videos

regrecion multivariable

- El modelo lineal general extiende la regresión lineal simple (SLR) agregando términos linealmente en el modelo.

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ik}\beta_j + \epsilon_i$$

- aquí $X_{1i} = 1$ típicamente, por lo que se incluye una intersección.
- Mínimos cuadrados (y, por lo tanto, estimaciones de ML bajo iid Gaussianity de los errores) minimiza

$$\sum_{i=1}^{n}\left(Y_i - \sum_{k=1}^{p} X_{ki}\beta_j\right)^2$$

- Tenga en cuenta que la linealidad importante es la linealidad en los coeficientes. Por lo tanto

$$Y_i = \beta_1 X_{1i}^2 + \beta_2 X_{2i}^2 + \ldots + \beta_p X_{pi}^2 + \epsilon_i$$

  sigue siendo un modelo lineal. (Acabamos de cuadrar los elementos de las variables predictoras).
- Recuerde que la estimación LS para la regresión a través del origen, $E[Y_i] = X_{1i}\beta_1$, sea $\sum X_i Y_i / \sum X_i^2$.
- Let's consider two regressors, $E[Y_i] = X_{1i}\beta_1 + X_{2i}\beta_2 = \mu_i$.
- Mínimos cuadrados intenta minimizar

$$\sum_{i=1}^{n}(Y_i - X_{1i}\beta_1 - X_{2i}\beta_2)^2$$

Para mas formalizacion ver el video "regression part II"

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,X_1|X_2}^2}$$

* Es decir, la estimación de regresión para β_1 es la regresión a través de la estimación de origen habiendo retrocedido X_2 de ambos la respuesta y el predictor. * (De manera similar, la estimación de regresión para β_2 es la regresión a través de la estimación de origen habiendo regresado X_1 tanto de la respuesta como del predictor). * De manera más general, las estimaciones de regresión multivariante son exactamente aquellas que han eliminado la relación lineal de las otras variables tanto del regresor como de la respuesta.

ejemplo con dos varibles * $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i}$ donde $X_{2i} = 1$ es un término de intersección. * Observe el coeficiente ajustado de X_{2i} en Y_i es $\bar{Y}$ * Los residuales $e_{i,Y|X_2} = Y_i - \bar{Y}$ * Observe el coeficiente ajustado de X_{2i} en X_{1i} es $\bar{X}_1$ * Los residuales $e_{i,X_1|X_2} = X_{1i} - \bar{X}_1$ * Por lo tanto

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,X_1|X_2}^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = Cor(X,Y)\frac{Sd(Y)}{Sd(X)}$$

dem

```
n = 100; x = rnorm(n); x2 = rnorm(n); x3 = rnorm(n)
y = 1 + x + x2 + x3 + rnorm(n, sd = .1)
ey = resid(lm(y ~ x2 + x3))
ex = resid(lm(x ~ x2 + x3))
sum(ey * ex) / sum(ex ^ 2)
```

```
## [1] 0.9822965
```

```
coef(lm(ey ~ ex - 1))
```

```
##        ex
## 0.9822965
```

```
coef(lm(y ~ x+x2 + x3))
```

```
## (Intercept)           x          x2          x3
##   0.9864135   0.9822965   0.9964453   0.9908469
```

Todas nuestras cantidades de SLR se pueden ampliar a modelos lineales * Modelo $Y_i = \sum_{k=1}^{p} X_{ik}\beta_k + \epsilon_i$ donde $\epsilon_i \sim N(0, \sigma^2)$ * Respuestas ajustadas $\hat{Y}_i = \sum_{k=1}^{p} X_{ik}\hat{\beta}_k$ * Residuales $e_i = Y_i - \hat{Y}_i$ * estimacion de la varianza $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$ * Para obtener respuestas pronosticadas en nuevos valores, $x_1, \ldots, x_p$, simplemente conéctelos al modo lineal $\sum_{k=1}^{p} x_k\hat{\beta}_k$ * Los coeficientes tienen errores estándar, $\hat{\sigma}_{\hat{\beta}_k}$, y $\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$ sigue una distribucioj T con $n-p$ grados de libertad. * Las respuestas pronosticadas tienen errores estándar y podemos calcular los intervalos de respuesta pronosticados y esperados.

paradoja de simpson

En probabilidad y estadística, la paradoja de Simpson o efecto Yule-Simpson es una paradoja en la cual una tendencia que aparece en varios grupos de datos desaparece cuando estos grupos se combinan y en su lugar aparece la tendencia contraria para los datos agregados. Esta situación se presenta con frecuencia en las ciencias sociales, en los experimentos de Andre y en la estadística médica. y es causa de confusión cuando a la frecuencia de los datos se le asigna sin fundamento una interpretación causal. La paradoja desaparece cuando se analizan las relaciones causales presentes.

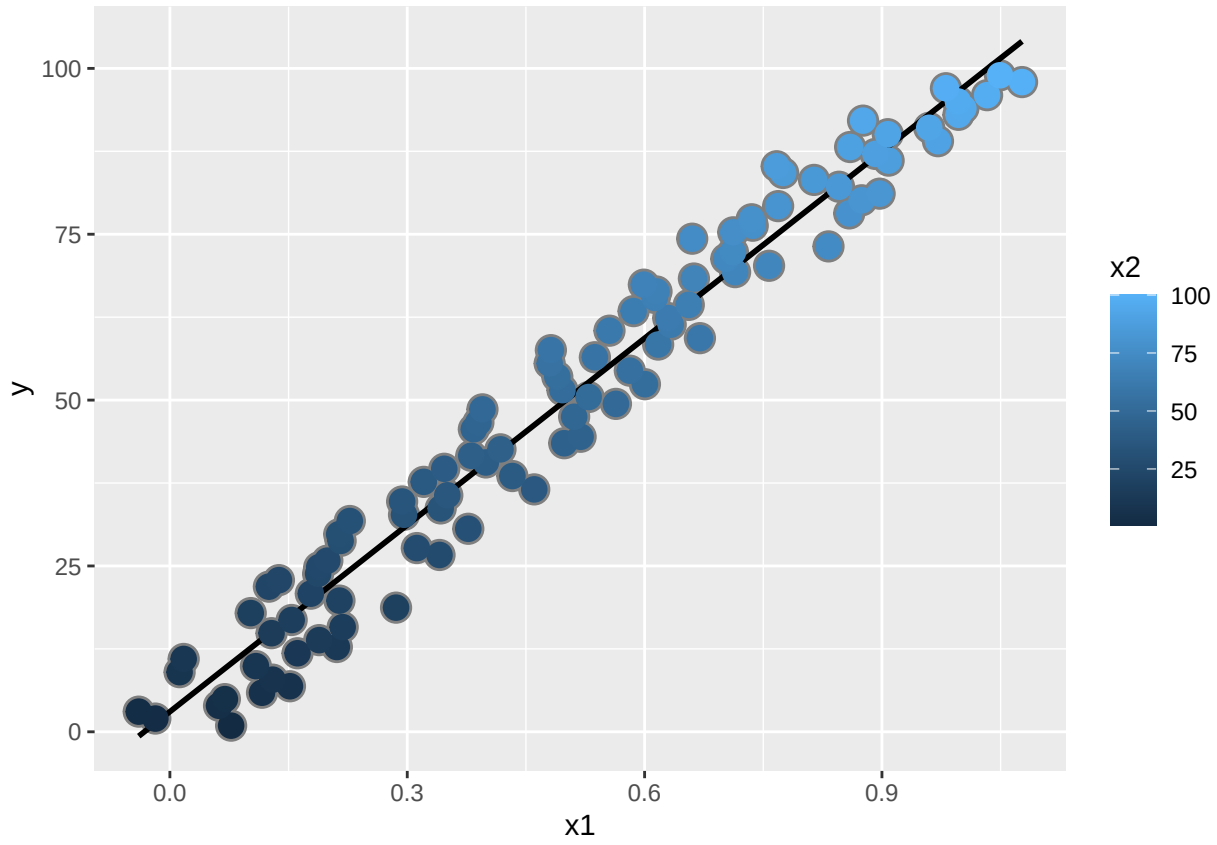
```
n <- 100; x2 <- 1 : n; x1 <- .01 * x2 + runif(n, -.1, .1); y = -x1 + x2 + rnorm(n, sd = .01)
summary(lm(y ~ x1))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.035096   1.081079  2.807468 6.025634e-03
## x1          93.809441   1.854052 50.596981 4.782533e-72
```

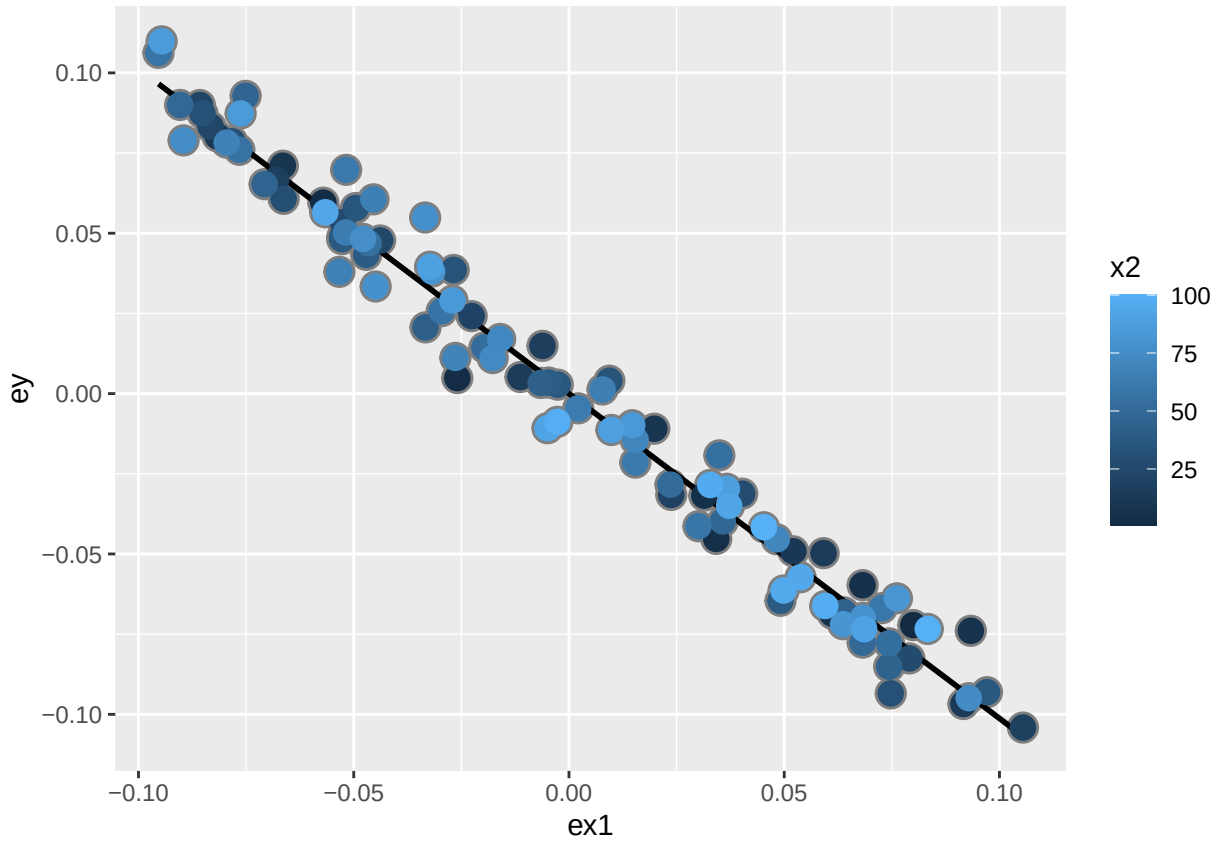
```
summary(lm(y ~ x1 + x2))$coef
```

```
##                   Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept) -0.0006456894 0.0017788471   -0.362982  7.174081e-01
## x1          -1.0127126205 0.0154440386  -65.573044  3.787594e-82
## x2           1.0001489341 0.0001599295 6253.686246 1.098024e-273
```

```
dat = data.frame(y = y, x1 = x1, x2 = x2, ey = resid(lm(y ~ x2)), ex1 = resid(lm(x1 ~ x2)))
library(ggplot2)
g = ggplot(dat, aes(y = y, x = x1, colour = x2))
g = g + geom_point(colour="grey50", size = 5) + geom_smooth(method = lm, se = FALSE, colour = "black")
g = g + geom_point(size = 4)
g
```




```
g2 = ggplot(dat, aes(y = ey, x = ex1, colour = x2))
g2 = g2 + geom_point(colour="grey50", size = 5) + geom_smooth(method = lm, se = FALSE, colour = "black"
g2
```



variables ficticias

- Considere el modelo lineal

$$Y_i = \beta_0 + X_{i1}\beta_1 + \epsilon_i$$

donde cada $X_ {i1} $ es binario, por lo que es un 1 si la medida i está en un grupo y 0 en caso contrario. (Tratados versus no en un ensayo clínico, por ejemplo). * Luego para las personas del grupo $E[Y_i] = \beta_0 + \beta_1$ * Y para las personas que no están en el grupo $E[Y_i] = \beta_0$ * Los ajustes de LS resultan ser $\hat{\beta}_0 + \hat{}beta_1$ es la media para aquellos en el grupo y $\hat{\beta}_0$ es la media para aquellos que no están en el grupo. * β_1 se interpreta como el aumento o disminución de la media comparando a los del grupo con los que no. * Tenga en cuenta que incluir una variable binaria que sea 1 para aquellos que no están en el grupo sería redundante. Crearía tres parámetros para describir dos medios.

mas de dos niveles

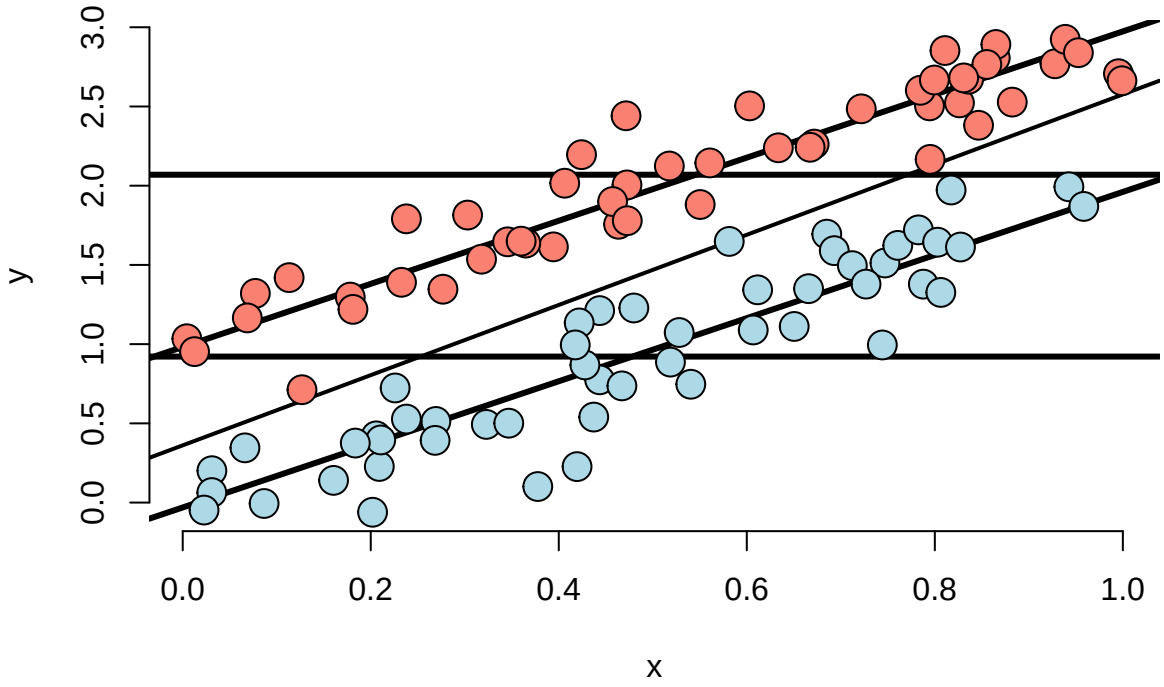
- Considere un nivel de factor multinivel. Por razones didácticas, digamos un factor de tres niveles (ejemplo, afiliación a un partido político estadounidense: republicano, demócrata, independiente)
- $Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i$.
- X_{i1} i1 para republicanos y 0 en caso contrario.
- X_{i2} Es 1 para los demócratas y 0 en caso contrario.
- si i es republicano $E[Y_i] = \beta_0 + \beta_1$

- si i es Democrata $E[Y_i] = \beta_0 + \beta_2$.
- si i es Independente $E[Y_i] = \beta_0$.
- β_1 Compara republicanos a independientes.
- β_2 Compara a los demócratas con los independientes.
- $\beta_1 - \beta_2$ compara a los republicanos con los demócratas.
- (La elección de la categoría de referencia cambia la interpretación).

ajustamiento

simulacion 1

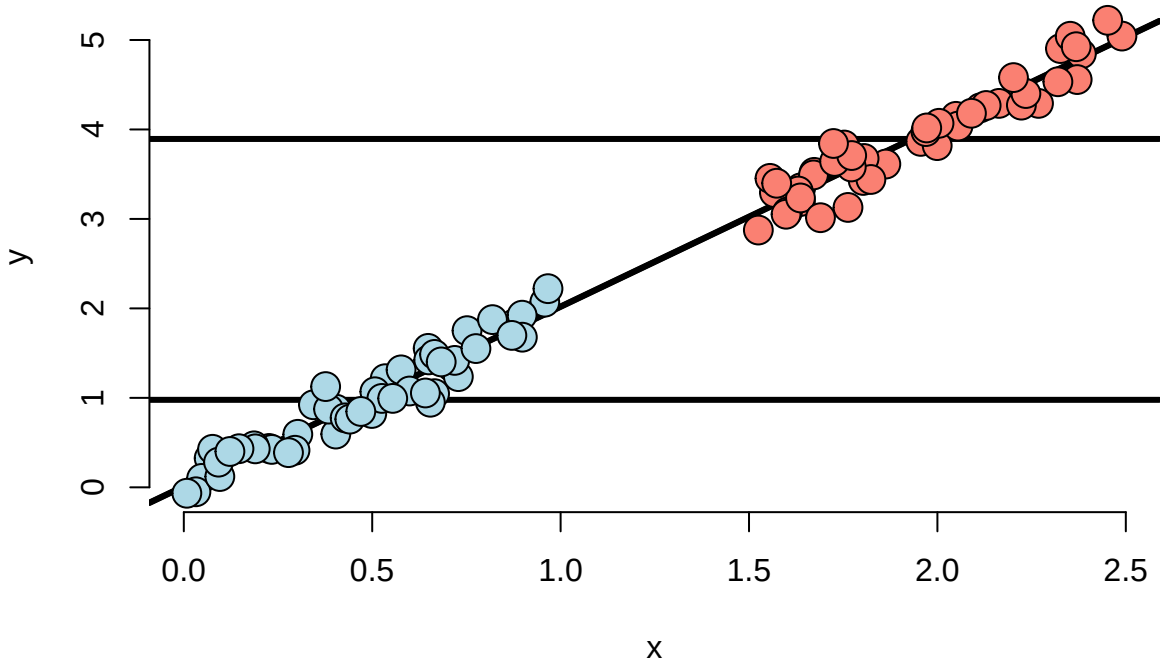
```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(runif(n/2), runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 1; sigma <- .2
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```



- La variable X no está relacionada con el estado del grupo
- La variable X está relacionada con Y, pero la intersección depende sobre el estado del grupo.
- La variable de grupo está relacionada con Y.
  - La relación entre el estado del grupo e Y es constante dependiendo de X.
  - La relación entre grupo e Y sin tener en cuenta X es aproximadamente la misma que mantener X constante

simulacion 2

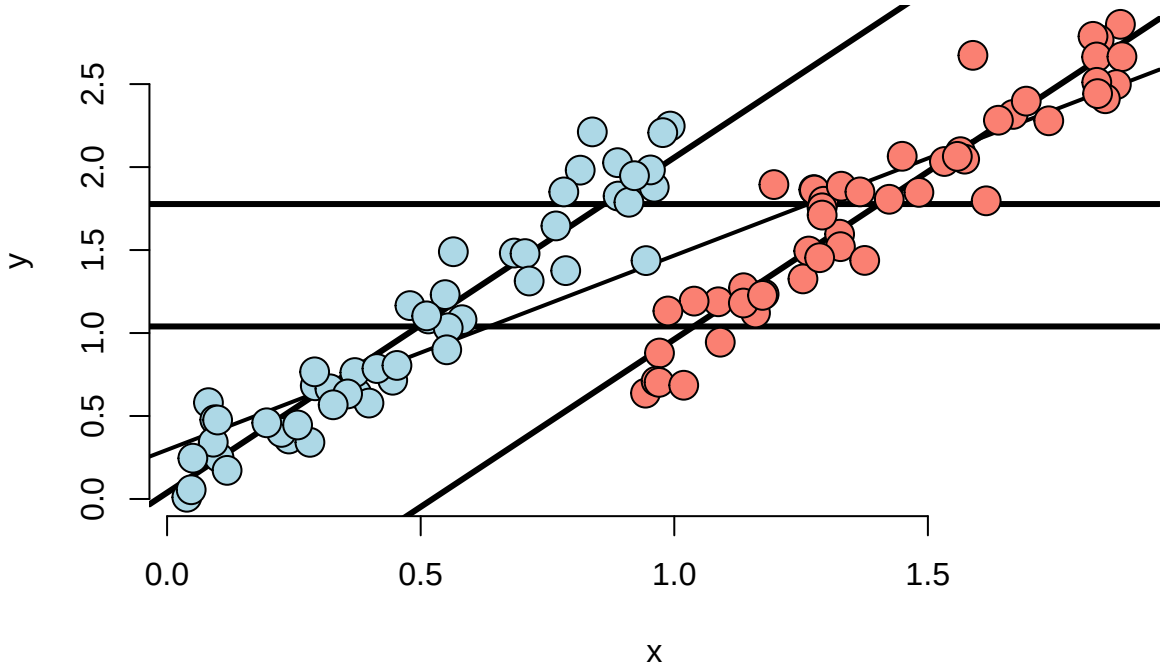
```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(runif(n/2), 1.5 + runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 0; sigma <- .2
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```



- La variable X está muy relacionada con el estado del grupo.
- La variable X está relacionada con Y, la intersección no depende de la variable de grupo.
  - La variable X permanece relacionada con el estado del grupo de retención Y constante
- La variable de grupo está marginalmente relacionada con Y sin tener en cuenta X.
- El modelo no estimaría ningún efecto ajustado debido al grupo.
  - No hay datos para informar la relación entre grupo e Y.
  - Esta conclusión se basa completamente en el modelo.

simulacion 3

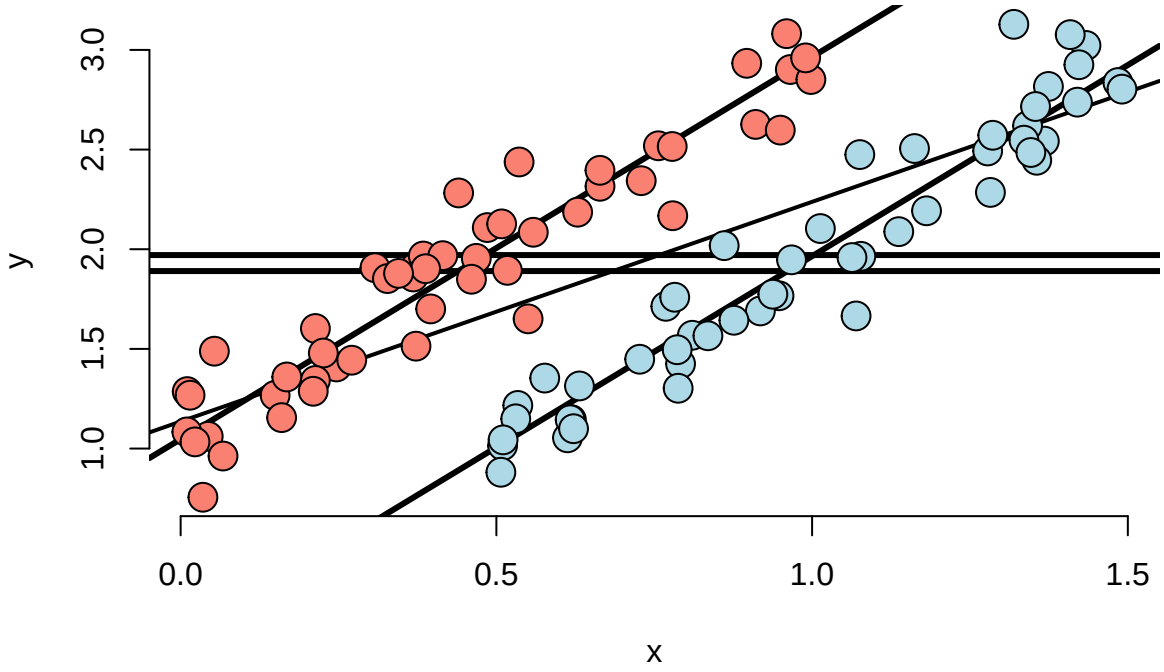
```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(runif(n/2), .9 + runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- -1; sigma <- .2
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```



- La asociación marginal tiene un grupo rojo más alto que el azul.
- La relación ajustada tiene un grupo azul más alto que rojo.
- Estado del grupo relacionado con X.
- Existe alguna evidencia directa para comparar rojo y azul manteniendo X fija.

simulacion 4

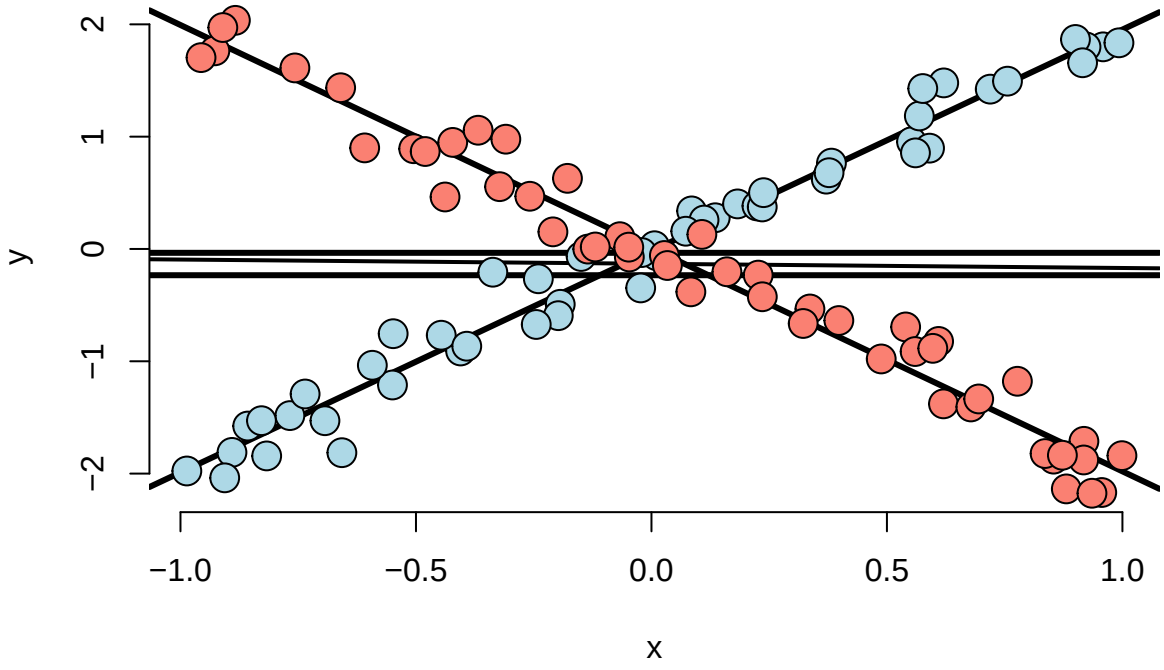
```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(.5 + runif(n/2), runif(n/2));
beta0 <- 0; beta1 <- 2; tau <- 1; sigma <- .2
y <- beta0 + x * beta1 + t * tau + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t)
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```



- No marginal association between group status and Y.
- Strong adjusted relationship.
- Group status not related to X.
- There is lots of direct evidence for comparing red and blue holding X fixed.

simulacion 5

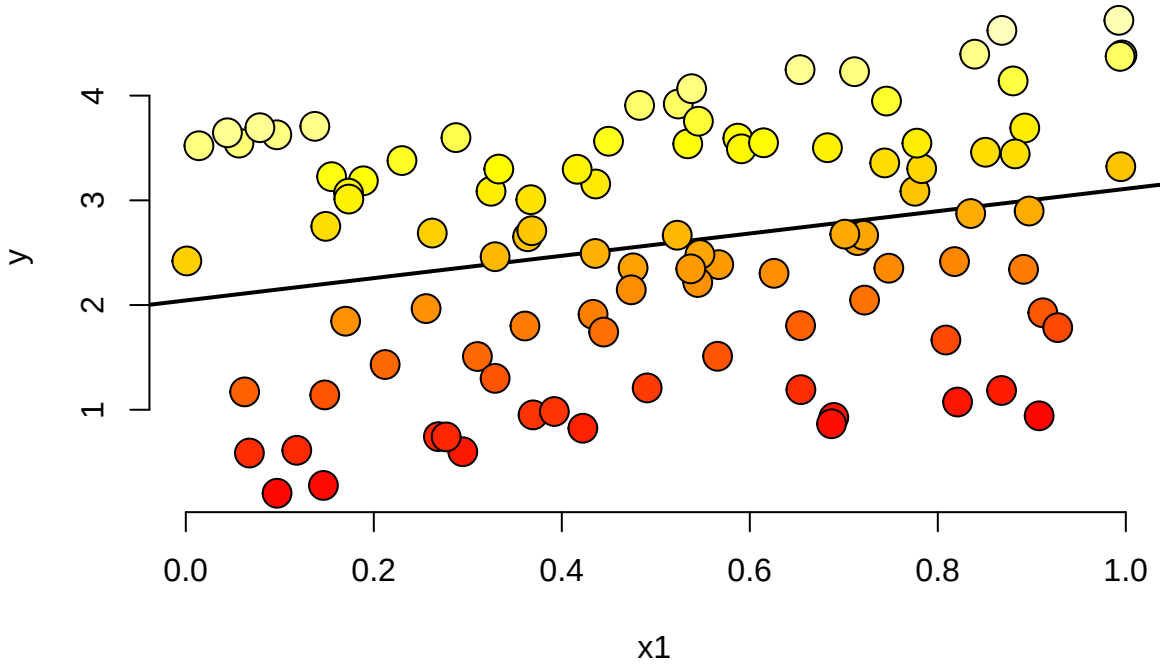
```
n <- 100; t <- rep(c(0, 1), c(n/2, n/2)); x <- c(runif(n/2, -1, 1), runif(n/2, -1, 1));
beta0 <- 0; beta1 <- 2; tau <- 0; tau1 <- -4; sigma <- .2
y <- beta0 + x * beta1 + t * tau + t * x * tau1 + rnorm(n, sd = sigma)
plot(x, y, type = "n", frame = FALSE)
abline(lm(y ~ x), lwd = 2)
abline(h = mean(y[1 : (n/2)]), lwd = 3)
abline(h = mean(y[(n/2 + 1) : n]), lwd = 3)
fit <- lm(y ~ x + t + I(x * t))
abline(coef(fit)[1], coef(fit)[2], lwd = 3)
abline(coef(fit)[1] + coef(fit)[3], coef(fit)[2] + coef(fit)[4], lwd = 3)
points(x[1 : (n/2)], y[1 : (n/2)], pch = 21, col = "black", bg = "lightblue", cex = 2)
points(x[(n/2 + 1) : n], y[(n/2 + 1) : n], pch = 21, col = "black", bg = "salmon", cex = 2)
```



- Aquí no existe el efecto de grupo.
  - El impacto del grupo se invierte dependiendo de X.
  - Tanto la intersección como la pendiente dependen del grupo.
- Estado del grupo y X no relacionados.
  - Hay mucha información sobre los efectos de grupo que mantienen X fijo.

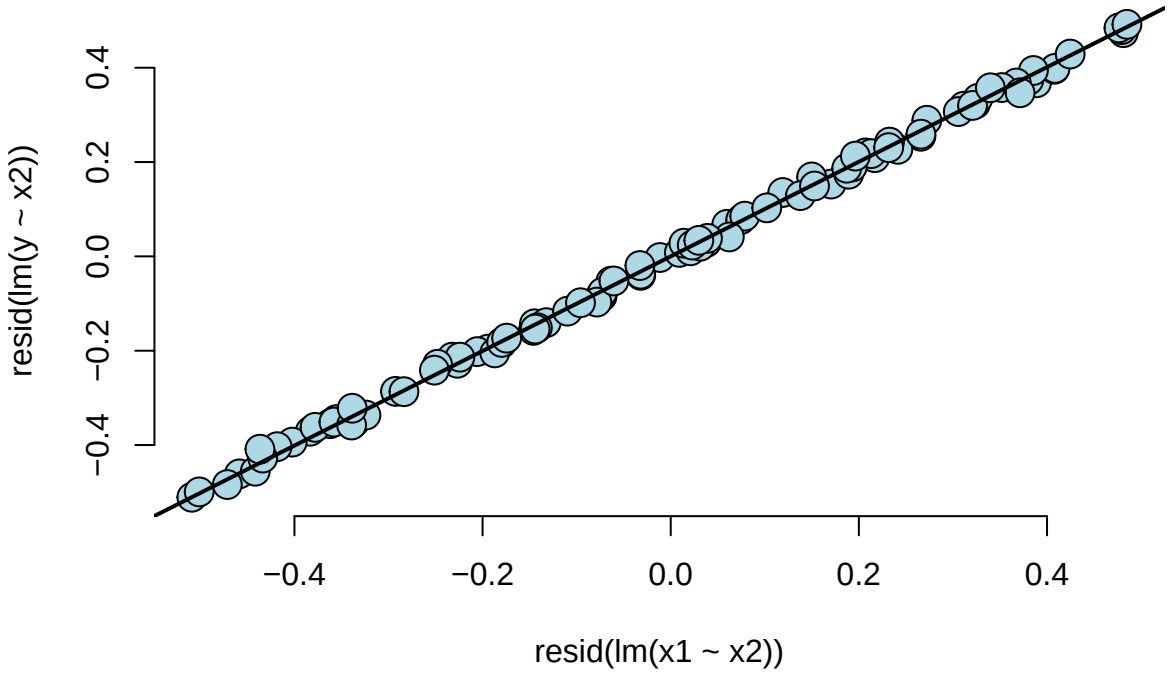
Simulation 6

```
p <- 1
n <- 100; x2 <- runif(n); x1 <- p * runif(n) - (1 - p) * x2
beta0 <- 0; beta1 <- 1; tau <- 4 ; sigma <- .01
y <- beta0 + x1 * beta1 + tau * x2 + rnorm(n, sd = sigma)
plot(x1, y, type = "n", frame = FALSE)
abline(lm(y ~ x1), lwd = 2)
co.pal <- heat.colors(n)
points(x1, y, pch = 21, col = "black", bg = co.pal[round((n - 1) * x2 + 1)], cex = 2)
```




```
library(rgl)
plot3d(x1, x2, y)
```

```
plot(resid(lm(x1 ~ x2)), resid(lm(y ~ x2)), frame = FALSE, col = "black", bg = "lightblue", pch = 21, c
abline(lm(I(resid(lm(x1 ~ x2))) ~ I(resid(lm(y ~ x2)))), lwd = 2)
```



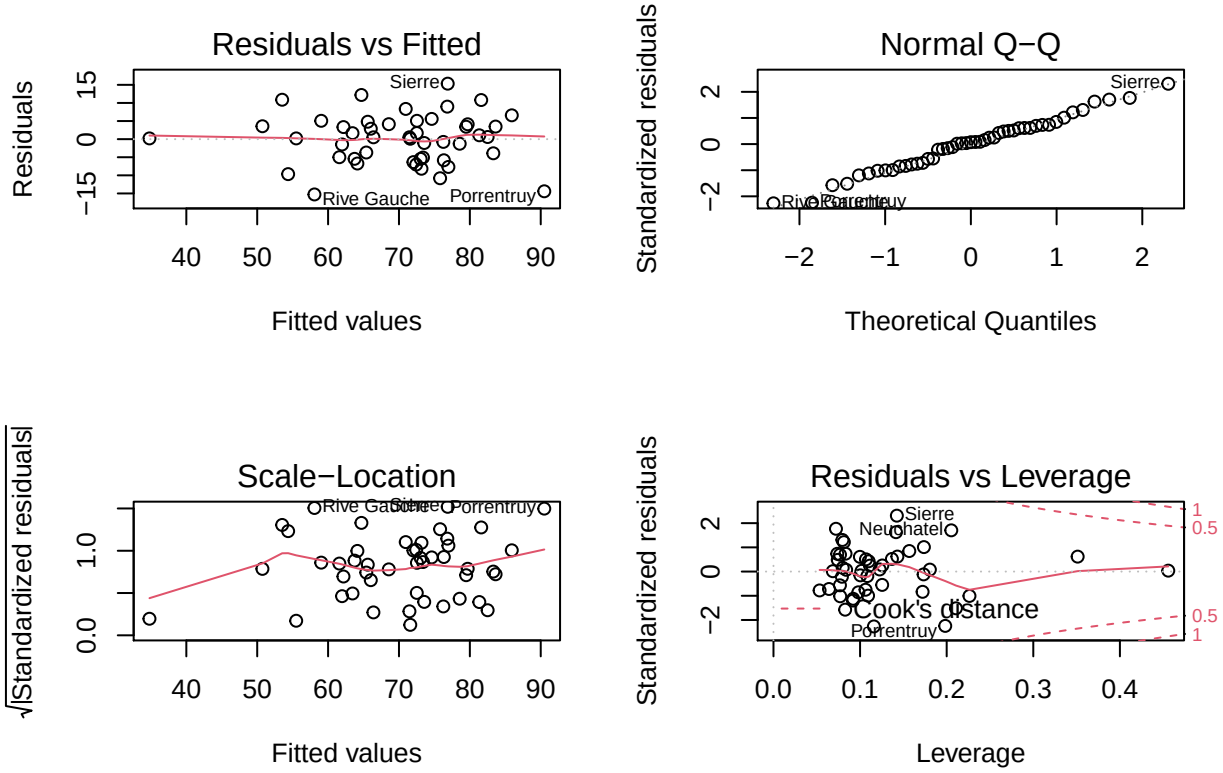
- X1 no relacionado con X2
- X2 fuertemente relacionado con Y
- Relación ajustada entre X1 e Y prácticamente sin cambios considerando X2.
  - Casi sin variabilidad residual después de considerar X2.

Algunos pensamientos finales

- Modelar relaciones multivariadas es difícil.
- Juega con simulaciones para ver cómo la inclusión o exclusión de otra variable puede análisis de cambios.
- Los resultados de estos análisis se refieren a la impacto de las variables en las asociaciones.
  - Determinar los mecanismos o la causa son temas difíciles. para agregar a la dificultad para comprender las asociaciones multivariadas.

residuals

```
data(swiss); par(mfrow = c(2, 2))
fit <- lm(Fertility ~ . , data = swiss); plot(fit)
```




```
par(mfrow = c(1, 1))
```

- Haga `? Influence.measures` para ver el conjunto completo de medidas de influencia en las estadísticas. Las medidas incluyen
  - `rstandard` - residuales estandarizados, residuales divididos por sus desviaciones estándar)
  - `rstudent`: residuales estandarizados, residuales divididos por sus desviaciones estándar, donde el i-ésimo punto de datos se eliminó en el cálculo de la desviación estándar para que el residual siga una distribución t
  - `hatvalues` - medidas de apalancamiento
  - `dffits`: cambio en la respuesta prevista cuando se elimina el punto i^{th} al ajustar el modelo.
  - `dfbetas`: cambio en los coeficientes individuales cuando se elimina el punto i^{th} al ajustar el modelo.
  - `cooks.distance` - cambio general en los coeficientes cuando se elimina el punto $i^{th} $.
  - `resid` - devuelve los residuos ordinarios
  - `resid (ajuste) / (1 - hatvalues(ajuste))` donde `ajuste` es el ajuste del modelo lineal devuelve los residuos de PRESS, es decir, los residuos de validación cruzada que se dejan fuera - la diferencia en la respuesta y la respuesta predicha en los datos punto i, donde no se incluyó en el ajuste del modelo.
- Tenga cuidado con las reglas simplistas para los diagramas y medidas de diagnóstico. El uso de estas herramientas depende del contexto. Es mejor comprender lo que están tratando de lograr y usarlos con prudencia.
- No todas las medidas tienen escalas absolutas significativas. Puede verlos en relación con los valores de los datos.
- Sondean sus datos de diferentes formas para diagnosticar diferentes problemas.

- Los patrones en sus gráficos de residuos generalmente indican algún aspecto deficiente del ajuste del modelo. Estos pueden incluir:
  - Heteroscedasticidad (varianza no constante).
  - Faltan términos del modelo.
  - Patrones temporales (trazar residuos versus orden de recolección).
- Los gráficos de QQ residual investigan la normalidad de los errores.
- Las medidas de apalancamiento (valores de sombrero) pueden ser útiles para diagnosticar errores de entrada de datos.
- Las medidas de influencia llegan a la conclusión, 'cómo la eliminación o la inclusión de este punto impacta un aspecto particular del modelo'.

siempre checar residuos

seleccion de modelos

- Tenemos una clase completa sobre predicción y aprendizaje automático, por lo que nos centraremos en el modelado.
  - La predicción tiene un conjunto diferente de criterios, necesidades de interpretación y estándares de generalización.
  - En el modelado, nuestro interés radica en representaciones parsimoniosas e interpretables de los datos que mejoran nuestra comprensión de los fenómenos en estudio.
  - Un modelo es una lente a través de la cual mirar sus datos. (Atribuyo esta cita a Scott Zeger)
  - Bajo esta filosofía, ¿cuál es el modelo correcto? Cualquiera sea el modelo que conecte los datos con una declaración verdadera y parsimoniosa sobre lo que está estudiando.
- Hay formas casi incontables en las que un modelo puede estar equivocado; en esta conferencia, nos centraremos en la inclusión y exclusión de variables.
- Como casi todos los aspectos de las estadísticas, las buenas decisiones de modelado dependen del contexto.
  - Un buen modelo de predicción versus uno para estudiar mecanismos versus uno para tratar de establecer efectos causales puede no ser el mismo
- Hay conocidos conocidos. Estas son cosas que sabemos que sabemos. Hay incógnitas conocidas. Es decir, hay cosas que sabemos que no sabemos. Pero también hay incógnitas desconocidas. Hay cosas que no sabemos que no sabemos. * Donald Rumsfeld

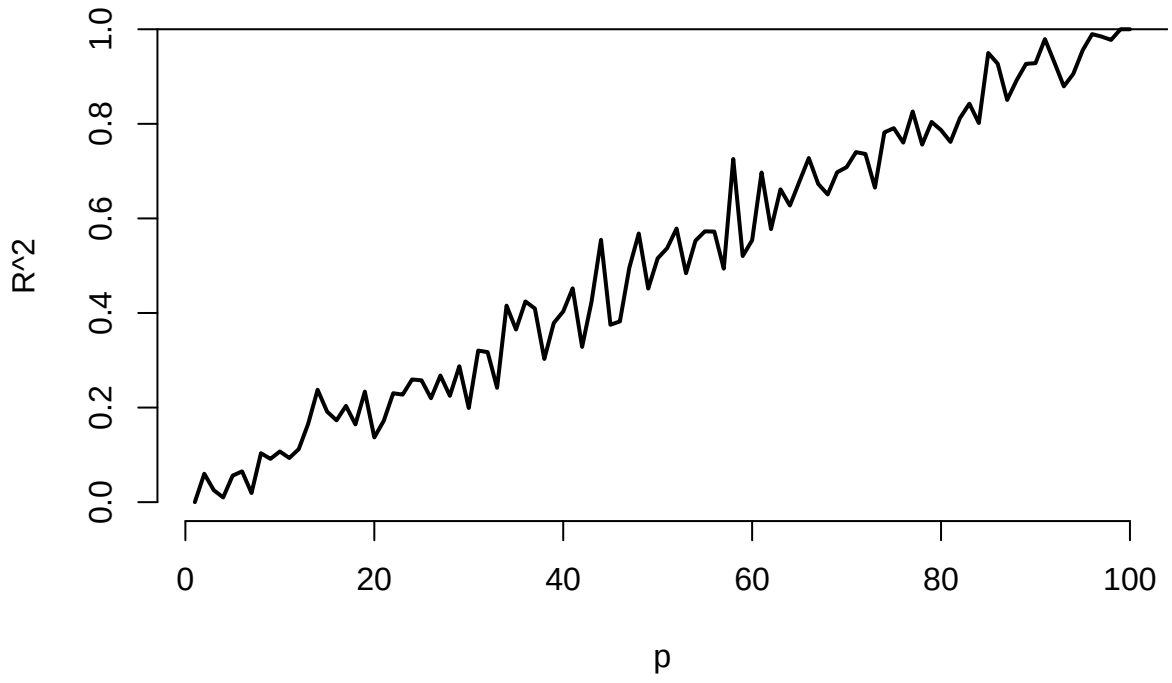
En nuestro contexto * (Conocimientos conocidos) Regresores que sabemos que debemos verificar para incluir en el modelo y tener. * (Desconocidas conocidas) Regresores que nos gustaría incluir en el modelo, pero que no tenemos. * (Desconocidas desconocidas) Regresores que ni siquiera conocemos que deberíamos haber incluido en el modelo.

- La omisión de variables da como resultado un sesgo en los coeficientes de interés, a menos que sus regresores no estén correlacionados con los omitidos.
  - Es por eso que aleatorizamos los tratamientos, intenta descorrelacionar nuestro indicador de tratamiento con variables que no tenemos que poner en el modelo.
  - (Si hay demasiadas variables de confusión no observadas, ni siquiera la aleatorización le ayudará).
- Incluir variables que no deberíamos tener aumenta los errores estándar de las variables de regresión.
  - De hecho, incluir cualquier variable nueva aumenta los errores estándar (reales, no estimados) de otros regresores. Por tanto, no queremos introducir variables en el modelo sin hacer nada.

- El modelo debe tender a un ajuste perfecto a medida que el número de regresores no redundantes se acerca a $ n $.
- R^2 aumenta monótonamente a medida que se incluyen más regresores.
- El SSE disminuye monótonamente a medida que se incluyen más regresores.

Para simulaciones, ya que el número de variables incluidas es igual a $ n = 100 $. No existe una relación de regresión real en ninguna simulación

```
n <- 100
plot(c(1, n), 0 : 1, type = "n", frame = FALSE, xlab = "p", ylab = "R^2")
r <- sapply(1 : n, function(p)
      {
        y <- rnorm(n); x <- matrix(rnorm(n * p), n, p)
        summary(lm(y ~ x))$r.squared
      }
    )
lines(1 : n, r, lwd = 2)
abline(h = 1)
```



inlfacion de varianza

```
n <- 100; nosim <- 1000
x1 <- rnorm(n); x2 <- rnorm(n); x3 <- rnorm(n);
betas <- sapply(1 : nosim, function(i){
  y <- x1 + rnorm(n, sd = .3)
  c(coef(lm(y ~ x1))[2],
```

```
    coef(lm(y ~ x1 + x2))[2],
    coef(lm(y ~ x1 + x2 + x3))[2])
})
round(apply(betas, 1, sd), 5)
```

```
##      x1      x1      x1
## 0.02882 0.02980 0.03029
```

```
n <- 100; nosim <- 1000
x1 <- rnorm(n); x2 <- x1/sqrt(2) + rnorm(n) /sqrt(2)
x3 <- x1 * 0.95 + rnorm(n) * sqrt(1 - 0.95^2);
betas <- sapply(1 : nosim, function(i){
  y <- x1 + rnorm(n, sd = .3)
  c(coef(lm(y ~ x1))[2],
    coef(lm(y ~ x1 + x2))[2],
    coef(lm(y ~ x1 + x2 + x3))[2])
})
round(apply(betas, 1, sd), 5)
```

```
##      x1      x1      x1
## 0.03076 0.03736 0.08812
```

- Observe que la inflación de la varianza fue mucho peor cuando incluimos una variable que estaba altamente relacionada con "x1".
- No sabemos $ sigma $, por lo que solo podemos estimar el aumento en el error estándar real de los coeficientes para incluir un regresor.
- Sin embargo, $ sigma $ elimina los errores estándar relativos. Si se agregan variables secuencialmente, se puede verificar la inflación de la varianza (o sd) para incluir cada una.
- Cuando los otros regresores son realmente ortogonales al regresor de interés, entonces no hay inflación de varianza.
- El factor de inflación de la varianza (VIF) es el aumento de la varianza para el i-ésimo regresor en comparación con el escenario ideal donde es ortogonal a los otros regresores.
  - (La raíz cuadrada del VIF es el aumento de la sd . . .)
- Recuerde, la inflación de la varianza es solo una parte de la imagen. Queremos incluir ciertas variables, incluso si aumentan drásticamente nuestra varianza.

revisando la simulacion anterior

```
##doesn't depend on which y you use,
y <- x1 + rnorm(n, sd = .3)
a <- summary(lm(y ~ x1))$cov.unscaled[2,2]
c(summary(lm(y ~ x1 + x2))$cov.unscaled[2,2],
  summary(lm(y~ x1 + x2 + x3))$cov.unscaled[2,2]) / a
```

```
## [1] 1.508155 8.813392
```

```
temp <- apply(betas, 1, var); temp[2 : 3] / temp[1]
```

```
##       x1       x1
## 1.475079 8.206240
```

modelos anidados

```
data(swiss);
fit1 <- lm(Fertility ~ Agriculture, data = swiss)
a <- summary(fit1)$cov.unscaled[2,2]
fit2 <- update(fit, Fertility ~ Agriculture + Examination)
fit3 <- update(fit, Fertility ~ Agriculture + Examination + Education)
  c(summary(fit2)$cov.unscaled[2,2],
    summary(fit3)$cov.unscaled[2,2]) / a
```

```
## [1] 1.891576 2.089159
```

Swiss data VIFs

```
library(car)
fit <- lm(Fertility ~ . , data = swiss)
vif(fit)
```

```
##      Agriculture      Examination        Education         Catholic
##         2.284129         3.675420         2.774943         1.937160
## Infant.Mortality
##         1.107542
```

```
sqrt(vif(fit))
```

```
##      Agriculture      Examination        Education         Catholic
##         1.511334         1.917138         1.665816         1.391819
## Infant.Mortality
##         1.052398
```

- Suponiendo que el modelo es lineal con errores iid aditivos (con varianza finita), podemos describir matemáticamente el impacto de omitir las variables necesarias o incluir las innecesarias.
  - Si no ajustamos el modelo, la estimación de la varianza está sesgada.
  - Si ajustamos correctamente o sobreajustamos el modelo, incluidas todas las covariables necesarias y / o las covariables innecesarias, la estimación de la varianza es insesgada.
    - Sin embargo, la varianza de la varianza es mayor si incluimos variables innecesarias.
- La selección automática de covariables es un tema difícil. Depende en gran medida de la riqueza del espacio covariable que se quiera explorar.
  - El espacio de modelos explota rápidamente a medida que agrega interacciones y términos polinomiales.
- En la clase de predicción, cubriremos muchos métodos modernos para atravesar espacios de modelos grandes con fines de predicción.
- Los componentes principales o modelos analíticos de factores sobre covariables suelen ser útiles para reducir espacios de covariables complejos.
- Un buen diseño a menudo puede eliminar la necesidad de realizar búsquedas complejas de modelos en los análisis; aunque a menudo el control sobre el diseño es limitado.
- Si los modelos de interés están anidados y sin muchos parámetros que los diferencien, es bastante poco controvertido usar pruebas de razón de verosimilitud anidadas. (Ejemplo a seguir).

- Mi enfoque favorito es el siguiente. Dado un coeficiente que me interesa, me gusta usar el ajuste de covariables y múltiples modelos para probar ese efecto para evaluar su robustez y ver qué otras covariables lo eliminan. Este no es un enfoque terriblemente sistemático, pero tiende a enseñarle mucho sobre los datos a medida que se ensucia las manos.

```
fit1 <- lm(Fertility ~ Agriculture, data = swiss)
fit3 <- update(fit, Fertility ~ Agriculture + Examination + Education)
fit5 <- update(fit, Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality)
anova(fit1, fit3, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Agriculture
## Model 2: Fertility ~ Agriculture + Examination + Education
## Model 3: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     45 6283.1
## 2     43 3180.9  2    3102.2 30.211 8.638e-09 ***
## 3     41 2105.0  2    1075.9 10.477 0.0002111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

multivariable ejemplo 2 jjj

```
library(UsingR)
data(InsectSprays)
```

Esta es la segunda lección en la que veremos algunos modelos de regresión con más de una variable independiente. Comenzaremos con los datos de InsectSprays que nos hemos tomado la libertad de cargar para usted. Estos datos son parte del paquete de conjuntos de datos de R. Muestra la efectividad de diferentes aerosoles para insectos. Hemos utilizado el código de las diapositivas para mostrarle un diagrama de caja de los datos.

```
dim(InsectSprays)
```

```
## [1] 72  2
```

Entonces, este conjunto de datos contiene 72 recuentos, cada uno asociado con un aerosol diferente en particular. Los recuentos están en la primera columna y una letra que identifica el aerosol en la segunda. Para ahorrarle algo de escritura, hemos creado 6 matrices con solo los datos de recuento de cada aplicación. Las matrices tienen los nombres sx, donde x es A, B, C, D, E o F. Escriba uno de los nombres (su elección) de estas matrices para ver de qué estamos hablando.

```
library(dplyr)
sA<-subset(InsectSprays,spray=="A")
sB<-subset(InsectSprays,spray=="B")
sC<-subset(InsectSprays,spray=="C")
sD<-subset(InsectSprays,spray=="D")
```

```
sE<-subset(InsectSprays,spray=="E")
sF<-subset(InsectSprays,spray=="F")
 sC
```

```
##    count spray
## 25     0     C
## 26     1     C
## 27     7     C
## 28     2     C
## 29     3     C
## 30     1     C
## 31     2     C
## 32     1     C
## 33     3     C
## 34     0     C
## 35     1     C
## 36     4     C
```

cada aplicacion de aereosol tiene 12 recuentos

```
sapply(InsectSprays,class)
```

```
##     count     spray
## "numeric"  "factor"
```

La clase de la segunda columna de "spray" es un factor. Recuerde de las diapositivas que la ecuación que representa la relación entre un resultado particular y varios factores contiene variables binarias, una para cada factor. Estos datos tienen 6 factores, por lo que necesitamos 6 variables ficticias. Cada uno indicará si un resultado en particular (un recuento) está asociado con un factor o categoría específica (insecticida en aerosol).

Usando la función lm de R, genere el modelo lineal en el que count es la variable dependiente y spray es la independiente. Recuerde que en R la fórmula tiene la forma y ~ x, donde y depende del predictor x. El conjunto de datos es InsectSprays. Almacene el modelo en el ajuste variable

```
fit <- lm(count ~ spray, InsectSprays)
summary(fit)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  14.5000000   1.132156 12.8074279 1.470512e-19
## sprayB        0.8333333   1.601110  0.5204724 6.044761e-01
## sprayC      -12.4166667   1.601110 -7.7550382 7.266893e-11
## sprayD       -9.5833333   1.601110 -5.9854322 9.816910e-08
## sprayE      -11.0000000   1.601110 -6.8702352 2.753922e-09
## sprayF        2.1666667   1.601110  1.3532281 1.805998e-01
```

Observe que R devuelve una matriz de 6 por 4. Por conveniencia, almacene la primera columna de esta matriz, la columna Estimación, en una variable llamada est. Recuerde que la construcción R para acceder a la primera columna es x [, 1].

```
est <- summary(fit)$coef[,1]
```

tambien existe ptra forma de hacer el analisis:

```
summary(lm(count ~
             I(1 * (spray == 'B')) + I(1 * (spray == 'C')) +
             I(1 * (spray == 'D')) + I(1 * (spray == 'E')) +
             I(1 * (spray == 'F'))
           , data = InsectSprays))$coef
```

```
##                           Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)             14.5000000   1.132156 12.8074279 1.470512e-19
## I(1 * (spray == "B"))    0.8333333   1.601110  0.5204724 6.044761e-01
## I(1 * (spray == "C")) -12.4166667   1.601110 -7.7550382 7.266893e-11
## I(1 * (spray == "D"))  -9.5833333   1.601110 -5.9854322 9.816910e-08
## I(1 * (spray == "E")) -11.0000000   1.601110 -6.8702352 2.753922e-09
## I(1 * (spray == "F"))   2.1666667   1.601110  1.3532281 1.805998e-01
```

Tenga en cuenta que sprayA no aparece explícitamente en la lista de estimaciones. Sin embargo, está ahí como la primera entrada en la columna Estimación. Está etiquetado como "(Intercepción)". Esto se debe a que sprayA es el primero en la lista alfabética de los niveles del factor, y R por defecto usa el primer nivel como referencia con la que se comparan los otros niveles o grupos al realizar sus pruebas t (que se muestran en la tercera columna).

¿Qué representan las estimaciones de este modelo? Por supuesto, son los coeficientes de las variables binarias o ficticias asociadas con los aerosoles. Más importante aún, la intersección es la media del grupo de referencia, en este caso sprayA, y las otras estimaciones son las distancias de las medias de los otros grupos de la media de referencia. Verifiquemos estas afirmaciones ahora. Primero calcule la media de los recuentos de sprayA. Recuerde que todos los recuentos se almacenan en los vectores llamados sx. Ahora estamos interesados en encontrar la media de sA.

```
mean(sA)
```

```
## [1] NA
```

```
mean(sB)-mean(sA)
```

```
## [1] NA
```

Generemos otro modelo de estos datos, esta vez omitiendo la intersección. Podemos usar fácilmente la función lm de R para hacer esto agregando "- 1" a la fórmula, por ejemplo, count ~ spray - 1. Esto le dice a R que omita el primer nivel. Haga esto ahora y almacene el nuevo modelo en la variable nfit.

```
nfit <- lm(count ~ spray - 1, InsectSprays)
summary(nfit)$coef
```

```
##         Estimate Std. Error   t value     Pr(>|t|)
## sprayA 14.500000   1.132156 12.807428 1.470512e-19
## sprayB 15.333333   1.132156 13.543487 1.001994e-20
## sprayC  2.083333   1.132156  1.840148 7.024334e-02
## sprayD  4.916667   1.132156  4.342749 4.953047e-05
## sprayE  3.500000   1.132156  3.091448 2.916794e-03
## sprayF 16.666667   1.132156 14.721181 1.573471e-22
```

Observe que sprayA ahora aparece explícitamente en la lista de estimaciones. Observe también cómo han cambiado los valores de las columnas. Las medias de todos los grupos ahora se muestran explícitamente en la columna Estimación. Recuerde que anteriormente, con una intersección, se excluyó la pulverización A, su media fue la intersección y los valores de las otras pulverizaciones (estimaciones, errores estándar y pruebas t) se calcularon en relación con la pulverización A, el grupo de referencia. Omitir la intersección claramente afectó al modelo.

Claramente, qué nivel es el primero es importante para el modelo. Si desea un grupo de referencia diferente, por ejemplo, para comparar sprayB con sprayC, puede reajustar el modelo con un grupo de referencia diferente.

La función R relevel hace precisamente esto. Reordena los niveles de un factor. Haremos esto ahora. Llamaremos relevel con dos argumentos. El primero es el factor, en este caso InsectSprays $ spray, y el segundo es el nivel que queremos que sea el primero, en este caso "C". Almacene el resultado en una nueva variable spray2.

```
spray2 <- relevel(InsectSprays$spray,"C")
```

ahora creando el modelo:

```
fit2 <- lm(count ~ spray2, InsectSprays)
summary(fit2)$coef
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  2.083333   1.132156 1.840148 7.024334e-02
## spray2A     12.416667   1.601110 7.755038 7.266893e-11
## spray2B     13.250000   1.601110 8.275511 8.509776e-12
## spray2D      2.833333   1.601110 1.769606 8.141205e-02
## spray2E      1.416667   1.601110 0.884803 3.794750e-01
## spray2F     14.583333   1.601110 9.108266 2.794343e-13
```

Recuerde que con este modelo sprayC es el grupo de referencia, por lo que las estadísticas de la prueba t (que se muestran en la columna 3 de los coeficientes de resumen) comparan las otras aplicaciones con sprayC. Estos se pueden calcular a mano utilizando las estimaciones y el error estándar del modelo original (ajuste) que utilizó sprayA como referencia.

Las diapositivas muestran los detalles de esto, pero aquí lo demostraremos calculando el valor t de spray2B. Reste el coeficiente sprayC de ajuste (ajuste$coef [3]) de sprayB (ajuste $ coef [2]) y divida por el error estándar que vimos fue 1,6011. El resultado es el valor t de spray2B. Hacer esto ahora.

```
(fit$coef[2]-fit$coef[3])/1.6011
```

```
##   sprayB
## 8.275561
```

Pasamos por alto algunos detalles en esta lección. Por ejemplo, los recuentos nunca pueden ser 0, por lo que se viola el supuesto de normalidad. Exploraremos más este tema cuando analicemos los GLM de Poisson. Por ahora, alégrate de haber concluido esta segunda lección sobre modelos lineales multivariables.

#multivariable ejemplo 3

Esta es la tercera y última lección en la que veremos modelos de regresión con más de una variable independiente o predictor. Comenzaremos con los datos de la OMS sobre el hambre que nos hemos tomado la libertad de cargar para usted. La OMS es la Organización Mundial de la Salud y estos datos se refieren

a niños pequeños de todo el mundo y las tasas de hambre entre ellos que la organización recopiló durante varios años. El archivo csv original era muy grande y hemos subconjunto solo las filas que identifican el género del niño como hombre o mujer. Hemos leído los datos en el marco de datos "hambre" por usted, para que pueda acceder a ellos fácilmente.

Como hicimos en la última lección, primero intentemos comprender mejor el conjunto de datos. Utilice la función R tenue para encontrar las dimensiones del hambre.

```
load("C:/Users/luism/Documents/hung.RData")
dim(hunger)
```

```
## [1] 948  13
```

```
names(hunger)
```

```
##  [1] "X"              "Indicator"      "Data.Source"    "PUBLISH.STATES"
##  [5] "Year"           "WHO.region"     "Country"        "Sex"
##  [9] "Display.Value"  "Numeric"        "Low"            "High"
## [13] "Comments"
```

```
head(hunger)
```

```
##    X                             Indicator Data.Source PUBLISH.STATES Year
## 1  8 Children aged <5 years underweight (%) NLIS_310044      Published 1986
## 2 11 Children aged <5 years underweight (%) NLIS_310095      Published 1989
## 3 13 Children aged <5 years underweight (%) NLIS_310138      Published 1988
## 4 16 Children aged <5 years underweight (%) NLIS_310044      Published 1986
## 5 18 Children aged <5 years underweight (%) NLIS_310095      Published 1989
## 6 21 Children aged <5 years underweight (%) NLIS_310138      Published 1988
##   WHO.region  Country    Sex Display.Value Numeric Low High Comments
## 1     Africa  Senegal   Male          19.3    19.3  NA   NA       NA
## 2     Africa   Uganda Female          19.1    19.1  NA   NA       NA
## 3     Africa Zimbabwe Female           7.2     7.2  NA   NA       NA
## 4     Africa  Senegal Female          15.3    15.3  NA   NA       NA
## 5     Africa   Uganda   Male          20.4    20.4  NA   NA       NA
## 6     Africa Zimbabwe   Male           8.7     8.7  NA   NA       NA
```

La columna Numérica de una fila en particular nos dice el porcentaje de niños menores de 5 años que tenían bajo peso cuando se tomó esa muestra. Esta es una de las columnas en las que nos centraremos en esta lección. Será el resultado (variable dependiente) de los modelos que generemos.

Primero veamos la tasa de hambre y veamos cómo ha cambiado con el tiempo. Utilice la función R lm para generar el modelo lineal en el que la tasa de hambre, numérica, depende del predictor, año. Ponga el resultado en el ajuste de la variable.

```
fit <- lm(hunger$Numeric ~ hunger$Year)
summary(fit)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 634.479660 121.1445995  5.237375 2.007699e-07
## hunger$Year  -0.308397   0.0605292 -5.095012 4.209412e-07
```

Ahora usemos la capacidad de subconjunto de R para observar las tasas de hambre de los diferentes géneros para ver cómo, o incluso si, difieren. Una vez más, use la función R lm para generar el modelo lineal en el que la tasa de hambre (numérica) para las niñas y niños depende del año. Pon el resultado en la variable lmF.

```
 lmF <- lm(Numeric[Sex=="Female"] ~ Year[Sex=="Female"],hunger)
lmM <- lm(Numeric[Sex=="Male"] ~ Year[Sex=="Male"],hunger)
```

Ahora trazaremos los puntos de datos y las líneas ajustadas usando diferentes colores para distinguir entre machos (azul) y hembras (rosa).

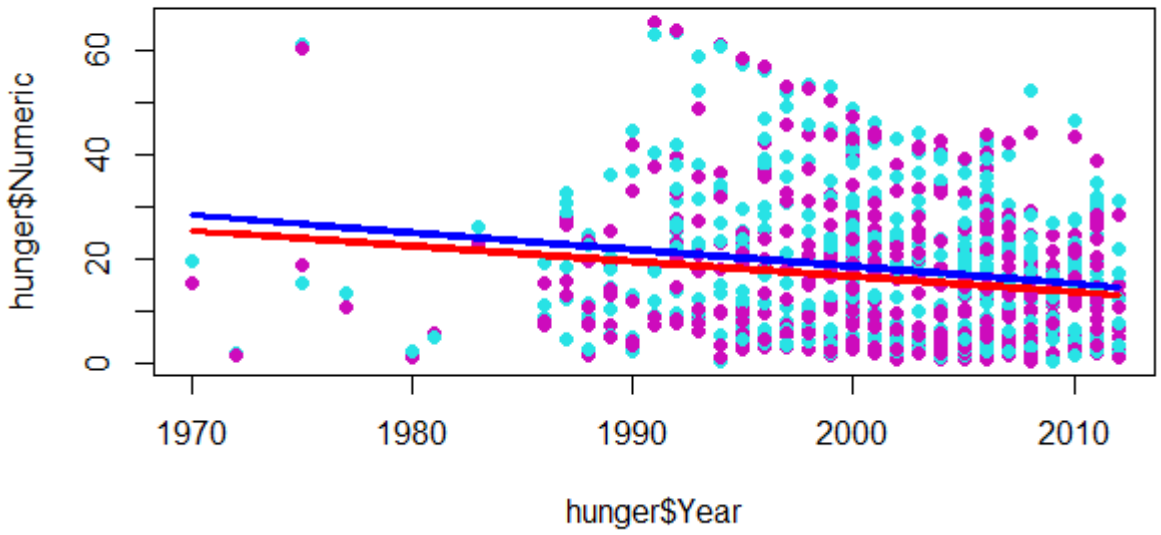


Figure 1: A caption

Podemos ver en la gráfica que las líneas no son exactamente paralelas. En el lado derecho del gráfico (alrededor del año 2010) están más juntos que en el lado izquierdo (alrededor de 1970). Ahora, en lugar de separar los datos subconjuntos de las muestras por género, usaremos el género como otro predictor para crear el modelo lineal lmBoth. Recuerde que para hacer esto en R colocamos un signo más "+" entre las variables independientes, por lo que la fórmula parece dependiente ~ independiente1 + independiente2.

Cree lmBoth ahora. Numérico es el dependiente, Año y Sexo son las variables independientes. Los datos son "hambre". Para lmBoth, asegúrese de que el año sea el primero y el sexo el segundo.

```
lmBoth <- lm(Numeric ~ Year+Sex,hunger)
summary(lmBoth)
```

```
##
## Call:
## lm(formula = Numeric ~ Year + Sex, data = hunger)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -25.472 -11.297  -1.848   7.058  45.990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 633.5283   120.8950   5.240 1.98e-07 ***
## Year         -0.3084     0.0604  -5.106 3.99e-07 ***
## SexMale       1.9027     0.8576   2.219   0.0267 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.2 on 945 degrees of freedom
## Multiple R-squared:  0.03175,    Adjusted R-squared:  0.0297
## F-statistic: 15.49 on 2 and 945 DF,  p-value: 2.392e-07
```

Ahora volveremos a trazar los puntos de datos junto con dos nuevas líneas usando diferentes colores. La línea roja tendrá la intersección femenina y la línea azul tendrá la intersección masculina, y la penddiente sera el valor de year en el modelo

Ahora consideraremos la interacción entre el año y el género para ver cómo afecta eso a los cambios en las tasas de hambre. Para hacer esto, agregaremos un tercer término a la parte del predictor de nuestra fórmula modelo, el producto del año y el género.

Cree el modelo lmInter. Numérico es el resultado y los tres predictores son Año, Sexo y Sexo * Año. Los datos son "hambre".

```
lmInter <- lm(Numeric ~ Year + Sex + Year*Sex, hunger)
summary(lmInter)
```

```
##
## Call:
## lm(formula = Numeric ~ Year + Sex + Year * Sex, data = hunger)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.913 -11.248  -1.853   7.087  46.146
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  603.50580  171.05519   3.528 0.000439 ***
## Year          -0.29340    0.08547  -3.433 0.000623 ***
## SexMale       61.94772  241.90858   0.256 0.797946
## Year:SexMale  -0.03000    0.12087  -0.248 0.804022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.21 on 944 degrees of freedom
## Multiple R-squared:  0.03181,    Adjusted R-squared:  0.02874
## F-statistic: 10.34 on 3 and 944 DF,  p-value: 1.064e-06
```

La estimación asociada con Year: SexMale representa la distancia entre el cambio anual en porcentaje de hombres y el de mujeres.

Finalmente, observamos que las cosas son un poco más complicadas cuando se trata de una interacción entre predictores que son continuos (y no factores). Las diapositivas muestran el álgebra subyacente, pero podemos resumir.

Suponga que tenemos dos predictores que interactúan y uno de ellos se mantiene constante. El cambio esperado en el resultado de un cambio de unidad en el otro predictor es el coeficiente de ese predictor cambiante + el coeficiente de la interacción * el valor del predictor mantenido constante.

Suponga que el modelo lineal es Hi = b0 + (b1 * Ii) + (b2 * Yi) + (b3 * Ii *Yi) + ei. Aquí, las H representan los resultados, las I y las Y los predictores, ninguno de los cuales es una categoría, y las b representan los coeficientes estimados de los predictores. Podemos ignorar las e que representan los residuos del modelo. Esta ecuación modela una interacción continua ya que ni I ni Y son una categoría o factor. Supongamos que fijamos I en algún valor y dejamos que Y varíe.

¿Qué expresión representa el cambio en H por cambio unitario en Y dado que I está fijo en 5?R=b2+b3*5

Diagnóstico y variación de residuos

En la figura adjunta hay un valor atípico bastante obvio. Por obvio que sea, no afecta mucho al ajuste, como se puede ver al comparar la línea naranja con la negra. La línea naranja representa un ajuste en el que se incluye el valor atípico en el conjunto de datos, y la línea negra representa un ajuste en el que se excluye el valor atípico. La inclusión de este valor atípico no cambia mucho el ajuste, por lo que se dice que carece de influencia.

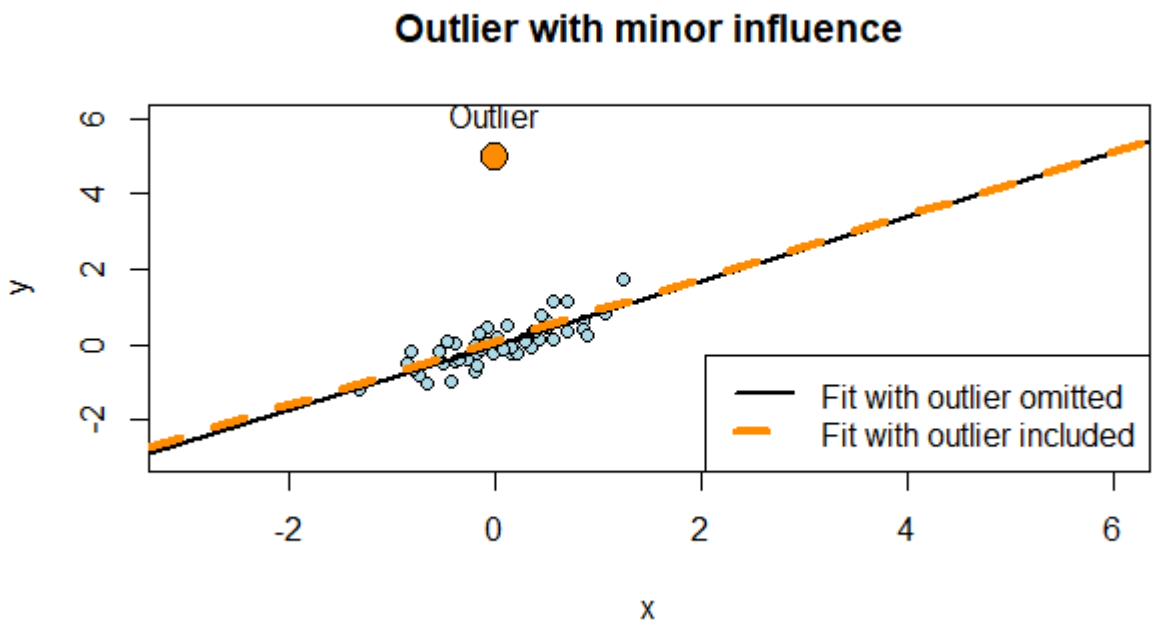


Figure 2: A caption

Esta siguiente cifra también tiene un valor atípico bastante obvio, pero en este caso, incluir el valor atípico cambia mucho el ajuste. La pendiente y los residuos de la línea naranja son muy diferentes a los de la línea negra. Se dice que este valor atípico es influyente.

Los valores atípicos pueden o no pertenecer a los datos. Pueden representar eventos reales o pueden ser falsos. En cualquier caso, deben examinarse. Para detectarlos, R proporciona varios gráficos de diagnóstico y medidas de influencia. En esta lección ilustraremos su significado y uso. La técnica básica es examinar los efectos de omitir una muestra, como hicimos al comparar las líneas negras y naranjas de arriba. Usaremos el valor atípico influyente para ilustrar, ya que omitirlo tiene efectos claros.

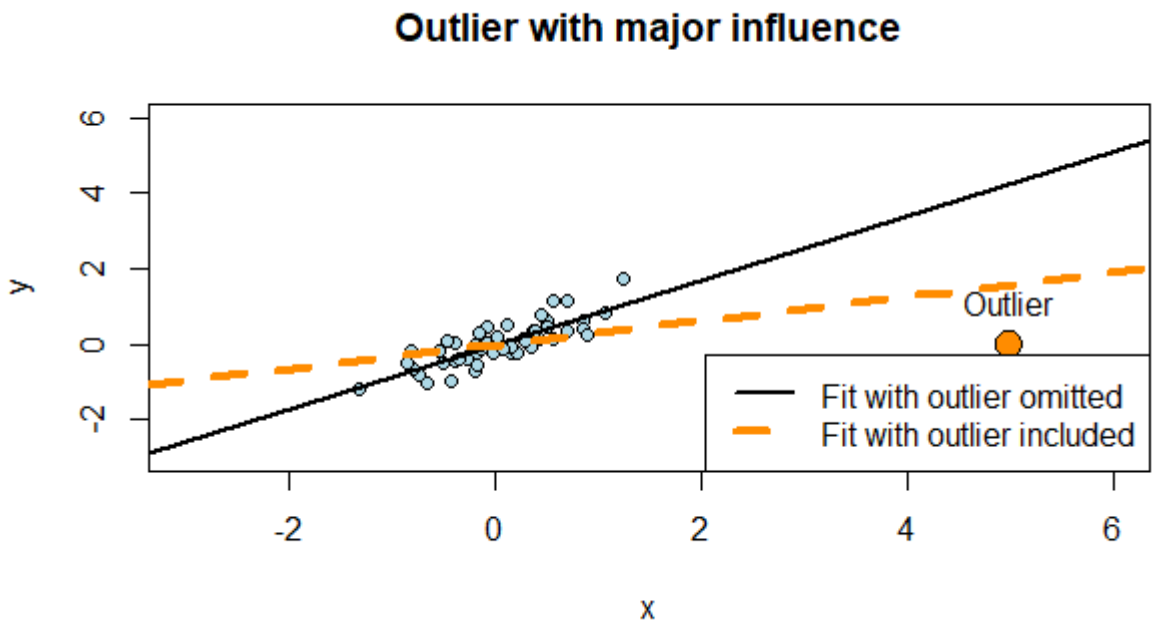


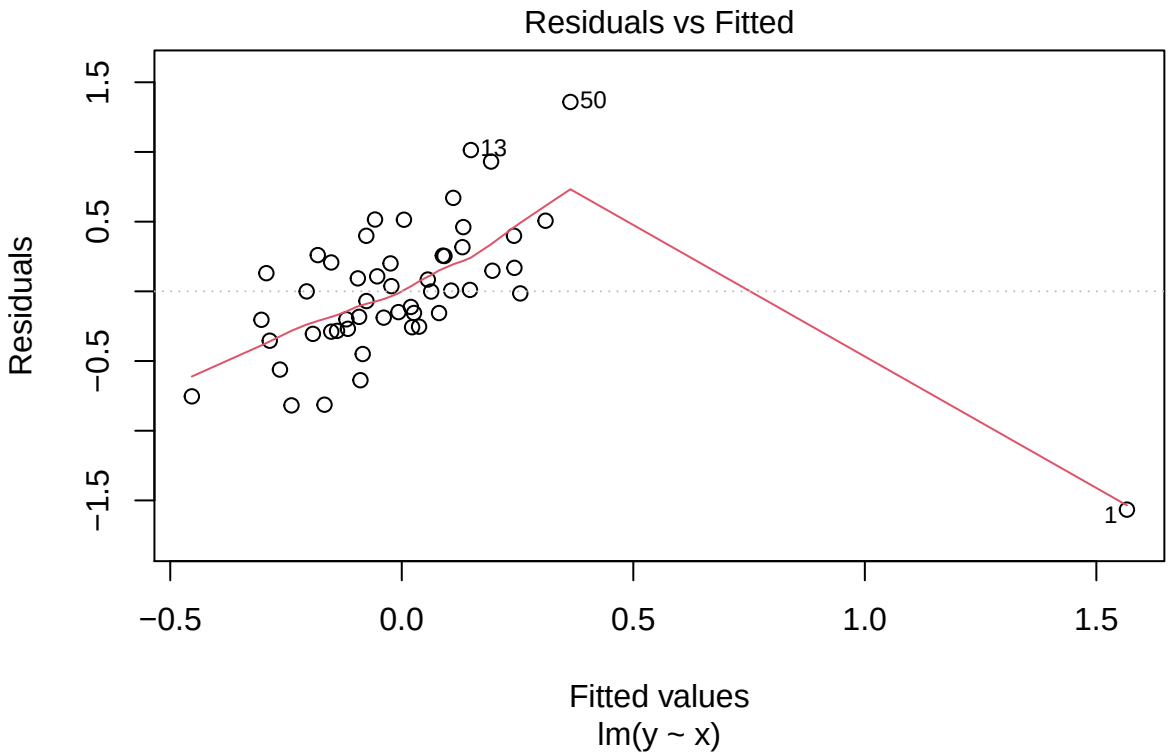
Figure 3: A caption

El valor atípico influyente está en un marco de datos denominado out2. Tiene dos columnas, etiquetadas y y x, respectivamente. Para comenzar, cree un modelo llamado ajuste usando ajuste <- lm (y ~ x, out2) o una expresión equivalente.

```
load("C:/Users/luism/Documents/res.RData")
fit <- lm(y ~ x, out2)
```

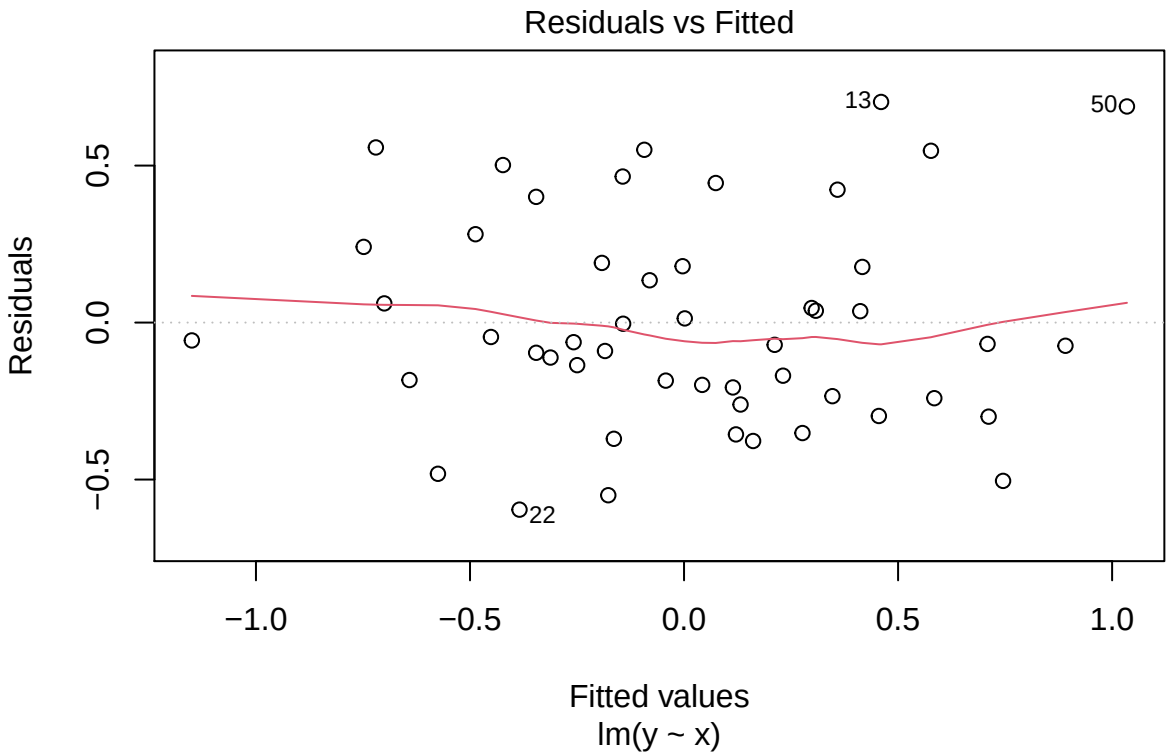
El gráfico de diagnóstico más simple muestra los valores residuales frente a los valores ajustados. Los residuos deben no estar correlacionados con el ajuste, ser independientes y (casi) idénticamente distribuidos con media cero. Ingrese plot (fit, which = 1) en el indicador R para ver si este es el caso.

```
plot (fit, which = 1)
```



Nuestro valor atípico influyente está en la fila 1 de los datos. Para excluirlo, es solo cuestión de usar out2 [-1,] en lugar de out2 como datos. Cree un segundo modelo, llamado fitno para 'ajustar sin valor atípico', que excluye el valor atípico.

```
fitno <- lm(y ~ x, out2[-1, ])
plot(fitno, which=1)
```



Esta trama no tiene la apariencia estampada de la primera. Se ve como cabría esperar si los residuos estuvieran distribuidos de forma independiente y (casi) idéntica con media cero, y no estuvieran correlacionados con el ajuste.

El cambio que la inclusión o exclusión de una muestra induce en los coeficientes es una simple medida de su influencia. Reste coef (fitno) de coef (ajuste) para ver el cambio inducido al incluir la primera muestra influyente.

```
coef(fit)-coef(fitno)
```

```
## (Intercept)           x
## -0.01167866 -0.53363019
```

dfbeta

la función, dfbeta, realiza el cálculo equivalente para cada muestra de los datos. La primera fila de dfbeta (ajuste) debe coincidir con la diferencia que acabamos de calcular. La segunda fila es un cálculo similar para la segunda muestra, y así sucesivamente. Dado que dfbeta devuelve una matriz grande, use head (dfbeta (fit)) o View (dfbeta (fit)) para examinar el resultado.

```
head(dfbeta(fit))
```

```
##    (Intercept)             x
## 1 -0.011678662 -0.5336301857
## 2  0.008636967  0.0045759242
```

```
## 3  0.010323864 -0.0003509441
## 4  0.003122096 -0.0033664451
## 5  0.001975966 -0.0008297575
## 6  0.002230518 -0.0005867041
```

Al comparar la primera fila con las que están debajo de ella, vemos que la primera muestra tiene un efecto mucho mayor en la pendiente (la columna x) que otras muestras. De hecho, la magnitud de su efecto es aproximadamente 100 veces mayor que la de cualquier otro punto. Su efecto en la intersección no es muy distintivo esencialmente porque su coordenada y es 0, la media de las otras muestras.

Cuando se incluye una muestra en un modelo, acerca la línea de regresión a sí misma (línea naranja) que la del modelo que la excluye (línea negra). Su residuo, la diferencia entre su valor y real y el de una línea de regresión. , es por lo tanto menor en magnitud cuando se incluye (puntos naranjas) que cuando se omite (puntos negros). La relación de estos dos residuos, naranja a negro, es por lo tanto pequeña en magnitud para una muestra influyente. Para una muestra que no es influyente, la razón sería cercana a 1. Por lo tanto, 1 menos la razón es una medida de influencia, cerca de 0 para los puntos que no son influyentes y cerca de 1 para los puntos que sí lo son.

Esta medida a veces se llama influencia, a veces apalancamiento y, a veces, valor de sombrero. Dado que es 1 menos la razón de dos residuos, para calcularlo desde cero primero debemos obtener los dos residuos. El numerador de la razón (puntos naranjas) es el residuo de la primera muestra del modelo que llamamos ajuste. El modelo fitno, que excluye esta muestra, también excluye su residual, por lo que tendremos que calcular su valor. Esto se hace fácilmente. Usamos la función de predicción de R para calcular el valor predicho de fitno de y y restarlo del valor real. Utilice la expresión resno <- out2 [1, “y”] - predict (fitno, out2 [1,]) para hacer el cálculo.

```
resno <- out2[1, "y"] - predict(fitno, out2[1,])
1-resid(fit)[1]/resno
```

```
##         1
## 0.6311547
```

hatvalues

la función hatvalues realiza para cada muestra un cálculo equivalente al que acaba de realizar. Por lo tanto, la primera entrada de hatvalues (ajuste) debe coincidir con el valor que acaba de calcular. Dado que hay bastantes muestras, use head (hatvalues (ajuste)) o View (hatvalues (ajuste)) para comparar la medida de influencia de nuestro valor atípico con la de algunas otras muestras.

```
head(hatvalues(fit))
```

```
##          1          2          3          4          5          6
## 0.63115474 0.02324999 0.01962520 0.04326099 0.02255531 0.02071441
```

Los residuos de muestras individuales a veces se tratan como si tuvieran la misma varianza, que se estima como la varianza muestral de todo el conjunto de residuos. Sin embargo, teóricamente, los residuos de muestras individuales tienen diferentes variaciones y estas diferencias pueden volverse grandes en presencia de valores atípicos. Los residuos estandarizados y estudentizados intentan compensar este efecto de dos formas ligeramente diferentes. Ambos usan valores de sombrero. Primero consideraremos los residuos estandarizados. Para comenzar, calcule la desviación estándar muestral del ajuste residual dividiendo la desviación del ajuste, es decir, su suma de cuadrados residual, por los grados de libertad residuales y tomando la raíz cuadrada. Almacene el resultado en una variable llamada sigma.

```
sigma <- sqrt(deviance(fit)/df.residual(fit))
```

Por lo general, dividiríamos el residuo de ajuste (que tiene una media de 0) por sigma. En el caso presente, multiplicamos sigma por sqrt (1-hatvalues (ajuste)) para estimar las desviaciones estándar de muestras individuales. Por lo tanto, en lugar de dividir resid (ajuste) por sigma, dividimos por sigma * sqrt (1-hatvalues (ajuste)). El resultado se denomina residual estandarizado. Calcule el residuo estandarizado del ajuste y almacénelo en una variable llamada rstd.

```
rstd <- resid(fit)/(sigma * sqrt(1-hatvalues(fit)))
```

rstandard

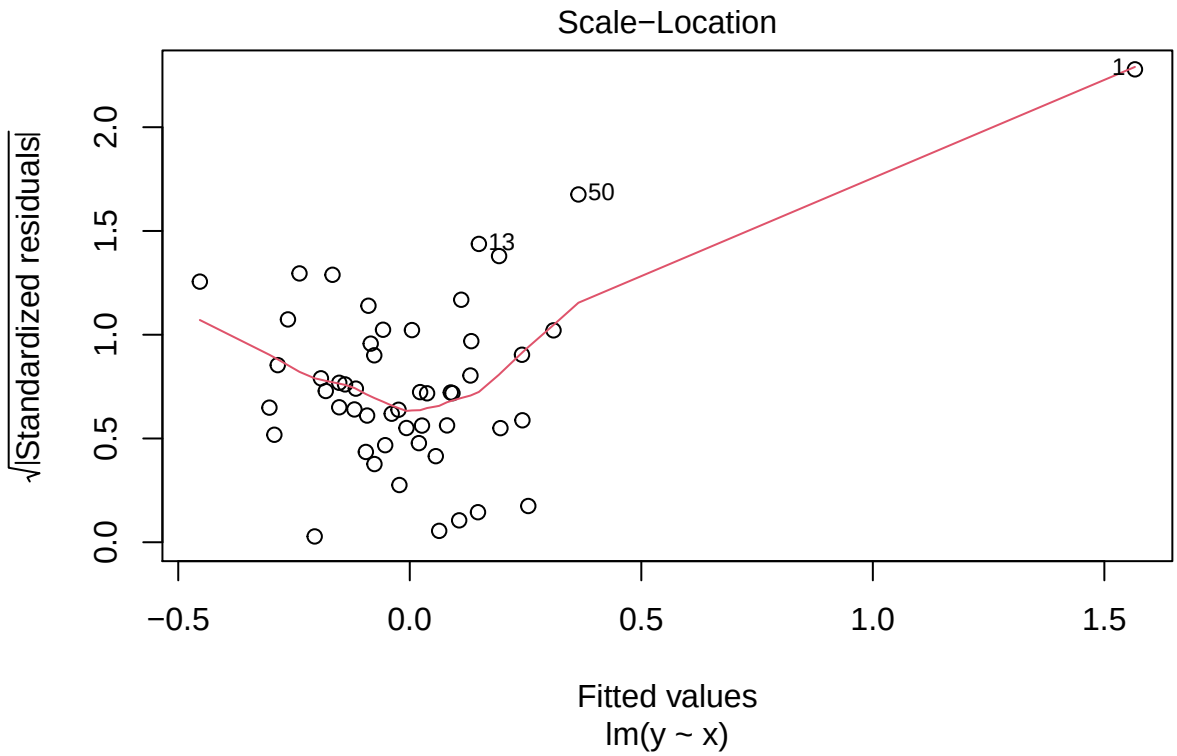
La función, rstandard, calcula el residuo estandarizado que acaba de calcular paso a paso. Utilice head (cbind (rstd, rstandard (ajuste))) o View (cbind (rstd, rstandard (ajuste))) para comparar los dos cálculos.

```
head(cbind(rstd, rstandard(fit)))
```

```
##         rstd
## 1 -5.1928156 -5.1928156
## 2  0.9389601  0.9389601
## 3  1.0450409  1.0450409
## 4  0.2682743  0.2682743
## 5  0.1893339  0.1893339
## 6  0.2186961  0.2186961
```

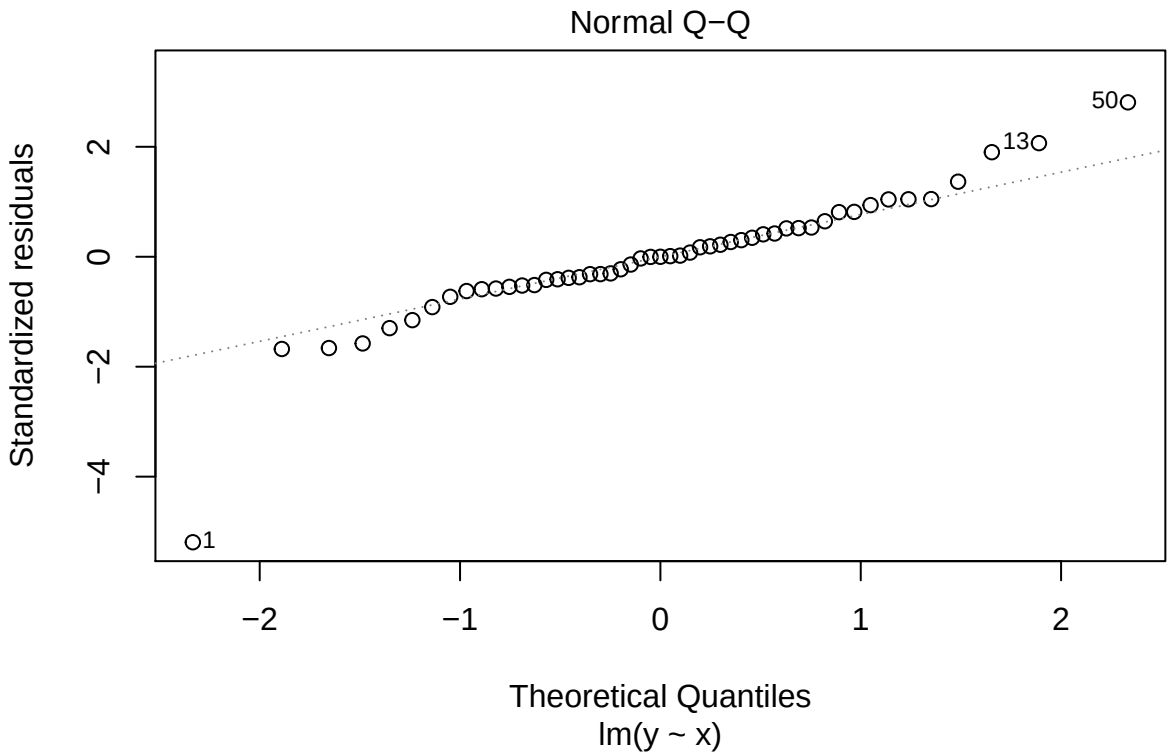
Una gráfica de ubicación de escala muestra la raíz cuadrada de los residuos estandarizados frente a los valores ajustados. Utilice plot (fit, which = 3) para mostrarlo.

```
plot(fit, which=3)
```



La mayoría de las estadísticas de diagnóstico en discusión se desarrollaron debido a las deficiencias percibidas de otros diagnósticos y porque se podían caracterizar sus distribuciones bajo una hipótesis nula. La suposición de que los residuos son aproximadamente normales está implícita en tales caracterizaciones. Dado que los residuos estandarizados se ajustan a las varianzas residuales individuales, es de interés un gráfico QQ de los residuos estandarizados contra lo normal con varianza constante. Utilice plot (fit, which = 2) para mostrar este gráfico de diagnóstico.

```
plot(fit, which=2)
```



observe el residuo estandarizado del valor atípico, etiquetado como 1 en el gráfico de QQ normal. Aproximadamente, ¿cuántas desviaciones estándar de la media es?R=5

Los residuales estudentizados, (a veces llamados residuales estudentizados externamente) estiman las desviaciones estándar de los residuos individuales utilizando, además de los valores de sombrero individuales, la desviación de un modelo que deja fuera la muestra asociada. Ilustraremos usando el valor atípico. Recordando que el modelo que llamamos fitno omite la muestra atípica, calcule la desviación estándar muestral del residuo de fitno dividiendo su desviación por sus grados de libertad residuales y tomando la raíz cuadrada. Almacene el resultado en una variable llamada sigma1.

```
sigma1 <- sqrt(deviance(fitno)/df.residual(fitno))
```

Calcule el residuo estudentizado para la muestra atípica dividiendo resid (ajuste) [1] por el producto de sigma1 y sqrt (1-hatvalues (ajuste) [1]). No es necesario almacenar esto en una variable.

```
resid(fit)[1]/(sigma1*sqrt(1-hatvalues(fit)[1]))
```

```
##         1
## -7.664261
```

rstudent

La función rstudent, calcula los residuales estudentizados para cada muestra usando un procedimiento equivalente al que acabamos de usar para el valor atípico. Por lo tanto, rstudent (fit) [1] debe coincidir con el valor

que calculamos en la pregunta anterior. Utilice head (rstudent (fit)) o View (rstudent (fit)) para verificar esto y comparar el residuo estudentizado del valor atípico con los de otras muestras.

La distancia de Cook es la última medida de influencia que consideraremos. Es esencialmente la suma de las diferencias al cuadrado entre los valores ajustados con y sin una muestra particular. Se normaliza (se divide por) la varianza de la muestra residual multiplicada por el número de predictores, que en nuestro caso es 2 (la intersección y x). Básicamente, indica cuánto cambia un modelo una muestra determinada. Ilustraremos una vez más calculando la distancia de Cook para el valor atípico.

Comenzaremos calculando la diferencia en los valores predichos entre fit y fitno, los modelos que incluyen y omiten respectivamente el valor atípico. Esto se hace más fácilmente restando predict (fit, out2) de predict (fitno, out2). Almacene la diferencia en una variable llamada dy.

```
dy <- predict(fitno, out2)-predict(fit, out2)
```

Recuerde que calculamos antes la desviación estándar muestral del ajuste residual, sigma. Divida los cuadrados sumados de dy por 2 * sigma ^ 2 para calcular la distancia de Cook del valor atípico. No es necesario almacenar el resultado en una variable.

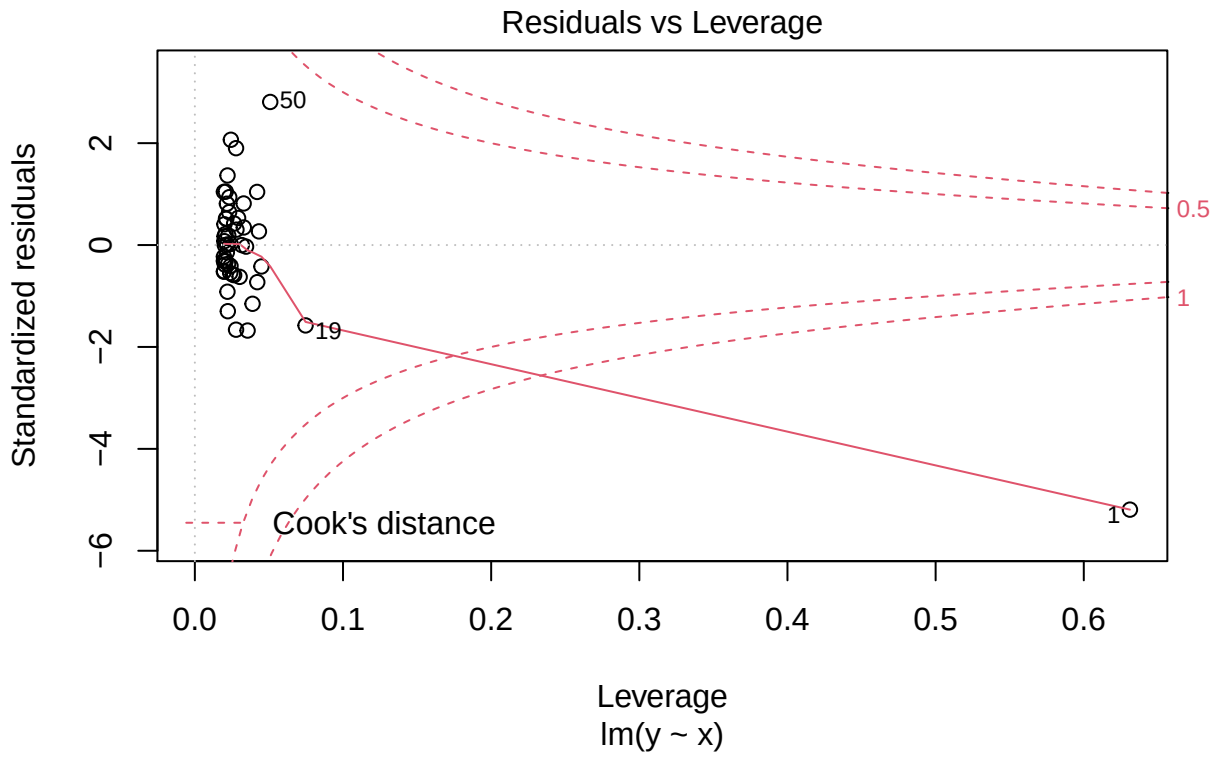
```
sum(dy^2)/(2*sigma^2)
```

```
## [1] 23.07105
```

cooks.distance:

La función cooks.distance calculará la distancia de Cook para cada muestra. En lugar de verificar que cooks.distance (fit) [1] es igual al valor que se acaba de calcular, porque ese tipo de cosas debe estar volviéndose tedioso a estas alturas, muestre una gráfica de diagnóstico que use la distancia de Cook usando plot (fit, which = 5) .

```
plot(fit, which=5)
```




```
influence.measures(fit)
```

```
## Influence measures of
##   lm(formula = y ~ x, data = out2) :
##
##       dfb.1_     dfb.x     dffit cov.r   cook.d    hat inf
## 1  -0.245002 -9.87e+00 -1.00e+01 0.571 2.31e+01 0.6312   *
## 2   0.122613  5.73e-02  1.45e-01 1.029 1.05e-02 0.0232
## 3   0.146882 -4.40e-03  1.48e-01 1.016 1.09e-02 0.0196
## 4   0.043954 -4.18e-02  5.65e-02 1.086 1.63e-03 0.0433
## 5   0.027808 -1.03e-02  2.85e-02 1.065 4.14e-04 0.0226
## 6   0.031394 -7.28e-03  3.15e-02 1.062 5.06e-04 0.0207
## 7   0.023243  3.75e-03  2.44e-02 1.062 3.03e-04 0.0201
## 8   0.099954  9.55e-02  1.50e-01 1.048 1.13e-02 0.0329
## 9  -0.061050  2.74e-02 -6.37e-02 1.060 2.07e-03 0.0241
## 10 -0.193443  6.81e-02 -1.97e-01 0.994 1.92e-02 0.0223
## 11  0.064203 -3.54e-02  6.92e-02 1.063 2.44e-03 0.0265
## 12 -0.054696  1.01e-02 -5.47e-02 1.057 1.52e-03 0.0203
## 13  0.277327  1.49e-01  3.38e-01 0.890 5.32e-02 0.0243
## 14  0.069284  1.99e-02  7.58e-02 1.053 2.92e-03 0.0211
## 15 -0.188287  1.64e-01 -2.33e-01 1.026 2.69e-02 0.0389
## 16 -0.042275 -1.08e-02 -4.57e-02 1.060 1.06e-03 0.0208
## 17  0.152566 -3.78e-02  1.53e-01 1.017 1.17e-02 0.0209
## 18  0.002686  1.42e-03  3.26e-03 1.068 5.42e-06 0.0242
## 19 -0.290023  3.91e-01 -4.55e-01 1.015 1.00e-01 0.0747
## 20 -0.097028  6.49e-02 -1.10e-01 1.058 6.07e-03 0.0302
```

```
## 21 -0.000403 -7.63e-05 -4.27e-04 1.064 9.30e-08 0.0203
## 22 -0.261184  1.55e-01 -2.86e-01 0.955 3.95e-02 0.0278
## 23  0.037784  2.80e-02  5.09e-02 1.068 1.32e-03 0.0281
## 24 -0.089793  4.95e-02 -9.68e-02 1.055 4.75e-03 0.0265
## 25 -0.274595  2.20e-01 -3.29e-01 0.960 5.20e-02 0.0356
## 26 -0.043446 -2.13e-03 -4.43e-02 1.059 1.00e-03 0.0197
## 27  0.247170  1.80e-01  3.31e-01 0.920 5.17e-02 0.0278
## 28  0.083988  3.86e-02  9.88e-02 1.049 4.94e-03 0.0231
## 29  0.001459  5.25e-04  1.64e-03 1.065 1.38e-06 0.0218
## 30  0.042019  4.03e-02  6.33e-02 1.072 2.04e-03 0.0331
## 31 -0.054719  1.98e-02 -5.59e-02 1.060 1.59e-03 0.0224
## 32 -0.072216 -2.48e-03 -7.35e-02 1.051 2.74e-03 0.0196
## 33 -0.020634  6.39e-03 -2.09e-02 1.064 2.23e-04 0.0216
## 34 -0.119435  1.11e-01 -1.52e-01 1.065 1.17e-02 0.0421
## 35 -0.070484 -6.34e-03 -7.26e-02 1.052 2.67e-03 0.0198
## 36 -0.003679 -3.76e-03 -5.71e-03 1.079 1.66e-05 0.0346
## 37  0.123255  1.60e-01  2.19e-01 1.040 2.38e-02 0.0420
## 38  0.082042 -5.24e-02  9.14e-02 1.061 4.24e-03 0.0292
## 39  0.118728 -3.69e-02  1.20e-01 1.037 7.28e-03 0.0216
## 40 -0.000122  8.63e-05 -1.40e-04 1.076 9.97e-09 0.0317
## 41  0.182368  6.89e-02  2.07e-01 0.986 2.10e-02 0.0221
## 42 -0.069428  6.80e-02 -9.05e-02 1.083 4.17e-03 0.0449
## 43 -0.031452 -7.72e-04 -3.20e-02 1.061 5.21e-04 0.0196
## 44  0.010723 -1.35e-03  1.07e-02 1.063 5.87e-05 0.0199
## 45 -0.042426  3.08e-03 -4.26e-02 1.059 9.23e-04 0.0197
## 46 -0.134848  4.53e-02 -1.37e-01 1.029 9.44e-03 0.0220
## 47  0.057656 -7.66e-03  5.77e-02 1.056 1.69e-03 0.0200
## 48 -0.081663  3.59e-02 -8.50e-02 1.055 3.67e-03 0.0239
## 49 -0.087358  4.48e-02 -9.30e-02 1.055 4.38e-03 0.0255
## 50  0.347285  5.51e-01  7.03e-01 0.773 2.11e-01 0.0509   *
## 51  0.068476  2.06e-02  7.53e-02 1.053 2.88e-03 0.0212
```