

## 7. Regresión lineal simple con datos de “performance”.

Consideraremos los datos en la base `performance.csv` y las variables:  $y$  = academic performance of the school (`api00`) y  $x$  = percentage of students receiving free meals (`meals`). Estos datos corresponden a una muestra aleatoria de 400 escuelas primarias en California, en donde por escuela se realizaron mediciones que tienen que ver con su desempeño en el año 2000.

### i) Regresión lineal simple y verificación de supuestos.

Ajustaremos un modelo de regresión lineal simple del desempeño escolar (`api00`) en función del porcentaje de estudiantes que recibieron desayunos gratuitos en las escuelas (`meals`).

Table 1: MODELO 1

<i>Dependent variable:</i>	
	<code>api00</code>
<code>meals</code>	-4.015*** s.e.(0.097) p-value: <2e-16
Constant	889.783*** (6.622) p-value: <2e-16
Observations	400
R <sup>2</sup>	0.811
Adjusted R <sup>2</sup>	0.811
Residual Std. Error	61.877 (df = 398)
F Statistic	1,710.691*** (df = 1; 398); p-value: < 2.2e-16
<i>Note:</i>	
*p<0.1; **p<0.05; ***p<0.01	

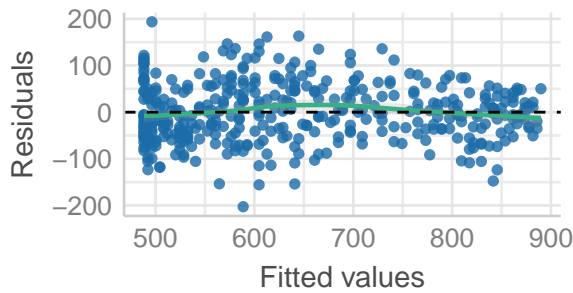
En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk y Breusch-Pagan y Durbin-Watson, en el primer caso de la normalidad el p-value asociado es mayor a 0.05, por lo que no hay evidencia para rechazar las hipótesis nulas de normalidad, sin embargo hay problemas de heterocedasticidad. Como la muestra se generó aleatoriamente, podemos asumir que no tenemos problemas de autocorrelación de los errores.

	1
Normality (Shapiro-Wilk)	0.618
Homoscedasticity (Breusch-Pagan)	0.002

También podemos observar de forma gráfica estos resultados. Observemos la gráfica de **Fitted values** contra **Residuals**, parece haber un problema de linealidad. En la gráfica **Standard Normal distribution Quantiles** contra **Sample Quantile Deviations** tenemos que la normalidad sí se preserva. En la gráfica de **Fitted Values** contra  $\sqrt{|Std.Residuals|}$  parece no haber homogeneidad de la varianza. Y finalmente, en la gráfica de **Leverage(hii)** contra **Std. Residuals** parece no haber valores atípicos influyentes. Entonces podemos concluir que nuestro modelo no cumple con dos supuestos importantes, la linealidad y la homocedasticidad.

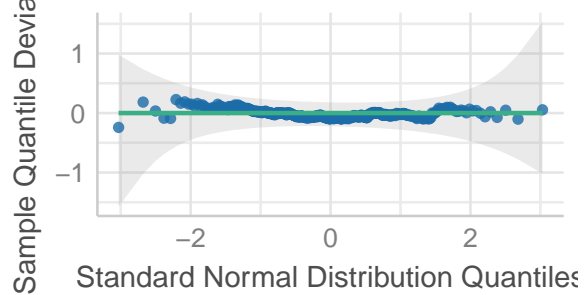
### Linearity

Reference line should be flat and horizontal



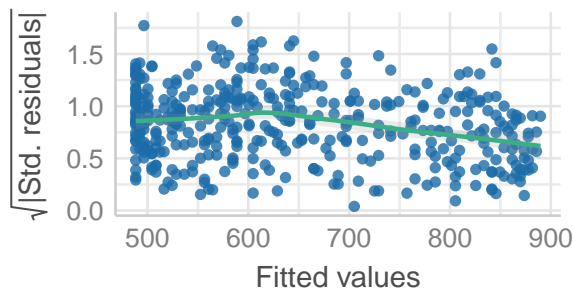
### Normality of Residuals

Points should fall along the line



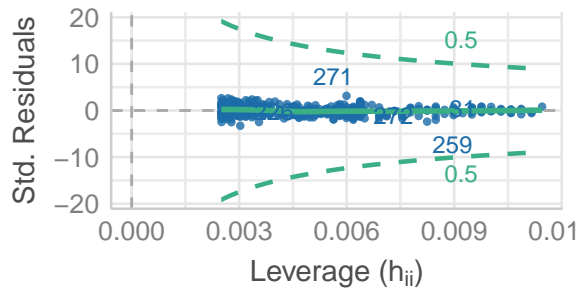
### Homogeneity of Variance

Reference line should be flat and horizontal



### Influential Observations

Points should be inside the contour lines



## ii) Ajuste de un mejor modelo que cumple los supuestos.

Debido a que con el primer modelo no se satisfacen los supuestos de linealidad y homocedasticidad procederemos a hacer algunas pruebas para corregir estos problemas. Haremos unas pruebas para ver qué transformación podría ser la adecuada para la variable dependiente e independiente.

```
## PowerTransform:
## Estimated transformation parameter
##      Y1
## 1.588293
## BoxTidwell:
## MLE of lambda Score Statistic (t) Pr(>|t|)
##      1.2132      -2.236  0.02591 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations = 3
```

El resultado de la prueba `Estimated transformation parameter` con la función `powerTransform` para transformación de tipo BoxCox, para conocer el exponente  $\lambda$  de la variable dependiente, muestra que el valor es de  $c(Y1 = 1.58829320628201)$ . Esto sugiere elevar a un exponente de  $3/2 = 1.5$  a la variable dependiente. Por otra parte, la prueba con la función `BoxTidwell` para la transformación de la variable independiente (modificada al sumarle +1) muestra un valor  $\lambda$  de 1.2132 con un p-value asociado de 0.02591, lo que implica que  $\lambda \neq 1$  y hay que transformar a la variable independiente.

Sin embargo, primero trataremos de hacer cambios a la variable independiente, ya que las transformaciones de la variable dependiente vuelve un poco más complicada la interpretación de los coeficientes estimados y en general el modelo. Primero probaremos con la transformación **cuadrática** a la variable independiente, en caso de que se cumplan los supuestos nos quedaremos con este modelo.

Table 3: MODELO 2

Dependent variable:	
api00	
I((meals)^(2))	-0.03592*** s.e. (0.0009269) t value: -38.75 Pr(> t ): <2e-16
Constant	814.761*** s.e. (5.407) t value: 150.70 Pr(> t ): <2e-16
Observations	400
R <sup>2</sup>	0.790
Adjusted R <sup>2</sup>	0.790
Residual Std. Error	65.195 (df = 398)
F Statistic	1,501.536*** (df = 1; 398); p-value: < 2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01

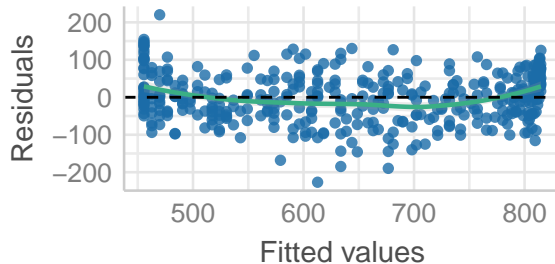
En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk y Breusch-Pagan, el p-value asociado es mayor a 0.05 para ambos casos, por lo que no hay evidencia para rechazar las hipótesis nulas de normalidad y homocedasticidad. Como se mencionó anteriormente, la muestra se generó aleatoriamente, por lo que podemos asumir que no tenemos problemas de autocorrelación de los errores.

	1
Normality (Shapiro-Wilk)	0.833
Homoscedasticity (Breusch-Pagan)	0.091

También podemos observar de forma gráfica estos resultados.

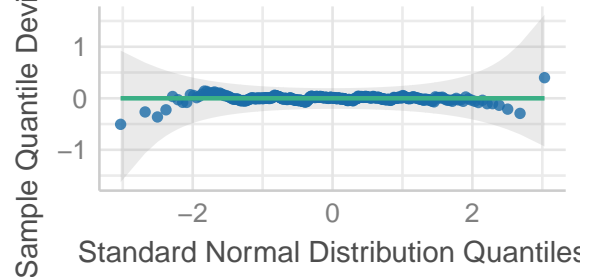
### Linearity

Reference line should be flat and horizontal



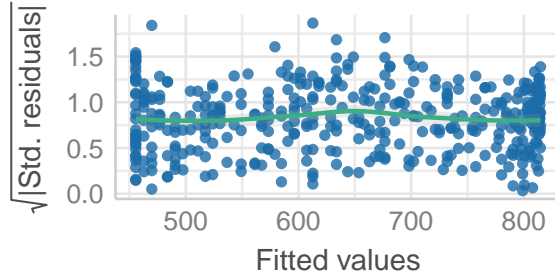
### Normality of Residuals

Points should fall along the line



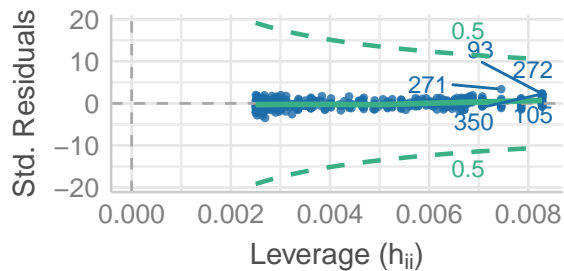
### Homogeneity of Variance

Reference line should be flat and horizontal



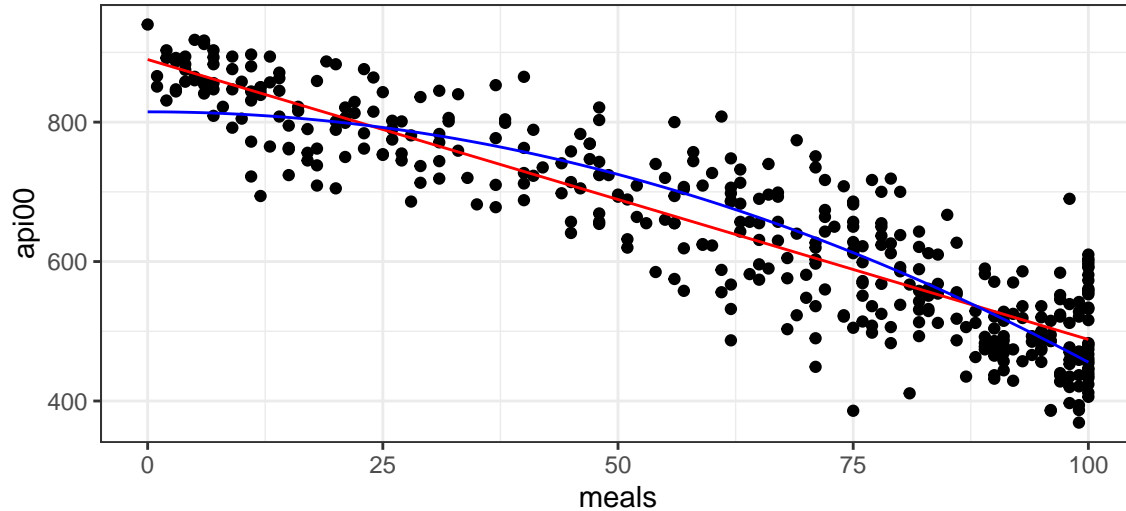
### Influential Observations

Points should be inside the contour lines



### iii) Gráfica de datos originales y las curvas ajustadas de ambos modelos.

A continuación se muestra la Gráfica de los datos originales y las curvas ajustadas tanto para el primero modelo sin tratamiento de las variables (recta roja) y la curva ajustada del segundo modelo con la variable independiente cuadrática (curva azul).



### iv) Interpretación de la prueba ANOVA y la $R^2$ .

En el Modelo 2, se tiene un  $R^2$  de 0.79, el cual es el coeficiente de determinación que en este caso se interpreta como que el 79% de la variabilidad del rendimiento académico en la escuela `api00` se explica por el modelo que incluye la variable del porcentaje de estudiantes que reciben desayuno en la escuela `meal`. Por otra parte, la prueba  $F$  asociada a la tabla ANOVA, contrasta en este caso de la regresión lineal simple las hipótesis nula  $H_0 : \beta_1 = 0$  contra la alternativa  $H_a : \beta_1 \neq 0$ . Como el p-value asociado es menor a  $2e - 16$  se rechaza  $H_0$  con una significancia estadística del 5%, podemos concluir que la inclusión de la variable explicativa `meal` ayuda a modelar  $E(\text{api00}; \text{meal})$ . Es decir, el rendimiento académico en la escuela `api00` se relaciona linealmente con la variable del porcentaje de estudiantes que reciben desayuno en la escuela `meal`.

### v) Prueba de hipótesis de investigación.

Para verificar el argumento de que “A mayor porcentaje de comidas gratis en la escuela es menor el desempeño de la escuela”, plantearemos una prueba de hipótesis. Planteamos la hipótesis nula  $H_0 : \beta_1 \geq 0$  contra la alternativa  $H_a : \beta_1 < 0$ , donde  $\beta_1$  es el parámetro estimado asociado a la variable independiente `meal`. A continuación se muestra la prueba **Simultaneous Tests for General Linear Hypotheses**, donde se rechaza esta hipótesis nula, pues el p-valor asociado es menor a 0.05, considerando un nivel de confianza del 95%.

Simultaneous Tests for General Linear Hypotheses				
Fit:lm(formula = api00 ~ I((meals)^(2)), data = datos7)				
Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 $\geq 0$	-0.0359152	0.0009269	-38.75	<2e-16 ***
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.' 0.1 ' ' 1
(Adjusted p values reported – single-step method)				