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 procentaje de estudiantes que recibieron desayunos gratuitos en las escuelas (meals).

Table 1: MODELO 1

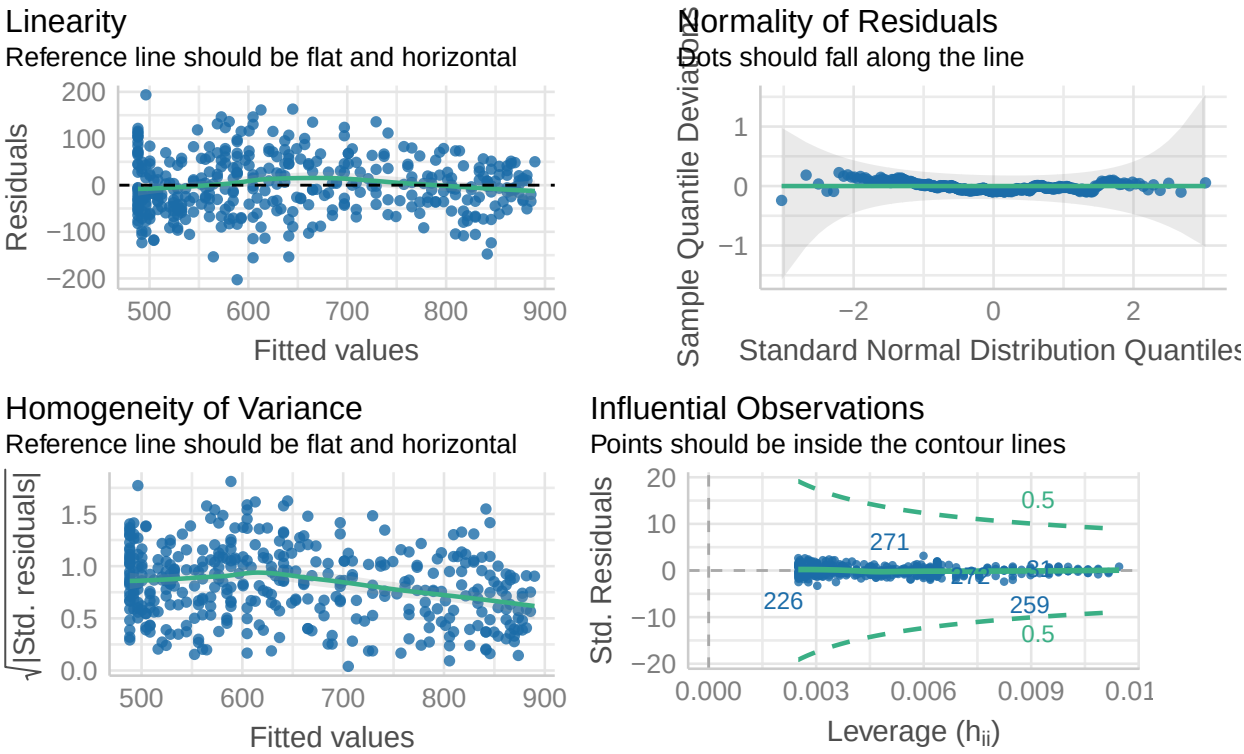
	Dependent variable:
	api00
meals	-4.015^{***}
	s.e.(0.097)
	p-value: <2e-16
Constant	889.783^{***}
	(6.622)
	p-value: <2e-16
Observations	400
R^2	0.811
Adjusted R^2	0.811
Residual Std. Error	61.877 (df = 398)
F Statistic	$1{,}710.691^{***}$ (df = 1; 398); p-value: < 2.2e-16
Note:	*p<0.1; **p<0.05; ***p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

En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk y Breusch-Pagan 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 rechzar 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

La normalidad de los errores se confirma con la prueba `Anderson-Darling normality test` con la función `ad.test` que muestra un p-value de 0.276497297460581. La heterocedasticidad se confirma con la prueba `Non-constant Variance Score Test` con la función `ncvTest`, que muestra un p-value de 0.00171467461271069. Finalmente, con la función de `residualPlots` obtenemos para la prueba de `Tukey test` un p value de 0.01174, por lo que se rechaza la hipótesis nula de linealidad.

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.



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

Presentamos pruebas para ver qué transformación es adecuada para la variable dependiente e independiente.

```
## Estimated transformation parameter
##       Y1
## 1.588293
```

```
##  MLE of lambda Score Statistic (t) Pr(>|t|)
##        0.93552              0.7774   0.4374
##
## iterations =  2
```

El resultado de la prueba `Estimated transformation parameter` con la función `powerTransform` para transformación de tipo BoxCox, para conocer el exponente λ de la variable dependiente, muestra que el valor es de c(Y1 = 1.58829320628201). Esto sugiere elevar a un exponente de 1.6 a la variable dependiente, por simplicidad en la interpretación consideraremos un exponente de 2. Por otra parte, la prueba con la función `BoxTidwell` para la tranformación de la variable independiente (modificada al sumarle +1 y tomando en cuenta la variable dependiente al cuadrado) muestra un valor λ de 0.93552 con un p-value asociado de 0.4374, lo que implica que la hipótesis nula de que $\lambda = 1$ no se rechaza, i.e., no hay evidencia suficiente para rechazar la linealidad de la variable independiente. Entonces ajustamos el MODELO 2.

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 rechzar 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.

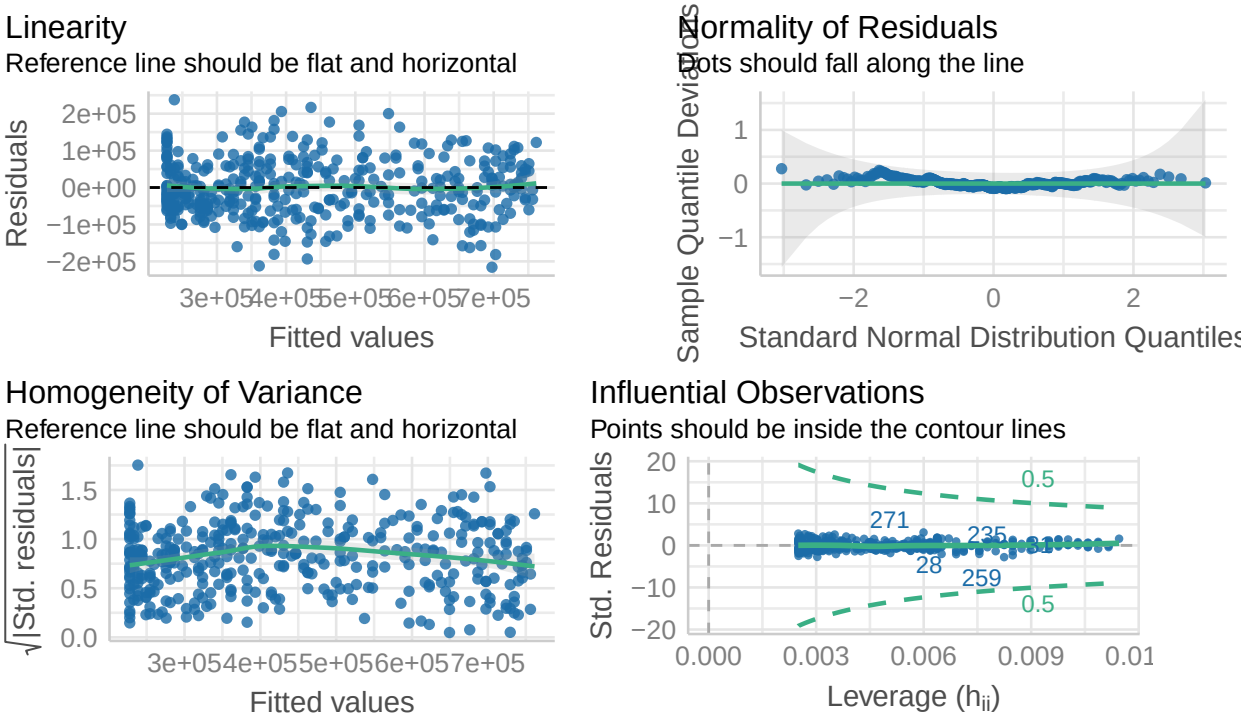
	1
Normality (Shapiro-Wilk)	0.384
Homoscedasticity (Breusch-Pagan)	0.419

Table 3: MODELO 2

	Dependent variable:
	I(api00^2)
meals	−5,337.734***
	s.e.(121.696)
	p-value: <2e-16
Constant	761,544.500***
	s.e. (8,301.866)
	p-value: <2e-16
Observations	400
R^2	0.829
Adjusted R^2	0.828
Residual Std. Error	77,573.340 (df = 398)
F Statistic	1,923.808*** (df = 1; 398); p-value: < 2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01

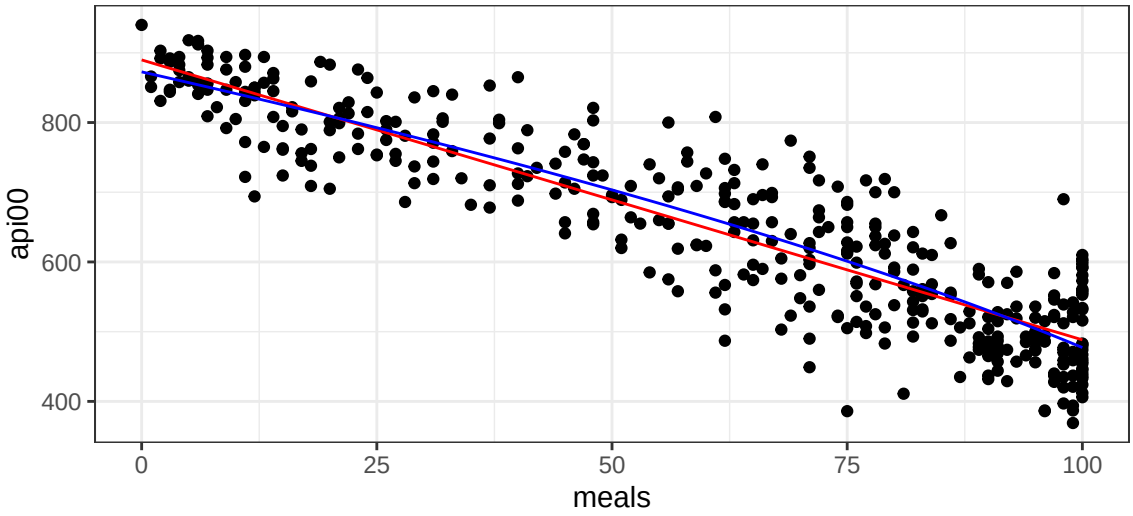
La normalidad de los errores se confirma con la prueba `Anderson-Darling normality test` con la función `ad.test` que muestra un p-value de 0.129178020851007. La homocedasticidad se confirma con la prueba `Non-constant Variance Score Test` con la función `ncvTest`, que muestra un p-value de 0.418222362546048. Finalmente, con la función de `residualPlots` obtenemos para la prueba de `Tukey test` un p value de 0.5969, por lo que no se rechaza la hipótesis nula de linealidad.

También podemos confirmar de forma gráfica estos resultados.



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 dependiente cuadrática (curva azul).



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

En el Modelo 2, se tiene un R^2 de 0.82, el cual es el coeficiente de determinación que en este caso se interpreta como que el 82% de la variabilidad del rendimiento acedémico en la escuela `api00` se explica por el modelo que incluye la variable del porcentaje de estudiantes que reciben desayuno en la esceula `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 $E(api00; meal)$. Es decir, el rendimiento acedémico en la escuela `api00` se relaciona linealmente con la variable del porcentaje de estudiantes que reciben desayuno en la esceula `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 β_1 es el parámetro estimado asociado a la variable independiente `meal`. A continuación se muestra la aprueba `Simultaneous Tests for General Linear Hypotheses`, donde se rechaza esta hipótesis nula, pues el p-valor asociado es menor a 0.05, con un nivel del confianza de 95%.

Simultaneous Tests for General Linear Hypotheses Fit:lm(formula = I(api00^2) ~ meals, data = datos7) Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 >= 0	-5337.7	121.7	-43.86	<2e-16 ***
Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
(Adjusted p values reported – single-step method)				