

Tarea 3 - Alejandro Hernández Farías

Parte teórica

Supongamos que queremos explicar una variable estadística Y , utilizando la información de p variables X^1, \dots, X^p . Si se toma una muestra de N individuos, cada variable está representada por un vector de tamaño N . La información de las variables explicativas se pueden juntar en una matriz $X = [X^1 | \dots | X^p]$ de tamaño $n \times p$ donde cada columna es una variable y cada fila uno de los individuos de la muestra.

-Plantear el problema de regresión como un problema de mínimos cuadrados. Encontrar el vector $\hat{\beta} = [\hat{\beta}_1, \dots, \hat{\beta}_p]^T$ que resuelva:

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2$$

Queremos minimizar $f(\beta) = \|Y - X\beta\|^2$, entonces:

$$\begin{aligned} \|Y - X\beta\|^2 &= (Y - X\beta)^T(Y - X\beta) = Y^T Y - Y^T X\beta - (X\beta)^T Y + (X\beta)^T(X\beta) \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta \end{aligned}$$

Derivamos con respecto a β :

$$\frac{\partial}{\partial \beta} f(\beta) = -2X^T Y + 2X^T X\beta$$

Igualamos a cero y despejamos:

$$\begin{aligned} -2X^T Y + 2X^T X\beta &= 0 \\ X^T Y &= X^T X\beta \end{aligned}$$

Por lo tanto:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

-¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos?

Por definición del modelo de regresión la variable dependiente es una combinación lineal de las variables independientes o explicativas y un término de error (i.e. $Y = X\beta + \epsilon$), donde los parámetros a estimar β son lineales, por lo que nuestra estimación va a ser un ajuste lineal de los datos.

Otro enfoque sería verlo como una composición de funciones lineales por lo que el ajuste a los datos es lineal.

También lo podemos pensar como que β_i es una combinación lineal de las covarianzas entre X_i y Y_i ponderada por los elementos de Σ^{-1} (i.e. $(X^T X)^{-1} X^T Y = \Sigma^{-1} \operatorname{Cov}(X, Y)$).

-¿Podríamos usarlo para ajustar polinomios (ej. $y = x^2$)?

Sí, siempre y cuando los parámetros sean lineales, la variable dependiente no necesariamente tiene que ser lineal.

-Argumentar la relación entre la solución encontrada y un problema de proyección en subespacios vectoriales de álgebra lineal. ¿Cuál es la relación particular con el teorema de Pitágoras?

Se relacionan pues el problema de regresión es a final de cuentas un problema de mínimos cuadrados, en el cual buscamos minimizar $\|Y - X\beta\|^2$, donde \hat{Y} es una proyección del vector Y de tal manera que $\hat{Y} = X(X^T X)^{-1} X^T Y$ y $\hat{\beta} = (X^T X)^{-1} X^T Y$. Es importante destacar que para que $\|Y - X\beta\|^2$ sea mínimo, $\|Y - X\beta\|$ tiene que ser perpendicular al espacio vectorial generado por las columnas de X .

Este último punto lo relaciona con el Teorema de Pitágoras, pues la proyección es la mínima distancia entre Y y \hat{Y} , lo que geométricamente se puede ver como la altura del triángulo rectángulo para que la distancia entre Y y \hat{Y} sea la mínima.

-¿Qué logramos al agregar una columna de unos en la matriz X ? Es decir, definir mejor $X = [1_n | X^1 | \dots | X^p]$ con $1_n = [1, 1, \dots, 1]^T$.

Esa columna sirve para que en la representación matricial de la regresión, consideremos el parámetro β_0 . El contar con el parámetro β_0 nos permite que nuestra regresión lineal esté desplazada del centro de gravedad (u origen).

Plantear el problema de regresión ahora como un problema de estadística

$$Y_i = \beta_0 + \beta_1 X_i^1 + \dots + \beta_p X_i^p + \epsilon_i$$

donde los errores son no correlacionados con distribución

$$\epsilon_i \sim N(0, \sigma^2)$$

-¿Cuál es la función de verosimilitud del problema anterior?

Si $\epsilon \sim N(0, \sigma^2 I_n)$, entonces $\epsilon + X\beta \sim N(X\beta, \sigma^2 I_n)$ y por lo tanto $Y \sim N(X\beta, \sigma^2 I_n)$, pues al sumar un valor a todos los datos se modifica la media en esa proporción, pero la varianza no cambia.

La función de verosimilitud es la siguiente:

$$L(\beta, \sigma^2) = f(Y|\beta, \sigma^2, X) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}$$

-Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados.

Consideremos $L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}$, entonces:

$$\begin{aligned} \log L(\beta, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} [Y^T Y - Y^T X\beta - (X\beta)^T Y + (X\beta)^T (X\beta)] \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} Y^T Y + \frac{1}{2\sigma^2} 2\beta^T X^T Y - \frac{1}{2\sigma^2} \beta^T X^T X\beta \end{aligned}$$

Derivamos con respecto a β e igualamos a cero:

$$\frac{\partial}{\partial \beta} \log L(\beta, \sigma^2) = \frac{1}{\sigma^2} X^T Y - \frac{1}{2\sigma^2} (2X^T X\beta) = \frac{1}{\sigma^2} (X^T Y - X^T X\beta) = 0$$

Entonces:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Por lo tanto, la solución coincide con la del problema de mínimos cuadrados.

-Investiga el contenido del Teorema de Gauss-Markov sobre mínimos cuadrados.

El Teorema de Gauss-Markov establece que los estimadores de mínimos cuadrados ordinarios son lineales, insesgados y óptimos, “Best Linear Unbiased Estimator (BLUE)”; es decir, tienen varianza mínima entre la clase de los estimadores lineales e insesgados.

El Teorema está basado en los siguientes supuestos:

- El modelo esta correctamente especificado.
- Debe ser lineal en los parámetros.
- El valor de la media condicional es cero.
- Hay homocedasticidad.
- No existe correlación entre las perturbaciones.
- La covarianza entre ϵ y X es cero.
- El número de observaciones es mayor que el de parámetros.
- Existe variabilidad entre los X .
- No hay multicolinealidad perfecta.
- Las X son no estocásticas; es decir, son fijas en muestras repetidas.

Parte aplicada

Primero utilizaremos la base `diamonds` y posteriormente probaremos con la base `ciudades`:

Base diamonds

Cargamos la base `diamonds` y el paquete `ggplot2`

```
library(ggplot2)
data(diamonds)
head(diamonds)
```

```
##   carat      cut color clarity depth table price     x     y     z
## 1  0.23    Ideal    E    SI2   61.5     55   326 3.95 3.98 2.43
## 2  0.21  Premium    E    SI1   59.8     61   326 3.89 3.84 2.31
## 3  0.23      Good    E    VS1   56.9     65   327 4.05 4.07 2.31
## 4  0.29  Premium    I    VS2   62.4     58   334 4.20 4.23 2.63
## 5  0.31      Good    J    SI2   63.3     58   335 4.34 4.35 2.75
## 6  0.24  Very Good   J   VVS2   62.8     57   336 3.94 3.96 2.48
```

-Regresion lineal para explicar la variable price usando las demas variables numéricas:

```
modelo1<-lm(price~carat+depth+table+x+y+z, diamonds)
summary(modelo1)

##
## Call:
## lm(formula = price ~ carat + depth + table + x + y + z, data = diamonds)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -23878.2   -615.0    -50.7   347.9 12759.2 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 20849.316   447.562  46.584 < 2e-16 ***
## carat       10686.309   63.201 169.085 < 2e-16 ***
## depth        -203.154   5.504 -36.910 < 2e-16 ***
## table        -102.446   3.084 -33.216 < 2e-16 ***
## x            -1315.668  43.070 -30.547 < 2e-16 ***
## y              66.322   25.523   2.599  0.00937 **  
## z              41.628   44.305   0.940  0.34744    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 1497 on 53933 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592 
## F-statistic: 5.486e+04 on 6 and 53933 DF, p-value: < 2.2e-16
```

Como la variable z no es significativa, no la vamos a considerar en el modelo. Con lo anterior, el modelo quedaría como sigue:

```
modelo1<-lm(price~carat+depth+table+x+y, diamonds)
summary(modelo1)

##
## Call:
## lm(formula = price ~ carat + depth + table + x + y, data = diamonds)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -23872.1   -614.8    -50.5   347.5 12759.4 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 20702.947   419.575  49.343 < 2e-16 ***
## carat       10686.707   63.199 169.095 < 2e-16 ***
## depth        -200.718   4.855 -41.344 < 2e-16 ***
## table        -102.490   3.084 -33.234 < 2e-16 *** 
## x            -1293.542  36.063 -35.869 < 2e-16 ***
```

```

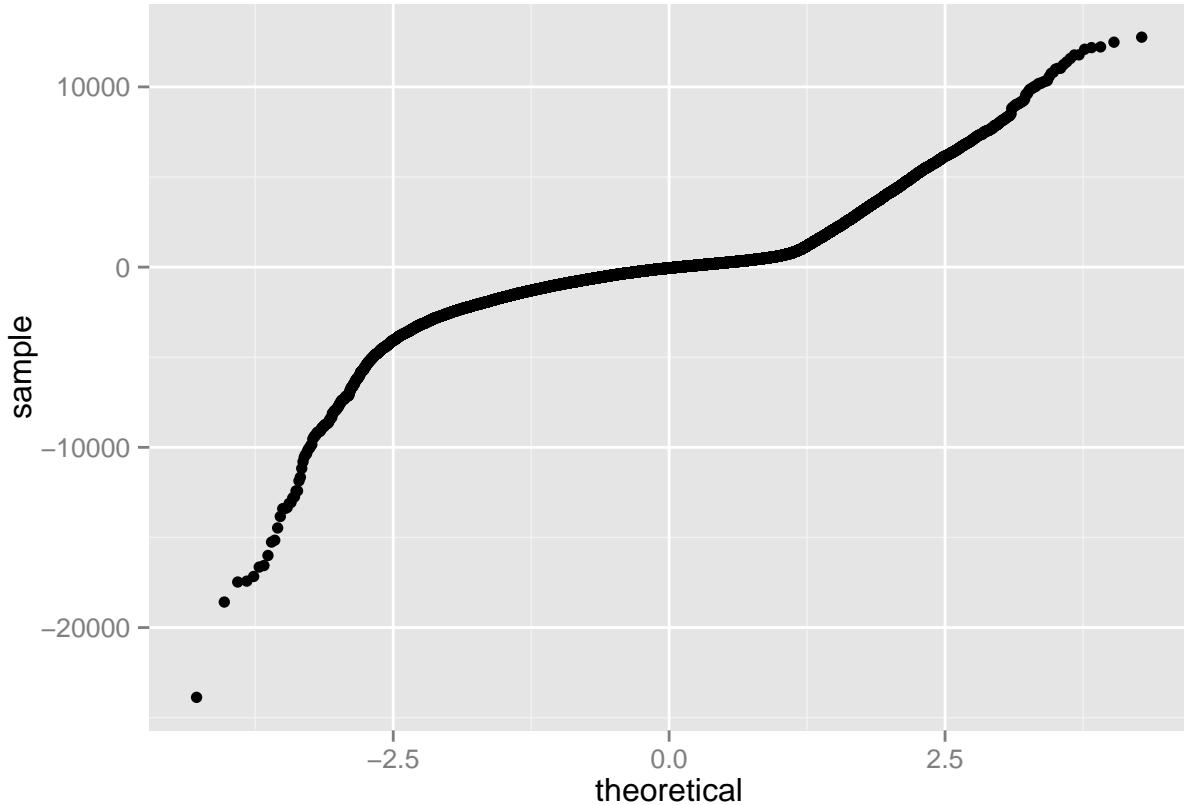
## y           69.575    25.287   2.751  0.00594 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1497 on 53934 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592
## F-statistic: 6.583e+04 on 5 and 53934 DF,  p-value: < 2.2e-16

```

-¿Qué tan bueno fue el ajuste?

En teoría los residuos tienen una distribución normal. Para verificar usamos la siguiente gráfica de qqplot:

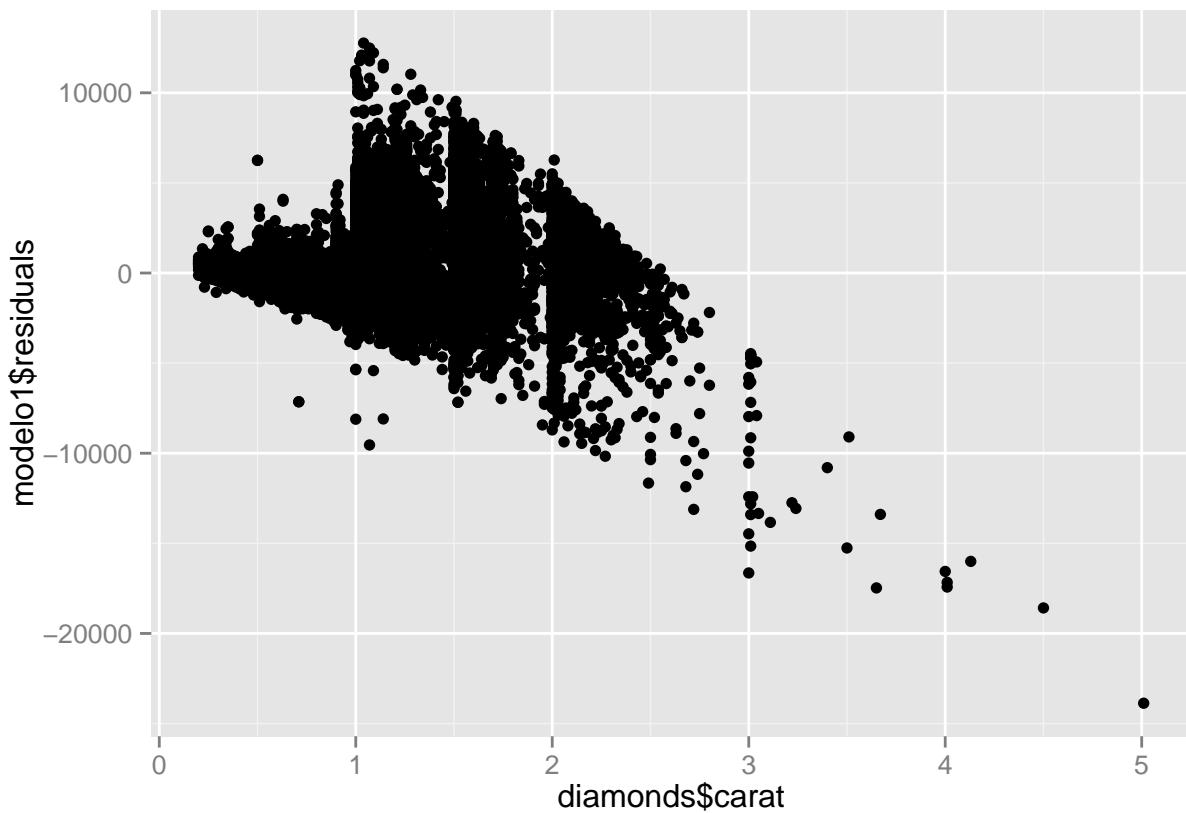
```
qqplot(sample = modelo1$residuals)
```



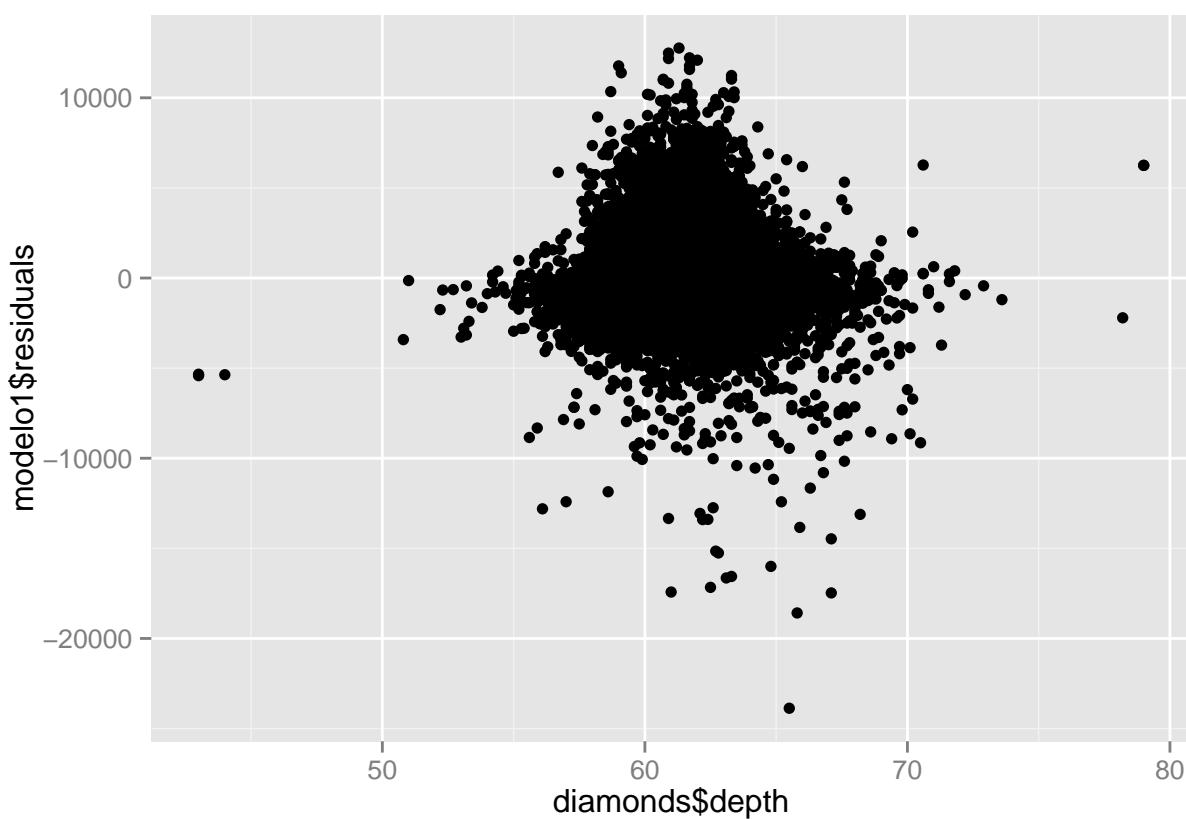
Esperaríamos una línea de 45 grados, en este caso no es así por lo que los residuos no cumplen con el supuesto de normalidad.

Para revisar si los errores son homocedásticos, los graficamos contra cada una de las variables independientes. Lo que estaríamos esperando es que no existan outliers, que no tenga un patrón ni tendencia.

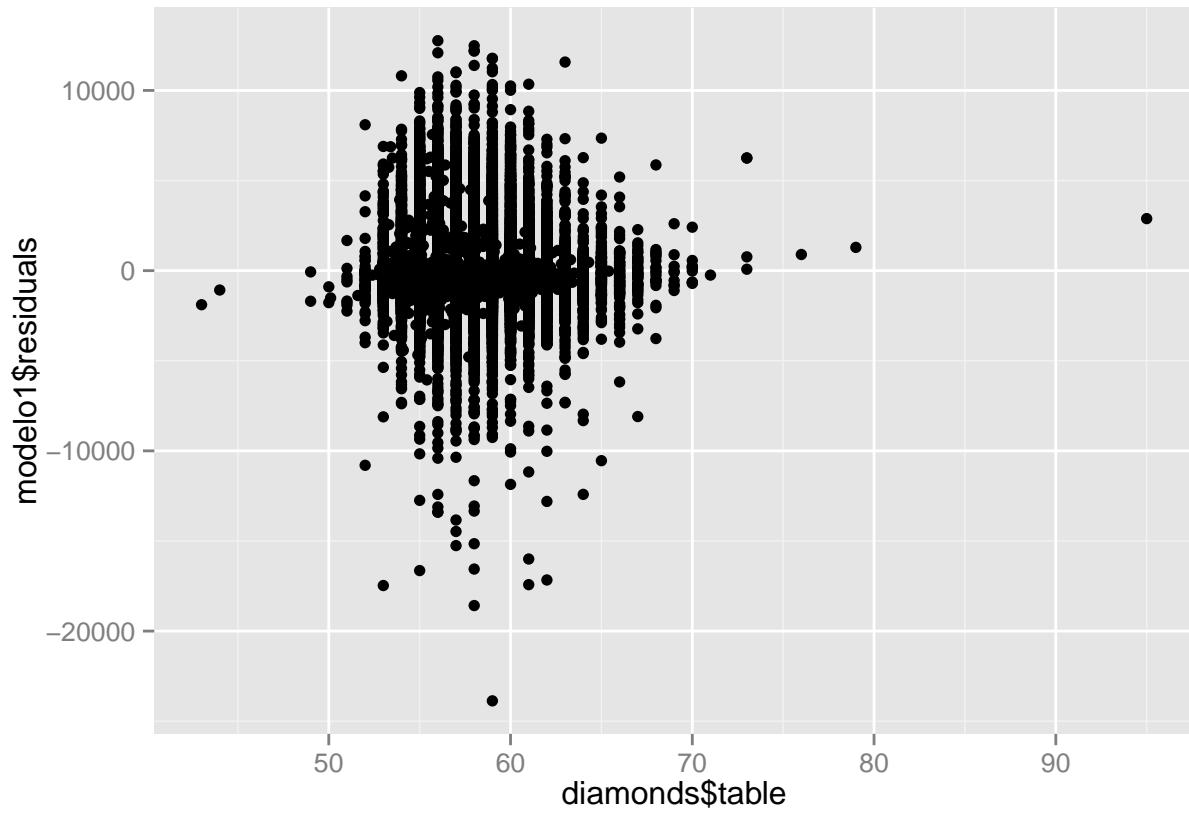
```
qqplot(diamonds$carat,modelo1$residuals)
```



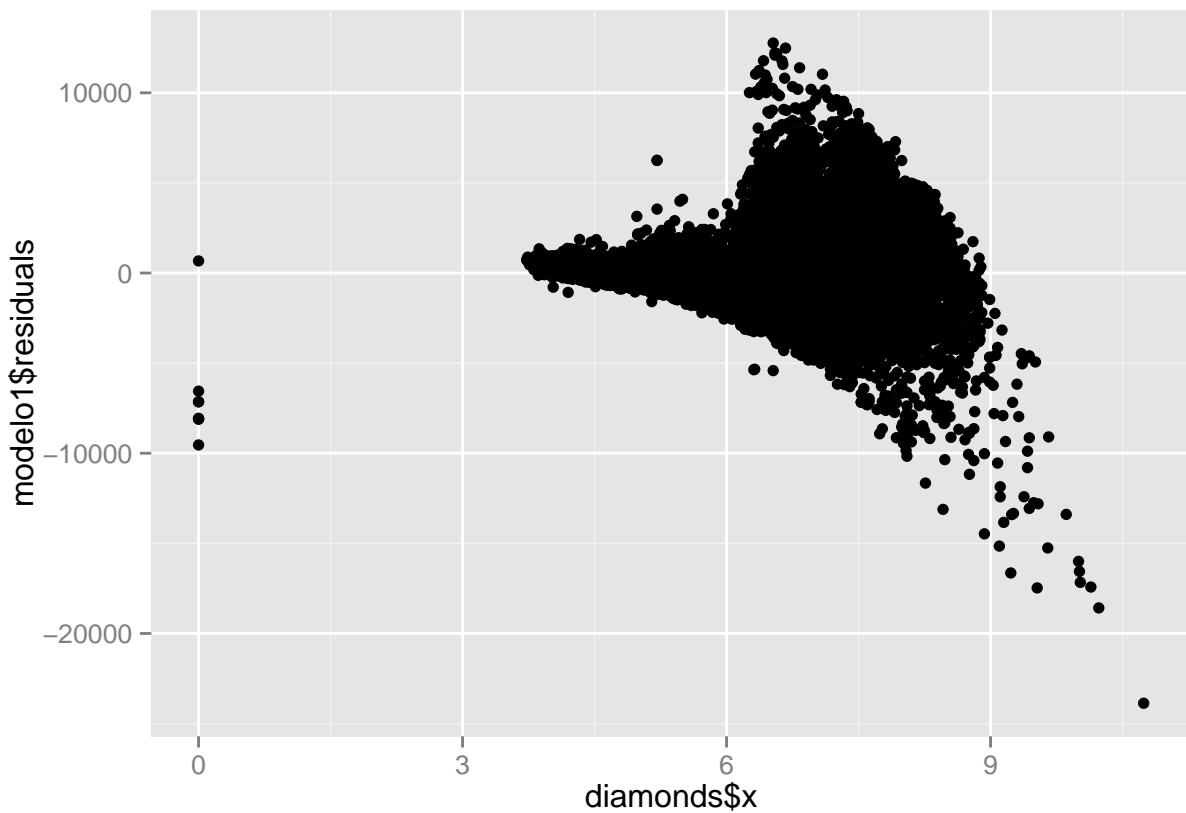
```
qplot(diamonds$depth,modelo1$residuals)
```



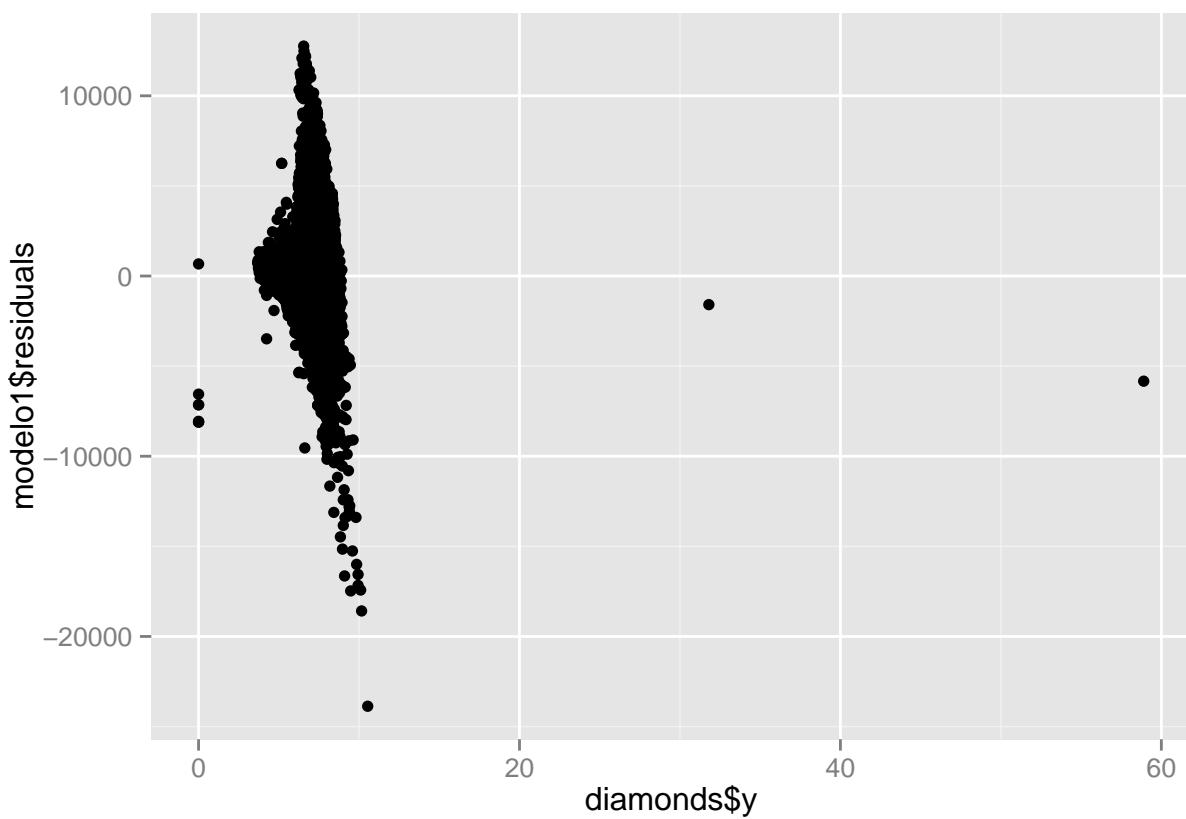
```
qplot(diamonds$table,modelo1$residuals)
```



```
qplot(diamonds$x,modelo1$residuals)
```



```
qplot(diamonds$y, modelo1$residuals)
```

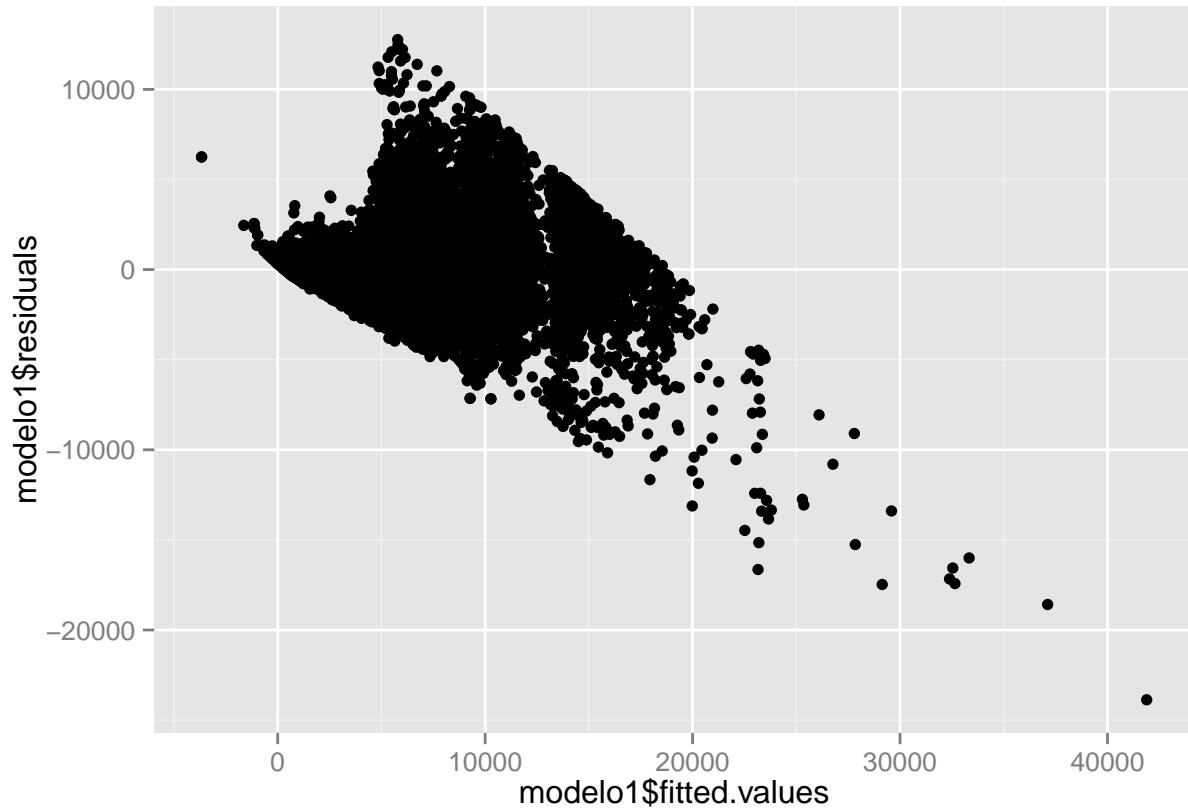


En las gráficas se puede observar tendencia de los residuos (diferencias de los residuos de acuerdo al valor de las variables independientes).

Por lo anterior, se puede concluir que en todos los casos se observa heterocedasticidad.

Teóricamente los valores fitted y los residuos no están correlacionados. Para comprobar los graficamos:

```
qplot(modelo1$fitted.values,modelo1$residuals)
```



Sin embargo, la gráfica nos muestra que hay una correlación negativa entre los residuos y los valores fitted.

Con todos estos elementos podemos concluir que el ajuste no es el adecuado pues no se cumplen varios de los supuestos que se requieren para utilizar el modelo de regresión lineal.

-¿Qué medida puede ayudarnos a saber la calidad del ajuste?

La medida R^2 nos ayuda a conocer la calidad del ajuste lineal determinado por los estimadores por mínimos cuadrados ordinarios. Esta medida en el caso de una regresión lineal múltiple se obtiene como la suma de cuadrados explicados entre la suma de cuadrados totales.

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

Nos va a indicar que proporción de la variación de Y es explicada por la regresión.

En nuestro ejemplo, el valor de la $R^2 = 0.8592$, indicando que el modelo es adecuado considerando únicamente la proporción de la variación explicada. No obstante lo anterior, esta medida no debe ser lo único a considerar para evaluar el ajuste, ya que como vimos en el punto anterior, al no cumplirse varios de los supuestos o hipótesis básicas, los resultados no son del todo fiables.

-¿Cuál fue el valor de σ^2 que ajustó su modelo y que relación tiene con la calidad del ajuste?

La varianza residual del modelo es:

```
var(modelo1$residuals)  
## [1] 2240660
```

La relación que tiene con la calidad de ajuste es que el inverso de la varianza nos sirve para ponderar las covarianzas entre las variables independientes y la variable dependiente. Lo anterior, pues nos interesan las β 's que son precisas (inverso de la varianza) y que están relacionadas con Y .

-¿Cuál es el ángulo entre Y y \hat{Y} ?

La $R^2 = \cos^2(\theta) = 0.8592$; es decir, la R^2 es el coseno al cuadrado del ángulo entre la variable dependiente Y y la variable estimada \hat{Y} , ambas centradas.

Por esta razón, el ángulo $\theta = \cos^{-1}(R) = \cos^{-1}(\sqrt{R^2})$:

```
acos(sqrt(0.8592))
```

```
## [1] 0.3846484
```

-Definan una función que calcule la logverosimilitud de unos parámetros β y σ^2 .

Definimos la función `logver` como sigue:

```
logver <- function(par, X, y) {  
  n <- length(y)  
  p <- dim(X)[2]  
  beta <- par[1:p]  
  sigma2 <- par[p+1]  
  mu <- X %*% beta  
  logl<- suppressWarnings(-0.5*n*log(2*pi)-0.5*n*log(sigma2)-(1/(2*sigma2))*sum((y-mu)**2))  
  return(-logl)  
}
```

-Utilicen la función `optim` de R para obtener numéricamente el máximo de la función de verosimilitud. La solución debe coincidir con la del método `lm`.

Usando la función `optim` de R :

```
beta_sigma<-as.matrix(c(1,1,1,1,1,1,100))  
X<-as.matrix(cbind(rep(1,53940), diamonds[,c(1,5,6,8,9)]))  
y<-diamonds[,7]  
optim(par=beta_sigma, logver, X=X, y=y, method="L-BFGS-B")
```

```

## $par
## [1] [,1]
## [1,] 21038.24864
## [2,] 10739.65119
## [3,] -204.00071
## [4,] -103.30346
## [5,] -1335.23385
## [6,] 88.92062
## [7,] 2266521.97025
##
## $value
## [1] 470902.5
##
## $counts
## function gradient
##      122      122
##
## $convergence
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"

```

La solución es muy cercana a la que se obtuvo utilizando el método `lm`.

Base ciudades

Ahora utilizaremos la base `ciudades` obtenida de <http://knoema.es> para definir una regresión lineal usando la función `lm`, ver si ajusta correctamente y finalmente utilizar nuevamente la función `optim` para ver si la solución coincide con la del metodo `lm`.

Leemos la base que tiene como variables el rating de la ciudad (`Overall_Rating`), estabilidad (`Stability`), salud (`Healthcare`), cultura (`Culture`) y educación (`Education`):

```

ciudades = read.table("/Users/alex/Documents/Maestría Ciencia de Datos/Tarea3/ciudades1.csv", row.names=1)
head(ciudades)

##          Overall_Rating Stability Healthcare Culture Education
## Abidjan           49.7        45     45.8    54.2     50.0
## Abu Dhabi         73.1        85     66.7    59.0     66.7
## Adelaide          96.6        95    100.0    94.2    100.0
## Al Khobar          54.2        65     62.5    34.0     58.3
## Algiers            40.9        40     45.8    42.6     50.0
## Almaty             65.3        75     66.7    57.6     66.7

```

Vamos a buscar explicar la variable `Overall_Rating` en función de las otras variables como sigue:

```

modelo2<-lm(Overall_Rating~Stability+Healthcare+Culture+Education, data=ciudades)
summary(modelo2)

```

```

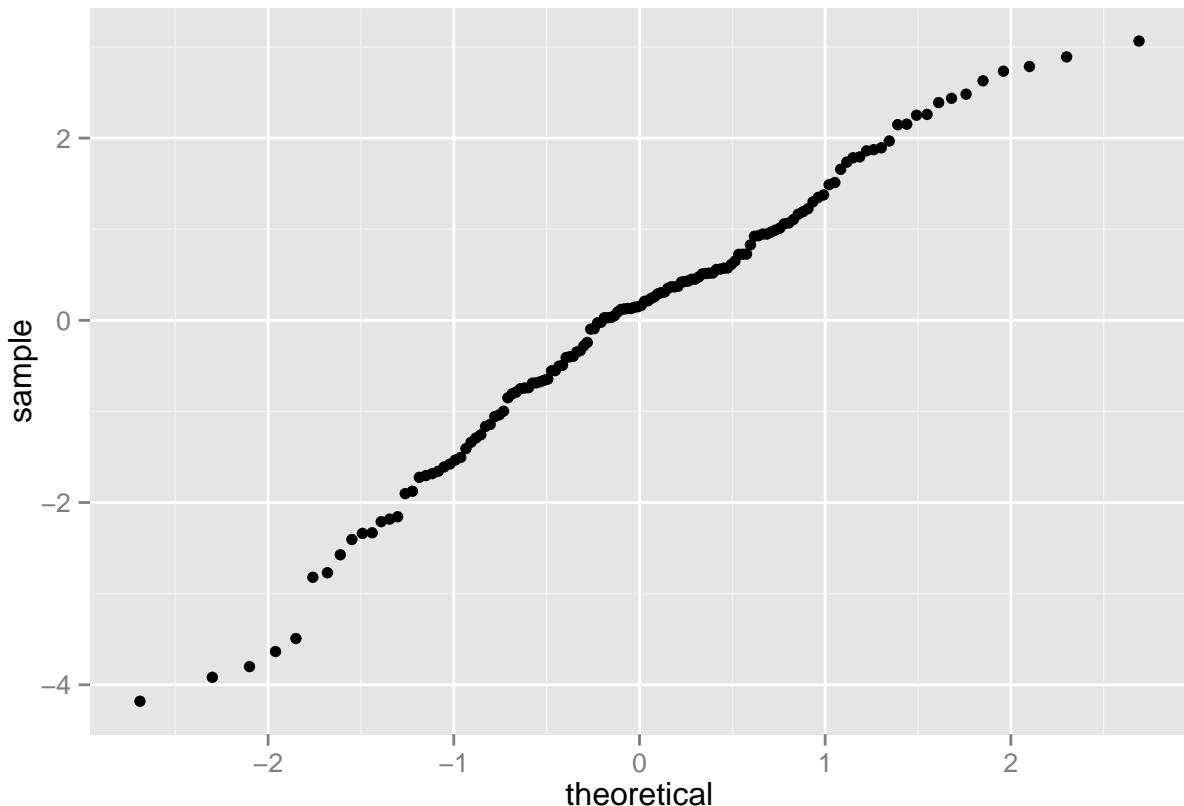
## 
## Call:
## lm(formula = Overall_Rating ~ Stability + Healthcare + Culture +
##     Education, data = ciudades)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -4.1807 -0.7924  0.1556  0.9467  3.0649 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.25395   0.65334  -0.389   0.698    
## Stability    0.31238   0.01020  30.640  <2e-16 ***
## Healthcare   0.25427   0.01475  17.239  <2e-16 *** 
## Culture      0.27531   0.01328  20.724  <2e-16 *** 
## Education    0.16211   0.01462  11.089  <2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 1.517 on 135 degrees of freedom
## Multiple R-squared:  0.9927, Adjusted R-squared:  0.9925 
## F-statistic:  4594 on 4 and 135 DF,  p-value: < 2.2e-16

```

Verificamos el ajuste del modelo:

En primer lugar vemos que la R^2 es muy cercana a 1, lo que es un indicativo de que el ajuste es adecuado.

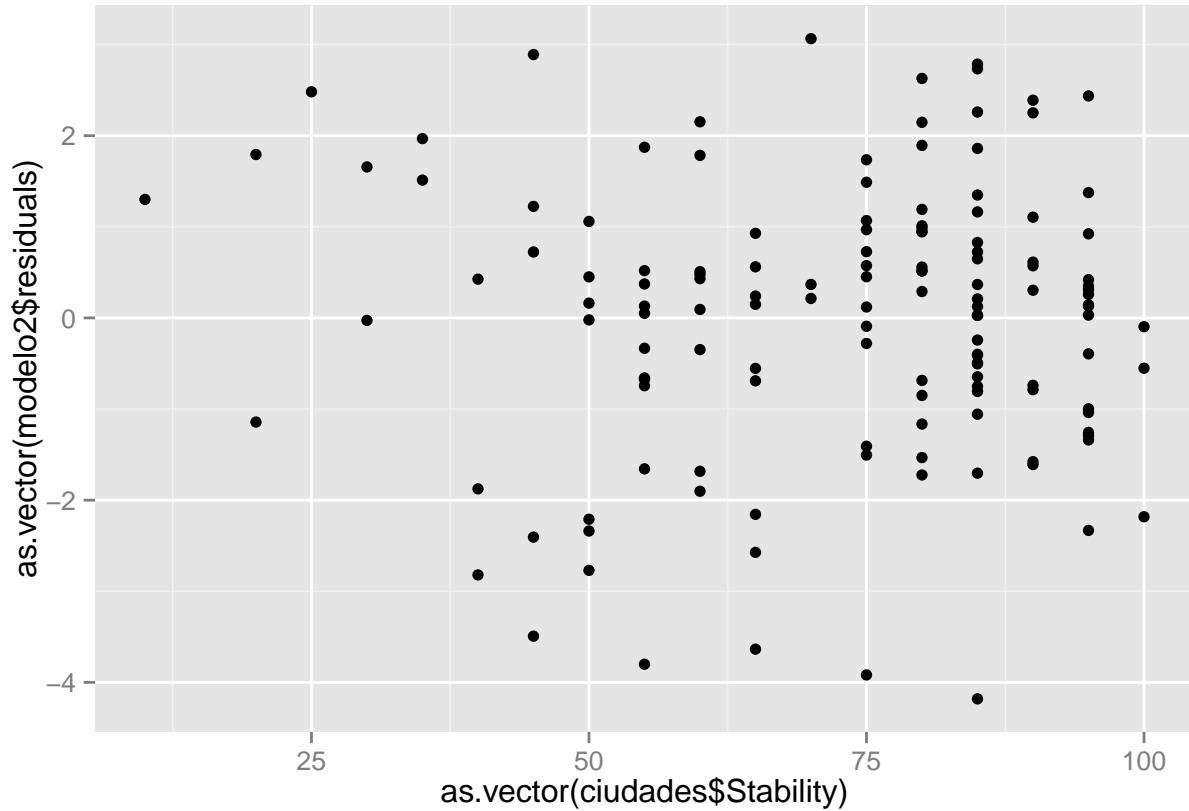
```
qplot(sample =as.vector(modelo2$residuals))
```



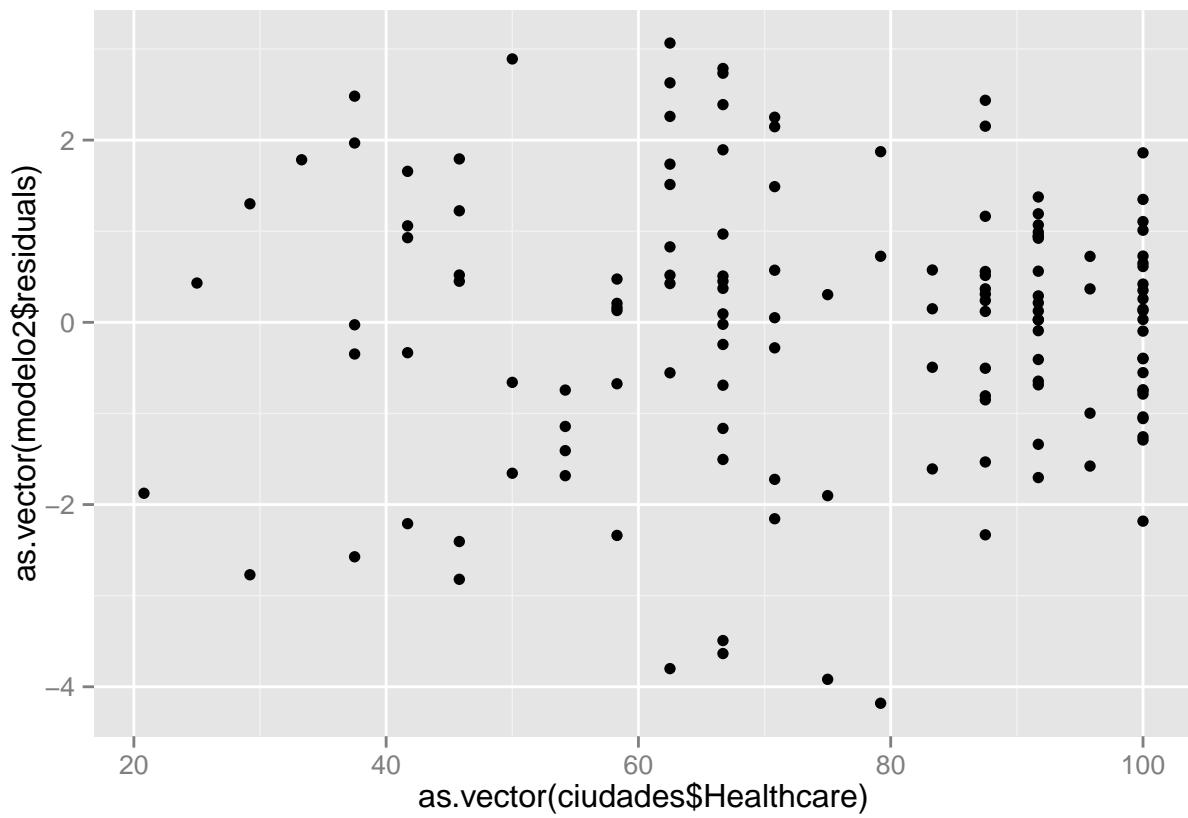
En la gráfica anterior podemos ver que los residuos cumplen con el supuesto de normalidad.

Asimismo, como se puede constatar en las siguientes gráficas los errores son homocedásticos, pues no se aprecian relaciones ni tendencias:

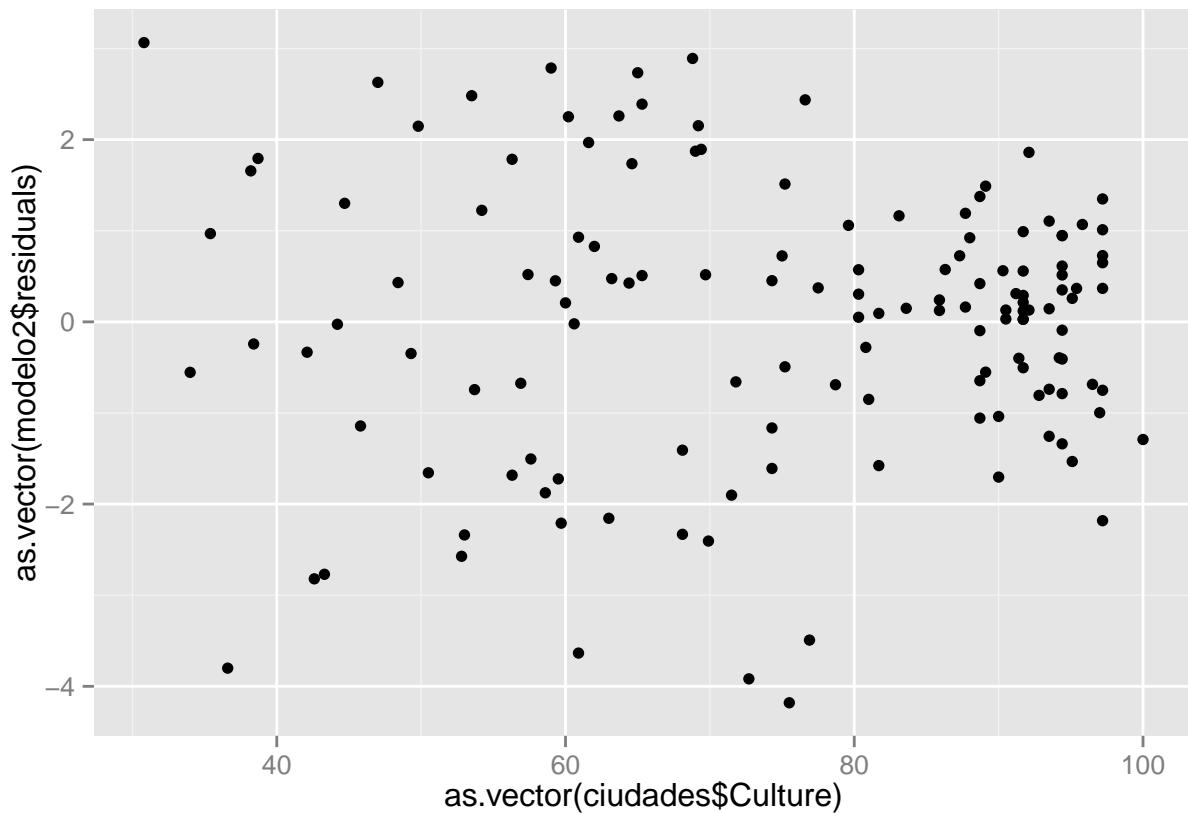
```
qplot(as.vector(ciudades$Stability),as.vector(modelo2$residuals))
```



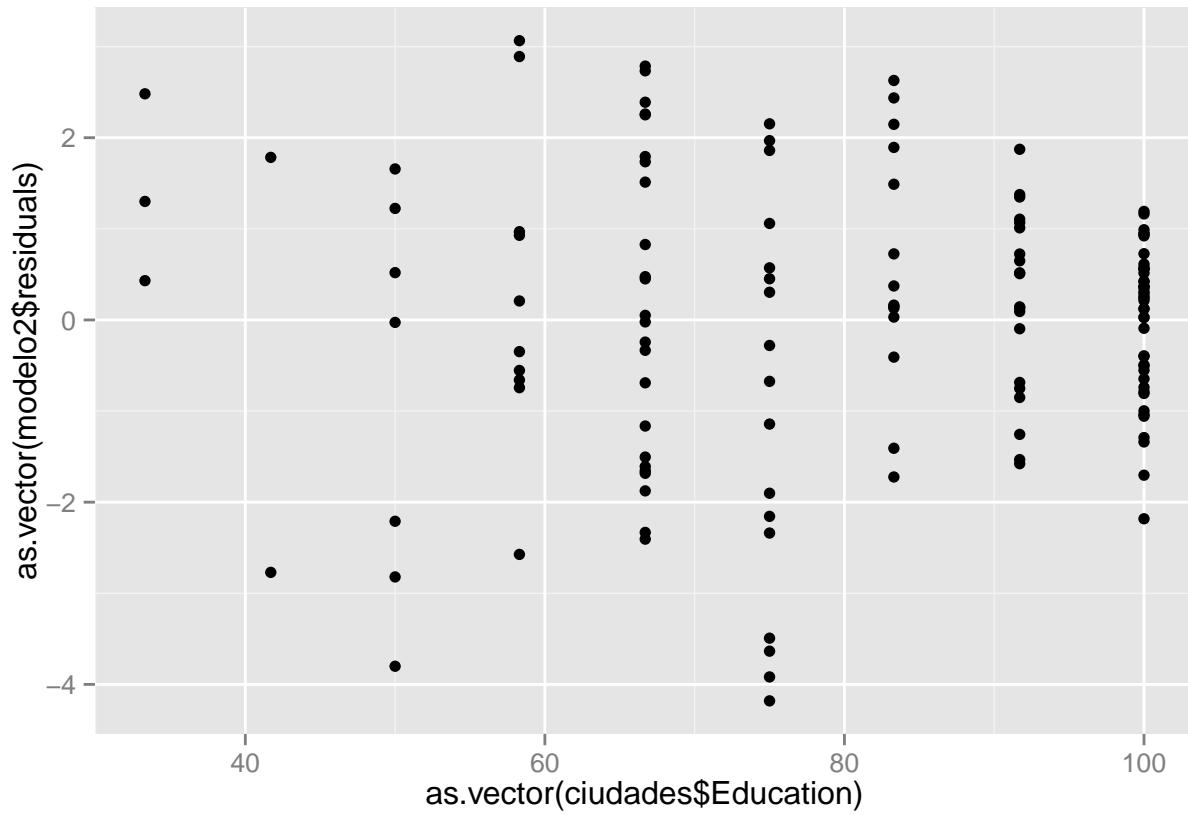
```
qplot(as.vector(ciudades$Healthcare),as.vector(modelo2$residuals))
```



```
qplot(as.vector(ciudades$Culture), as.vector(modelo2$residuals))
```

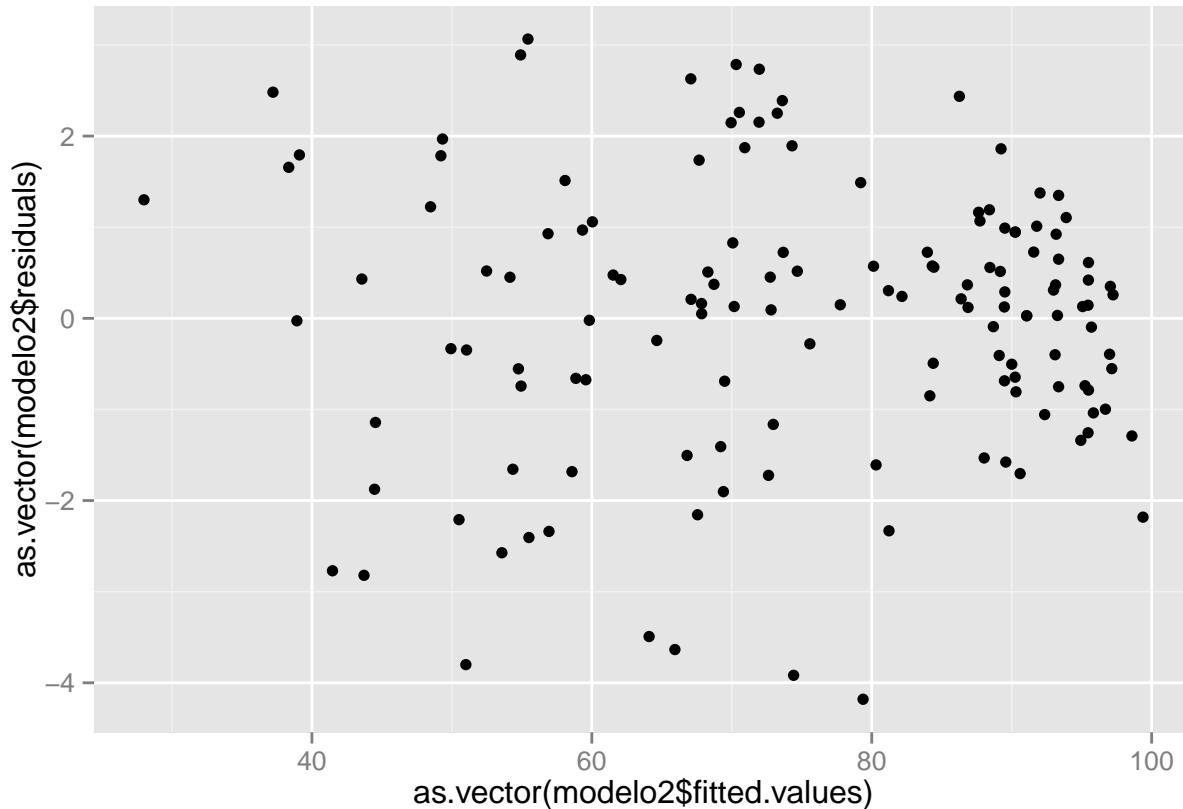


```
qplot(as.vector(ciudades$Education), as.vector(modelo2$residuals))
```



De manera similar, los valores fitted y los residuos no estan correlacionados:

```
qplot(as.vector(modelo2$fitted.values), as.vector(modelo2$residuals))
```



Por lo anterior, podemos concluir que el modelo ajusta bien.

Vamos a corroborar que la solución del método `lm` coincide con el máximo de la función de verosimilitud obtenido a partir de la función `optim`:

Método `lm`:

```
summary(modelo2)

##
## Call:
## lm(formula = Overall_Rating ~ Stability + Healthcare + Culture +
##     Education, data = ciudades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1807 -0.7924  0.1556  0.9467  3.0649
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.25395   0.65334 -0.389   0.698
## Stability    0.31238   0.01020 30.640 <2e-16 ***
## Healthcare   0.25427   0.01475 17.239 <2e-16 ***
## Culture      0.27531   0.01328 20.724 <2e-16 ***
## Education    0.16211   0.01462 11.089 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.517 on 135 degrees of freedom
```

```
## Multiple R-squared:  0.9927, Adjusted R-squared:  0.9925
## F-statistic:  4594 on 4 and 135 DF,  p-value: < 2.2e-16
```

Función optim:

```
X<-as.matrix(cbind(rep(1,dim(ciudades)[1]), ciudades[,2:5]))
y<-ciudades[,1]
beta_sigma2<-as.matrix(rep(1,6))
optim(par=beta_sigma2, logver, X=X, y=y, method="BFGS")
```

```
## $par
##      [,1]
## [1,] -0.2544537
## [2,]  0.3123800
## [3,]  0.2542704
## [4,]  0.2753100
## [5,]  0.1621157
## [6,]  2.2182886
##
## $value
## [1] 254.4175
##
## $counts
## function gradient
##       93        26
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Finalmente, podemos concluir que las soluciones son **identicas bajo ambos métodos**.