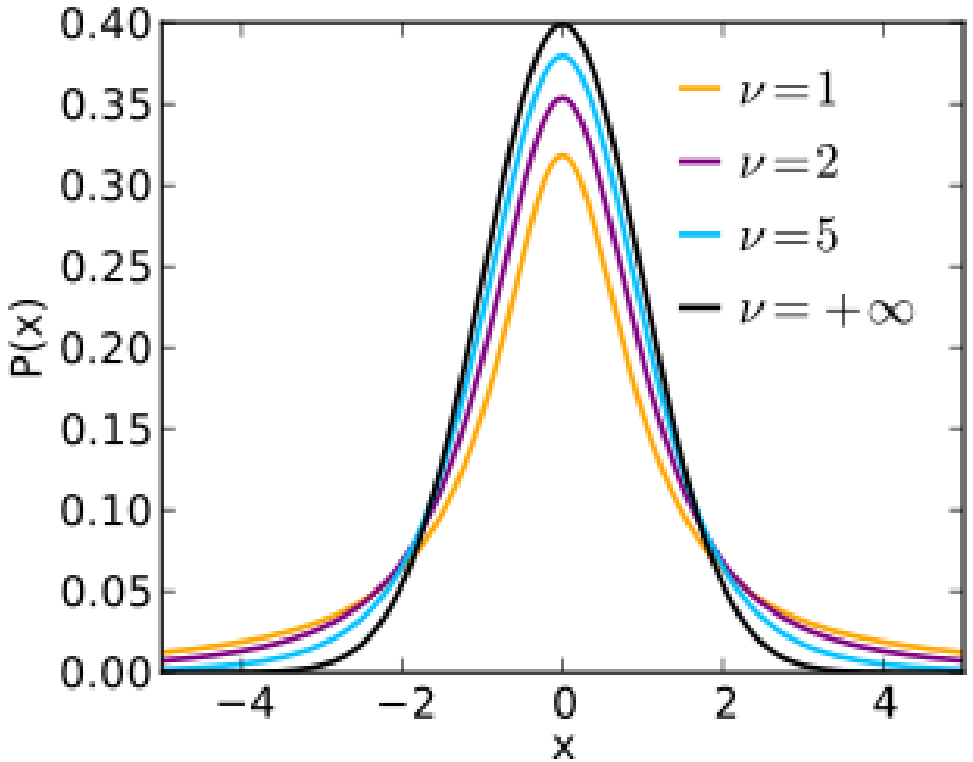
Inferencia en la regresión

Distribución t

- La distribución de t constituye la base para pruebas de significancia principalmente en casos en donde el tamaño de muestra es moderado
- Determina si la diferencia entre las medias de dos muestras con distribución normal es cero
- Construcción de intervalos de predicción para muestras de distribuciones normales con media y varianza desconocidas

Inferencia en la regresión

Distribución t



Inferencia en la regresión

Parámetros de la regresión

Los errores estándar de los parámetros de la regresión están dados por:

$$\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\sigma^2_{\hat{\beta}_0} = Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sigma^2$$

De esta forma:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

sigue una distribución t con $n-2$ grados de libertad y una distribución normal para valores grandes de n. Este estadístico puede ser utilizado para crear intervalos de confianza y realizar pruebas de hipótesis

Inferencia en la regresión

Parámetros de la regresión

```
y <- diamond$price; x <- diamond$carat; n <- length(y)
beta1 <- cor(y, x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
e <- y - beta0 - beta1 * x
sigma <- sqrt(sum(e^2) / (n-2))
ssx <- sum((x - mean(x))^2)
seBeta0 <- (1 / n + mean(x) ^ 2 / ssx) ^ .5 * sigma
seBeta1 <- sigma / sqrt(ssx)
tBeta0 <- beta0 / seBeta0
tBeta1 <- beta1 / seBeta1
```

Inferencia en la regresión

Parámetros de la regresión

```
pBeta0 <- 2 * pt(abs(tBeta0), df = n - 2, lower.tail = FALSE)
pBeta1 <- 2 * pt(abs(tBeta1), df = n - 2, lower.tail = FALSE)
coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0), c(beta1,
            seBeta1, tBeta1, pBeta1))
colnames(coefTable) <- c("Estimate", "Std. Error", "t value",
                         "P(>|t|)")
rownames(coefTable) <- c("(Intercept)", "x")
coefTable
```

```
##               Estimate Std. Error   t value      P(>|t|)
## (Intercept) -259.6259   17.31886 -14.99094 2.523271e-19
## x           3721.0249   81.78588  45.49715 6.751260e-40
```

Inferencia en la regresión

Intervalos de confianza para los parámetros

Los intervalos de confianza toman la forma de un valor estimado más o menos un cuantil multiplicado por un error estándar

```
sumCoef <- summary(fit)$coefficients
sumCoef[1,1] + c(-1, 1) * qt(.975, df = fit$df) *
  sumCoef[1, 2]

## [1] -294.4870 -224.7649

(sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) *
  sumCoef[2, 2])/10

## [1] 355.6398 388.5651
```

Lo cual se interpretaría como: "con un 95% de confianza, se estima que un incremento de 0.1 quilates en el tamaño de un diamante resultará en un incremento de precio entre 355.6 y 388.6 dólares"

Inferencia en la regresión

Intervalos de confianza para las predicciones

La predicción para el punto x_0 está dada por:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

Es necesario añadir un error estándar para crear el intervalo de confianza

El solo valor de la predicción no da indicación alguna sobre cuán preciso es ese valor que se está estimando

Inferencia en la regresión

Intervalos de confianza para las predicciones

Existe una diferencia sutil entre los intervalos de confianza para la línea de regresión en el punto x_0 y la predicción de cuál será el valor de y en el punto x_0. La diferencia está dada por el error estándar:

El error estándar para la línea en x_0 es:

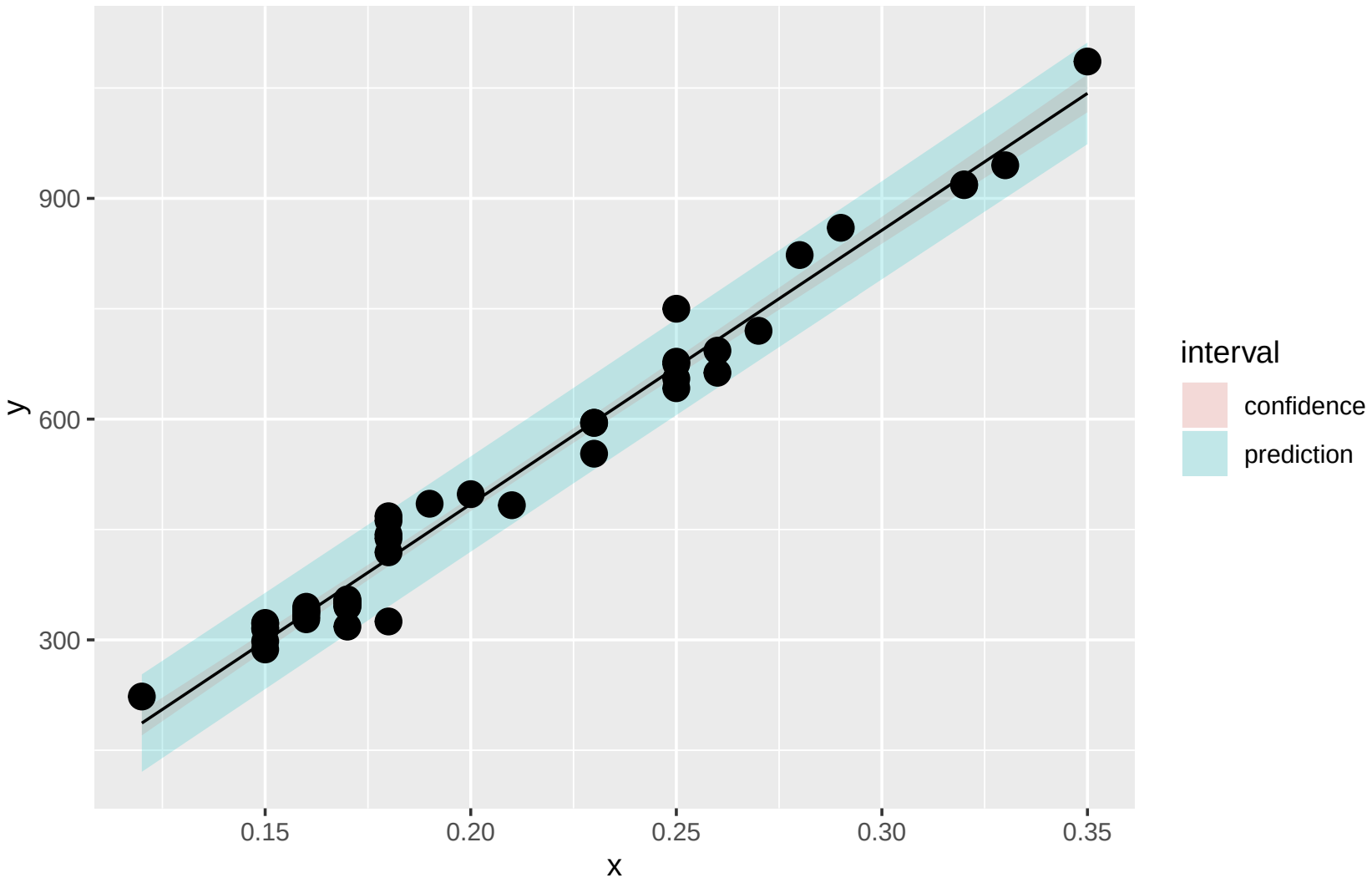
$$\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{X})^2}{\sum_{i=1} n(X_i-\bar{X})^2}}$$

El error estándar para el intervalo de predicción en el punto x_0 es:

$$\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{X})^2}{\sum_{i=1} n(X_i-\bar{X})^2}}$$

Inferencia en la regresión

Intervalos de confianza para las predicciones



Inferencia en la regresión

Intervalos de confianza para las predicciones

```
library(ggplot2)
newx <- data.frame(x = seq(min(x), max(x), length = 100))
p1 <- data.frame(predict(fit, newdata= newx,interval = ("confidence")))
p2 <- data.frame(predict(fit, newdata = newx,interval = ("prediction")))
p1$interval <- "confidence"
p2$interval <- "prediction"
p1$x <- newx$x
p2$x <- newx$x
dat <- rbind(p1, p2)
names(dat)[1] <- "y"
ggplot(dat, aes(x = x, y = y)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr, fill = interval), alpha = 0.2) +
  geom_line() +
  geom_point(data = data.frame(x = x, y = y), aes(x = x, y = y), size = 4)
```