Introducción a la analítica

Profesores César Augusto Gómez, Mauricio Alejandro Mazo y Juan Carlos Salazar Uribe

RLM con predictores con más de dos niveles. Cuando un predictor cualitativo tiene más de dos niveles, por ejemplo k, es necesario crear $k-1$ variables dummy. El nivel sin variable dummy se conoce como el nivel de base o nivel de referencia. Considere los datos de Credit y su predictor Ethnicity, el cual tiene tres niveles: Caucasian, African American y Asian.

En este caso se pueden definir las 2 siguientes dummies:

$$x_{i1} = \begin{cases} 1 & \text{Si la persona } i \text{ es de Asia} \\ 0 & \text{Si la persona } i \text{ no es de Asia} \end{cases}$$

y

$$x_{i2} = \begin{cases} 1 & \text{Si la persona } i \text{ es Caucásica} \\ 0 & \text{Si la persona } i \text{ no es Caucásica} \end{cases}$$

Con esta dos dummies se puede obtener el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{Si la persona } i \text{ es de Asia} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{Si la persona } i \text{ es Caucásica} \\ \beta_0 + \varepsilon_i & \text{Si la persona } i \text{ es Afrodescendiente} \end{cases}$$

β_0 se puede interpretar como el balance promedio en tarjeta de crédito para Afrodescendientes, β_1 se puede interpretar como la diferencia en el balance promedio en tarjeta de crédito entre asiáticos y Afrodescendientes, y , β_2 se puede interpretar como la diferencia en el balance promedio en tarjeta de crédito entre caucásicos y afrodescendientes.

Un modelo con Ethnicity como predictor se puede ajustar por medio de:

```
library(ISLR)
library(MASS)
Credit<-read.csv(file="Credit.csv",header=T,sep=',',dec='.')
Balance=Credit[,13]
Pred_Cuali<-Credit[,c(9,10,11,12)]
Ethnicity=as.factor(Pred_Cuali[,4])
fit=summary(lm(Balance~Ethnicity))$coefficients
fit
```

```
##                     Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)        531.00000   46.31868 11.4640565 1.774117e-26
## EthnicityAsian     -18.68627   65.02107 -0.2873880 7.739652e-01
## EthnicityCaucasian -12.50251   56.68104 -0.2205766 8.255355e-01
```

El balance estimado para afrodescendientes es 531.0US. Los asiáticos tiene 18.69US menos balance que los afrodescendientes y los caucásicos tendrán 12.5US menos balance que los afrodescendientes, pero los valores-p no permiten concluir una diferencia importante en el balance promedio de acuerdo a la raza. **La prueba F es invariante a la codificación de las dummies**.

Ejemplo Backward con datos Credit.

Backward Elimination Method . Variables Removed:

X Ethnicity

X Married

X Education

X Gender

No more variables to be removed. Call: lm(formula = paste(response, "~", paste(preds, collapse = - ")), data = l)

Coefficients: (Intercept) Income Limit Rating Cards Age StudentYes
-493.7342 -7.7951 0.1937 1.0912 18.2119 -0.6241 425.6099

Extensiones del modelo lineal. El MRLM hace dos supuestos fundamentales que no necesariamente se satisfacen en la práctica:

1. La relación entre los predictores y la respuesta es aditiva. El supuesto de aditividad significa que el efecto de los cambios en un predictor X_j sobre la respuesta Y es independiente de los valores de los otros predictores.

2. La relación entre los predictores y la respuesta es lineal. El supuesto de linealidad significa que los cambios en la respuesta Y por cambio de unidad en X_j es constante sin importar el valor de X_j

Estos supuestos podrían no cumplirse en algunas aplicaciones prácticas y se deben relajar de alguna manera a fin de extender el MLRM. Por ejemplo, en la aplicación del ejemplo de advertising, el MRLM asume que el efecto en las ventas al incrementar un medio publicitario es independiente del monto gastado en el otro medio y esto podría no ser del todo cierto.

Remoción del supuesto de aditividad.En el ejemplo de Advertising, suponga que invertir en Radio incrementa la efectividad de la publicidad por TV, de manera que la pendiente para TV se debe incrementar a medida que se incrementa el gasto en publicidad por Radio. Por ejemplo, si se cuenta solo con 100000US y se gasta la mitad en TV y la otra mitad en Radio, se observa un mayor incremento en las ventas que si se gastan los 100000US solo en TV o en Radio. Esto se conoce en marketing como Sinergia y en estadística se conoce como Interacción.

En el MRLM con dos predictores X_1 y X_2, un modelo con interacción es como sigue:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

que se puede reescribir como:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon \end{aligned}$$

con $\tilde{\beta}_1 = (\beta_1 + \beta_3 X_2)$. Puesto que $\tilde{\beta}_1$ cambia (o depende) con X_2, el efecto de X_1 sobre Y ya no es constante: ajustar X_2 cambiará el impacto de X_1 sobre Y.

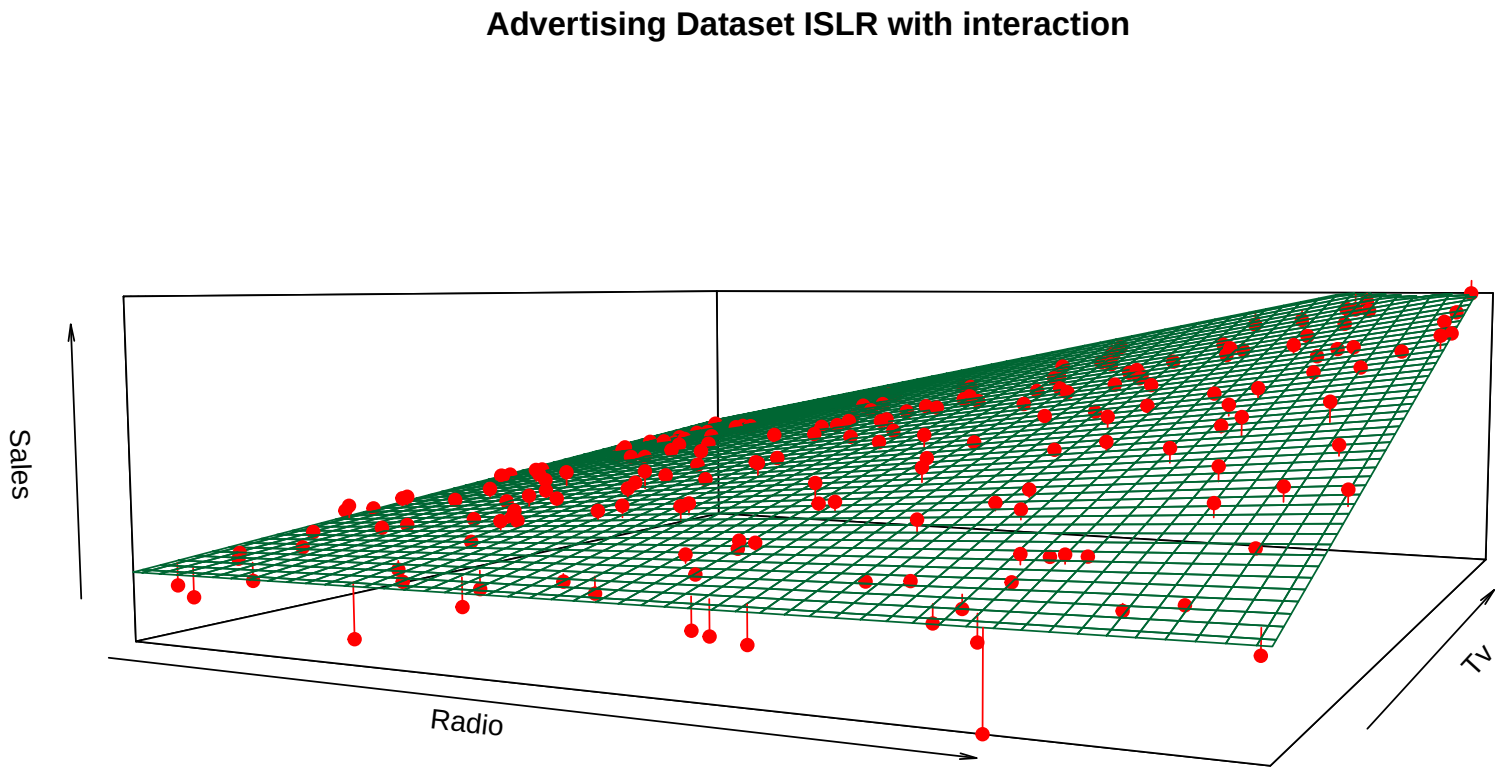
Un MRLM con interacción para los datos de Advertising considerando TV y Radio:

$$\begin{aligned} Sales &= \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 \times (TV \times Radio) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 Radio) \times TV + \beta_2 Radio + \varepsilon \end{aligned}$$

Este modelo se puede ajustar usando el software R:

```
## $Estimates
##                Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 6.750220203 0.2478713699 27.232755 1.541461e-68
## TV          0.019101074 0.0015041455 12.698953 2.363605e-27
## Radio       0.028860340 0.0089052729  3.240815 1.400461e-03
## TV:Radio    0.001086495 0.0000524204 20.726564 2.757681e-51
##
## $R2
## [1] 0.9677905
```

El gráfico del modelo ajustado:



De la tabla anterior se observa que hay mucha evidencia de un efecto importante de la interacción **lo que significa que el modelo no es aditivo**. Del $\hat{R}^2 = 0.9677905$ se observa una mejora muy notable en comparación al modelo de efectos fijos que solo contiene a TV y Radio como predictores. El modelo estimado:

$$\begin{aligned}\widehat{Sales} &= 6.75 + 0.019\,TV + 0.029\,Radio + 0.0011 \times (TV \times Radio) \\ &= 6.75 + (0.019 + 0.0011\,Radio) \times TV + 0.029\,Radio\end{aligned}$$

Un incremento en publicidad por TV de 1000US está asociado con un incremento en ventas de

$$\left(\hat{\beta}_1 + \hat{\beta}_3 \mathit{Radio}\right) \times 1000 = 19 + 1.1 \times \mathit{Radio} \quad \text{Unidades}$$

y un incremento en publicidad por Radio de 1000US estará asociado con un incremento en ventas de

$$\left(\hat{\beta}_2 + \hat{\beta}_3 \mathit{TV}\right) \times 1000 = 29 + 1.1 \times \mathit{TV} \quad \text{Unidades}$$

Notas:

- El principio de jerarquía establece que si se incluye una interacción en un modelo, se deben incluir también los efectos principales aún si sus valores-p no son importantes.
- El concepto de interacción aplica tambien para variables cualitativas o a una combinación de cualitativas y cuantitativas.

Para ilustrar este segundo punto, considere los datos de Credit. Se desea predecir Balance en términos de Income (cuantitativa) y Student (cualitativa). En ausencia de interacción:

$$\begin{aligned} Balance_i &\approx \beta_0 + \beta_1 Income_{i1} + \beta_2 Student_{i2} \\ &= \beta_1 Income_{i1} + \begin{cases} \beta_0 + \beta_2 & \text{Si la persona } i \text{ es estudiante} \\ \beta_0 & \text{Si la persona } i \text{ no es estudiante} \end{cases} \end{aligned}$$

Son dos lineas con distintos interceptos pero la misma pendiente, es decir, son rectas paralelas. Esto significa que el efecto promedio en el Balance de un incremento de una unidad en Income NO depende en si el sujeto es o no estudiante, lo cual puede ser una conclusión cuestionable.

Para formular un modelo más realista, se puede especificar un término de interacción entre Income y Student. En este caso, el modelo resultante es:

$$Balance_i = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times Income_i & \text{Si } i \text{ es estudiante} \\ \beta_0 + \beta_1 \times Income_i & \text{Si } i \text{ no es estudiante} \end{cases}$$

Estas son dos rectas para estudiantes y no estudiantes pero con distintas pendientes.

```
## $modelo1
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 211.142964 32.4572113  6.505271 2.338288e-10
## Income        5.984336  0.5566232 10.751143 7.817642e-24
## StudentYes  382.670539 65.3108082  5.859222 9.775720e-09
##
## $modelo2
##                      Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       200.623153  33.6983706  5.953497 5.789658e-09
## Income              6.218169   0.5920936 10.502003 6.340684e-23
## StudentYes        476.675843 104.3512235  4.567995 6.586095e-06
## Income:StudentYes  -1.999151   1.7312511 -1.154743 2.488919e-01
```

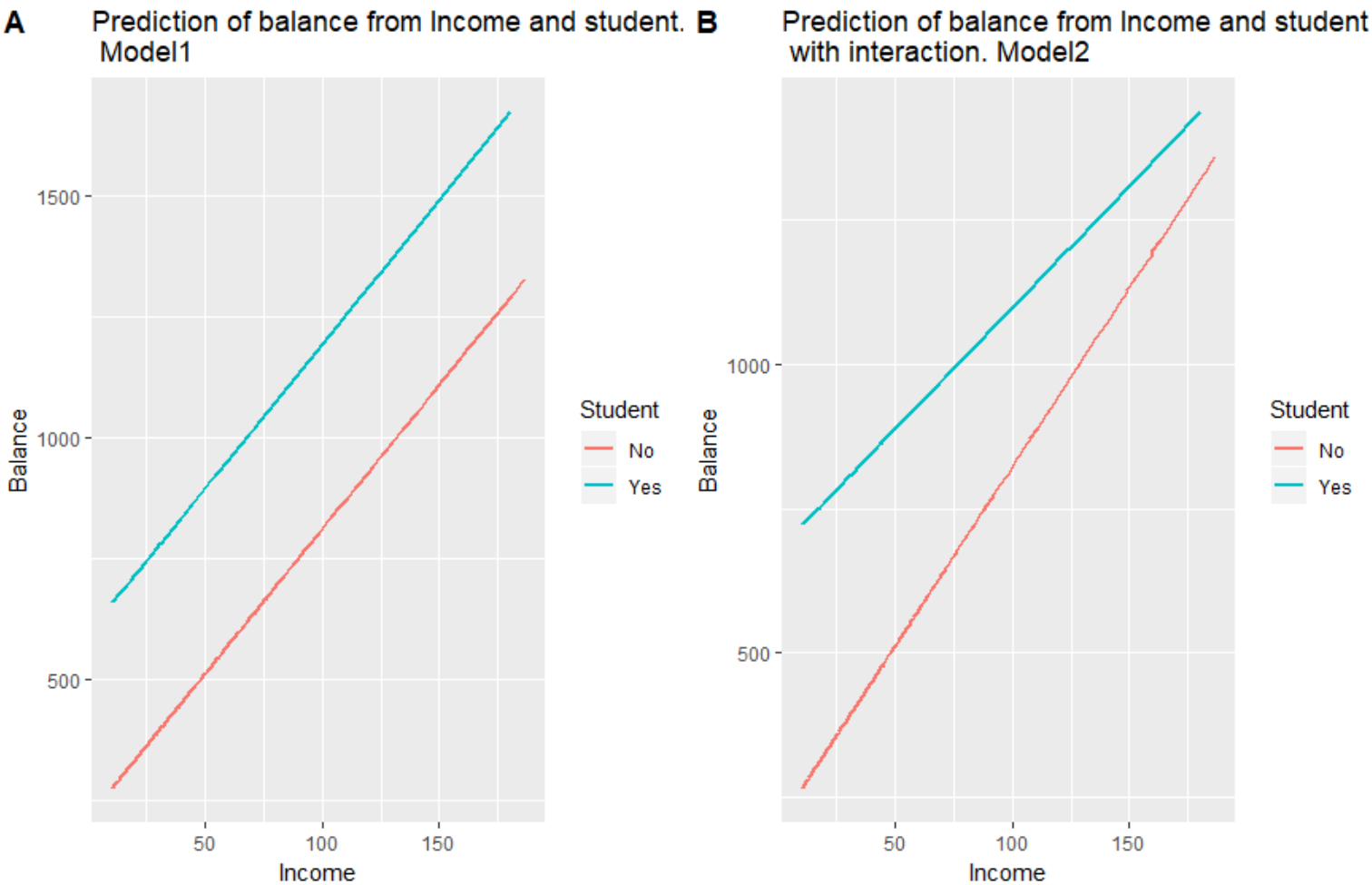
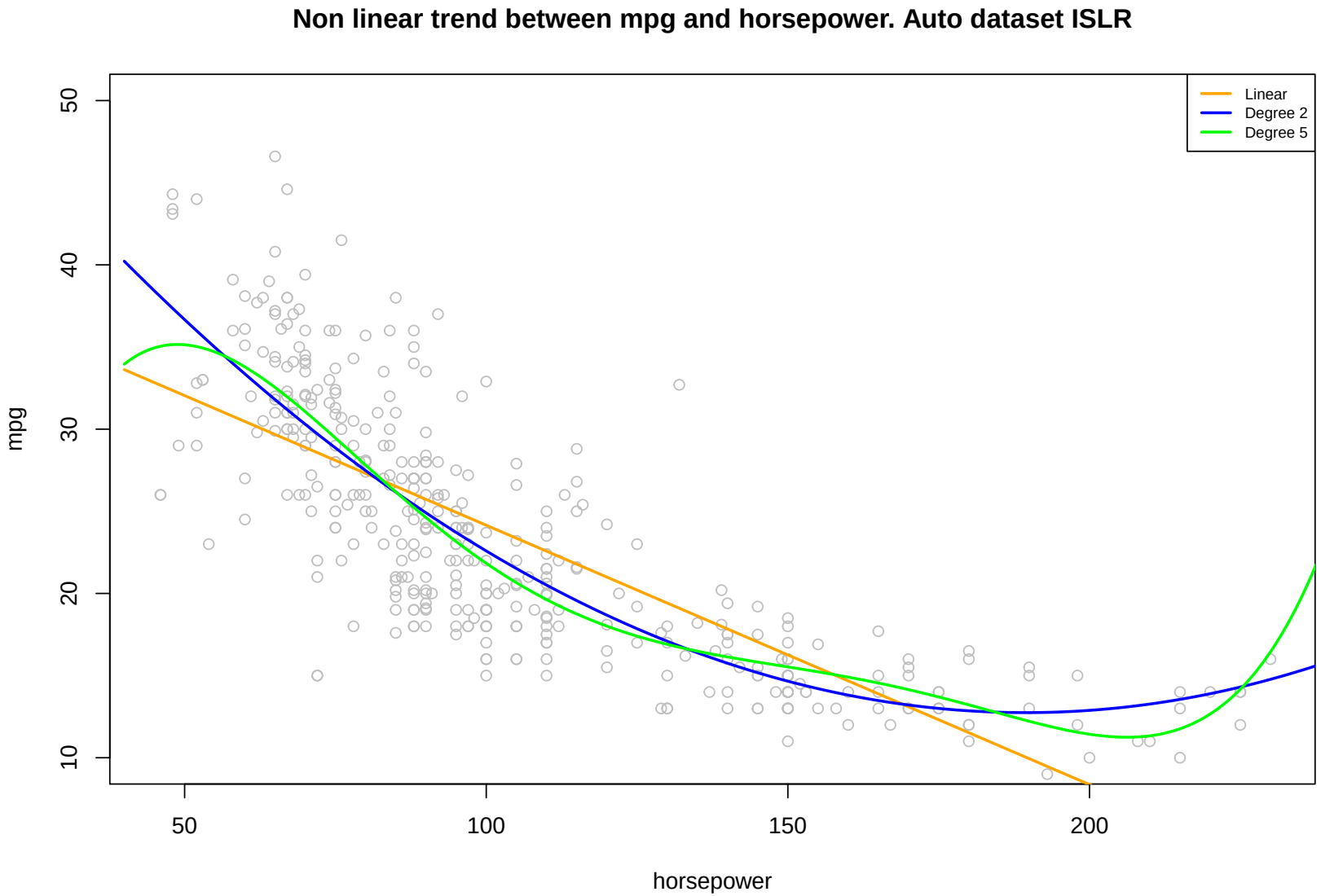


Figura 1: Interaction between a qualitative and a categorical predictor

En algunas ocasiones, la verdadera relación entre la respuesta y un predictor podría ser no lineal. Una forma simple de acomodar no linealidades es por medio de Regresión Polinomial. Considere la base de datos Auto del ISLR, que presenta las variables millas por galón (mpg. Esta es la Y) y caballos de fuerza (horsepower. Este es el predictor X). Para estos datos se ajustan varios modelos: una RLS, un polinomio de grado 2 y un polinomio de grado 5.


Non linear trend between mpg and horsepower. Auto dataset ISLR
Linear
Degree 2
Degree 5
mpg
10
20
30
40
50
50
100
150
200
horsepower

A pesar de que se observa una marcada relación entre mpg y horsepower, esta relación no parece ser líneal. La línea naranja representa un ajuste por RLS (polinomio de grado 1), la línea azul corresponde a un polinomio de grado 2 y la línea verde a un polinomio de grado 5. Estos dos últimos polinomios reconocen la no linealidad y parecen ajustar mejor que la RLS. Esto ilustra la flexibilidad de los polinomios para asociaciones no lineales.

En particular, la linea ajustada azul corresponde a un modelo no lineal de horsepower pero sigue siendo un modelo lineal y por lo tanto se puede ajustar con software estándar que ofrezca OLS:

$$\begin{aligned}\widehat{mpg} &= \widehat{\beta}_0 + \widehat{\beta}_1 \times horsepower + \widehat{\beta}_2 \times horsepower^2 \\ &= 23.45 - 120.14 \times horsepower + 44.09 \times horsepower^2\end{aligned}$$

```
## $degree1
## [1] 0.6059483
##
## $degree2
## [1] 0.687559
##
## $degree5
## [1] 0.696739
```

El R^2 del modelo lineal simple es 0.606. Si se incorpora el término cuadrático $horsepower^2$ hay un incremento en el R^2 que es ahora igual a 0.688 y si se incorpora el término de grado 5, $horsepower^5$, el R^2 llega a 0.70. Entonces ¿porqué no ajustar polinomios de mayor grado? pues esto conduciría a sobreajuste y se viola el principio de parsimonia. Observe que el polinomio de grado 5 es moderadamente rugoso y no parece mejorar notablemente el ajuste.

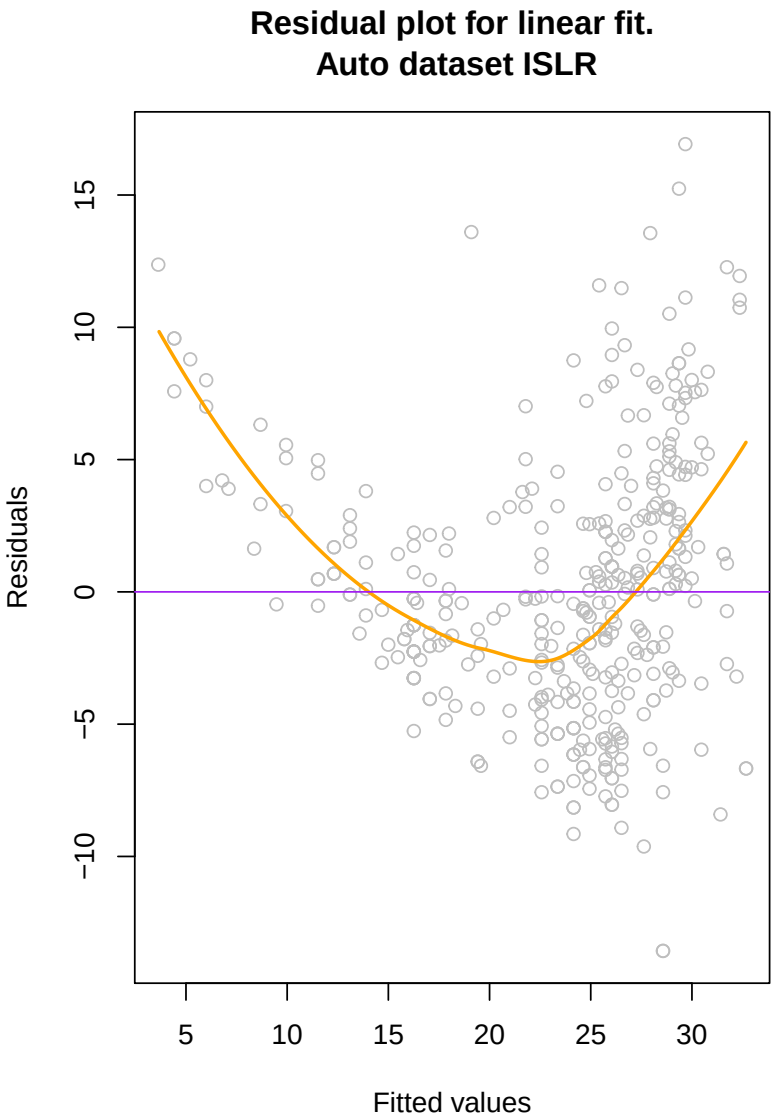
Problemas potenciales con modelo lineales:

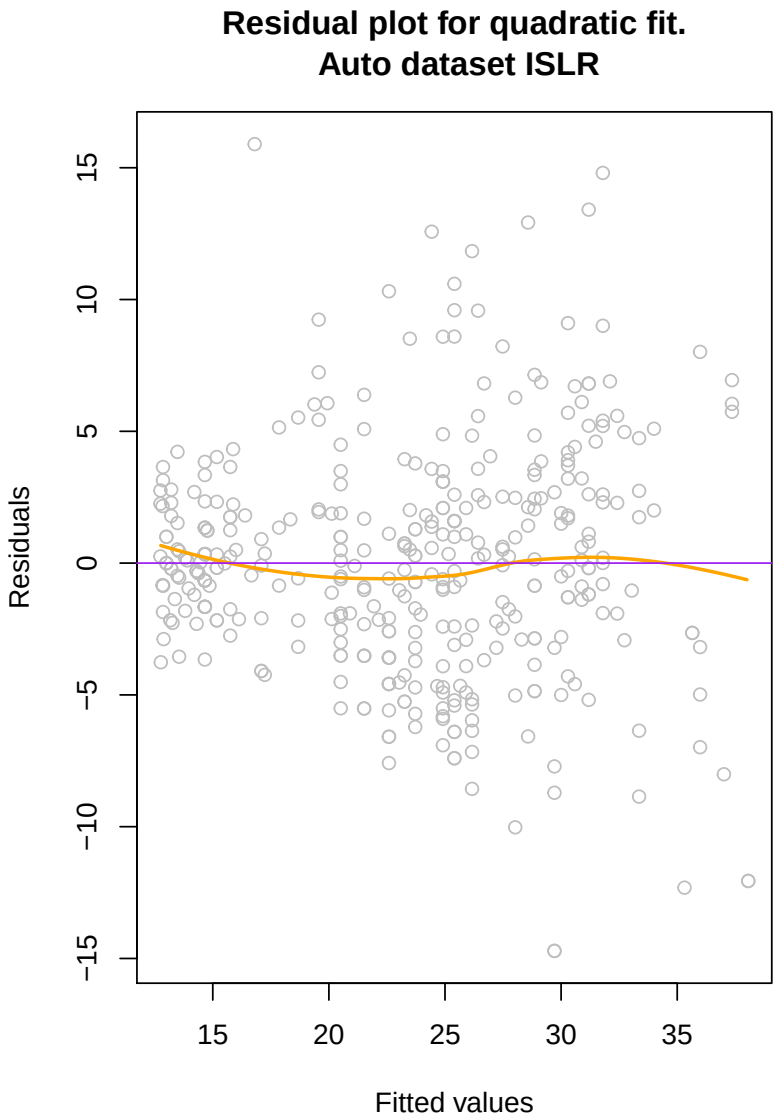
- No linearidad de la relación respuesta/predictores.
- Correlación de los términos de error.
- Varianza no constante de los errores.
- Outliers.
- Puntos con 'Leverage'[1] alto.
- Colinealidad.

[1] Leverage es una medida de qué tanto se aleja una observación de un predictor(por ejemplo x_{ki}) de la media de esa variable ($\bar{x}_k = \sum_{i=1}^{n} x_{ki}$).

- **No linearidad de la relación respuesta/predictores**. El modelo de RLS asume que hay una relación respuesta/predictor que se puede explicar con una línea recta. Si esto no se cumple no es buena idea concluir con dicho modelo. Hay métodos para evaluar qué tan bien o no ajusta un modelo lineal. Se presentan algunos.

Gráfico de residuales. Son una herramienta gráfica útil para identificar no linealidad. Consiste en gráficar los residuales $\hat{e}_i = y_i - \hat{y}_i$ versus el predictor x_i. En el caso de RLM se gráfican versus $\hat{y}_i$. Idealmente, el gráfico de residuales debe lucir aleatorio alrededor de cero. Alejamientos de este patrón aleatorio se asocian con problemas de algún aspecto del modelo.





El modelo lineal muestra una marcada tendencia no lineal, el modelo cuadrático exhibe un mejor ajuste y ningún patrón importante en su gráfico de residuales.

Si el gráfico de residuales indica que hay asociaciones no lineales en los datos, entonces una solución sencilla consiste en usar transformaciones no lineales de los predictores, tales como $\log X$, $\sqrt{X}$ y X^2 en el modelo de regresión. Sin embargo, y como se verá en este curso existen otras formas no lineales más avanzadas para enfrentar este problema.

- **Correlación de los términos de error**. Un supuesto del MRL es que los términos de error son incorrelacionados. Pero ¿qué significa que sean incorrelacionados? Por ejemplo, si los errores son incorrelacionados, entonces el hecho de que e_i sea positivo no proporciona mucha información acerca del signo de e_{i+1}. Los errores estándar que se calculan para los coeficientes de regresión estimados o para los valores ajustados se basan en el supuesto de que los errores son incorrelacionados.

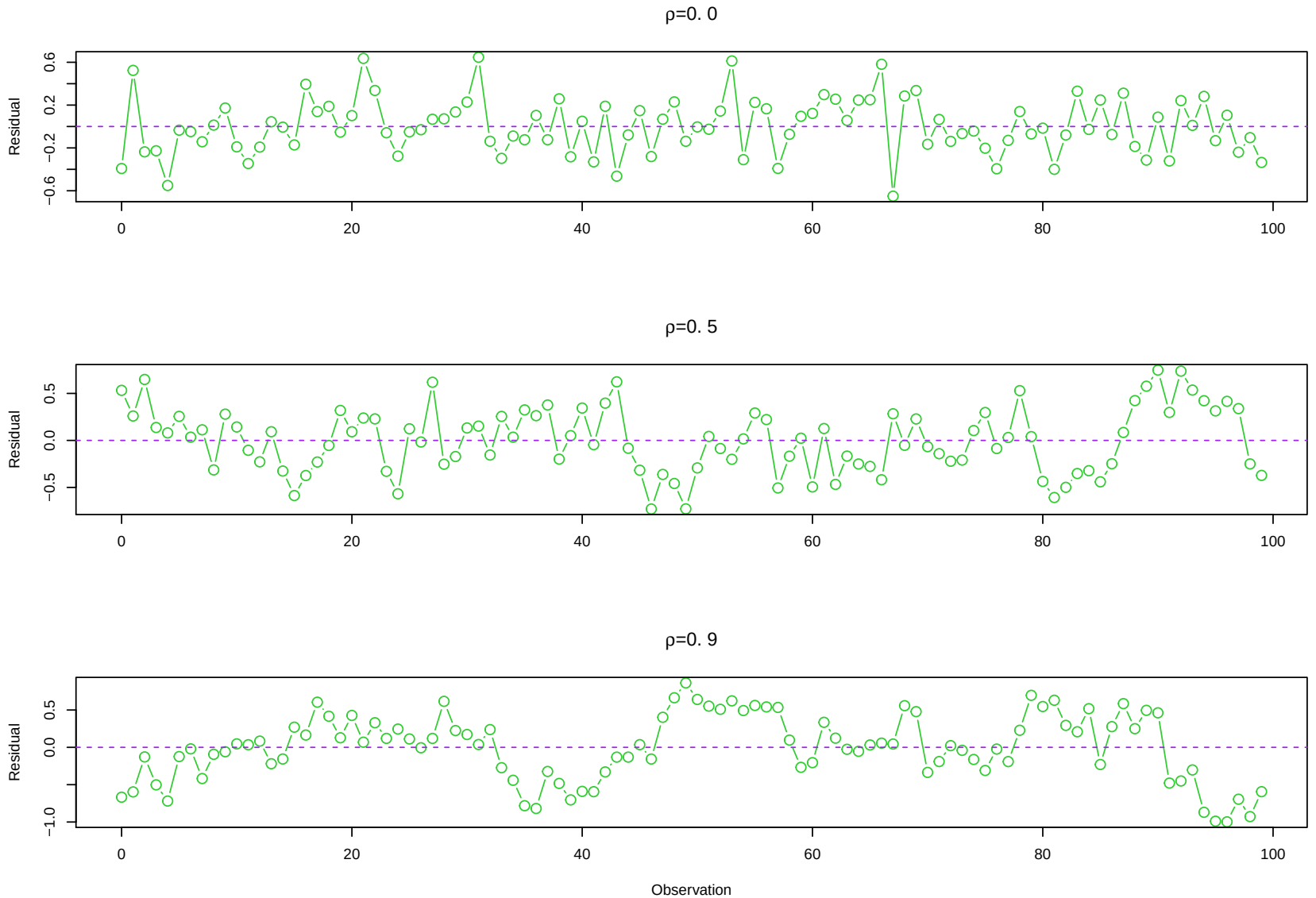
Si hay correlación entre estos términos de error, el error estándar de los estimadores tenderá a subestimar los verdaderos errores estándar. Como resultado de esto, los IC y los de predicción serán más angostos de lo que deberían ser. Los valores-p también serán menores de lo que deberían ser y así conducir a inferencia errónea.

Suponga que accidentalmente se doblan los datos, de manera que se tendrán observaciones y términos de error por pares. Si se ignora esto, los errores estándar estarán basados en una muestra, no de tamaño n, sino de tamaño $2n$. Las estimaciones serán las mismas para las $2n$ muestras que para las n muestras, pero los IC serán más angostos por un factor cercano a $\sqrt{2}$. Se comprobará esta afirmación con los datos de Auto dataset.

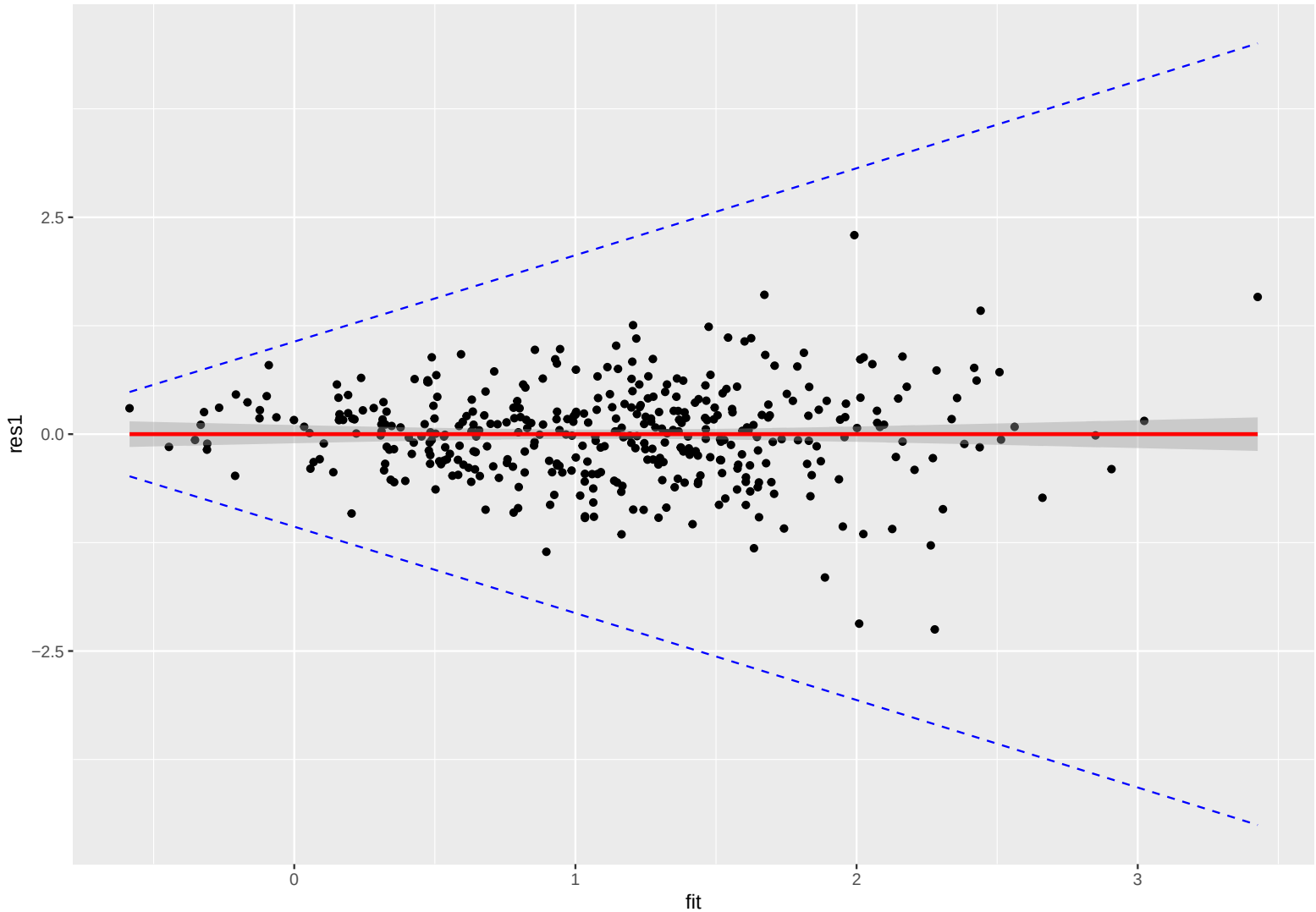
```
## [1] 392   9

## [1] 784   9

##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 39.9358610 0.717498656  55.65984 1.220362e-187
## horsepower  -0.1578447 0.006445501 -24.48914  7.031989e-81

##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 39.9358610 0.506698967  78.81575  0.00000e+00
## horsepower  -0.1578447 0.004551825 -34.67724 2.66037e-160

## $IC_n
##                  2.5 %     97.5 %
## (Intercept) 38.525212 41.3465103
## horsepower  -0.170517 -0.1451725
##
## $IC_2n
##                 2.5 %     97.5 %
## (Intercept) 38.94121 40.9305122
## horsepower  -0.16678 -0.1489095

## [1] 2.82129846 1.98930237 0.02534455 0.01787049
```

Note que $1.98930237 \times \sqrt{2} = 2.8133 \approx 2.821298$ y $0.01787049 \times \sqrt{2} = 0.0253 \approx 0.02534455$, lo cual confirma la afirmación anterior.

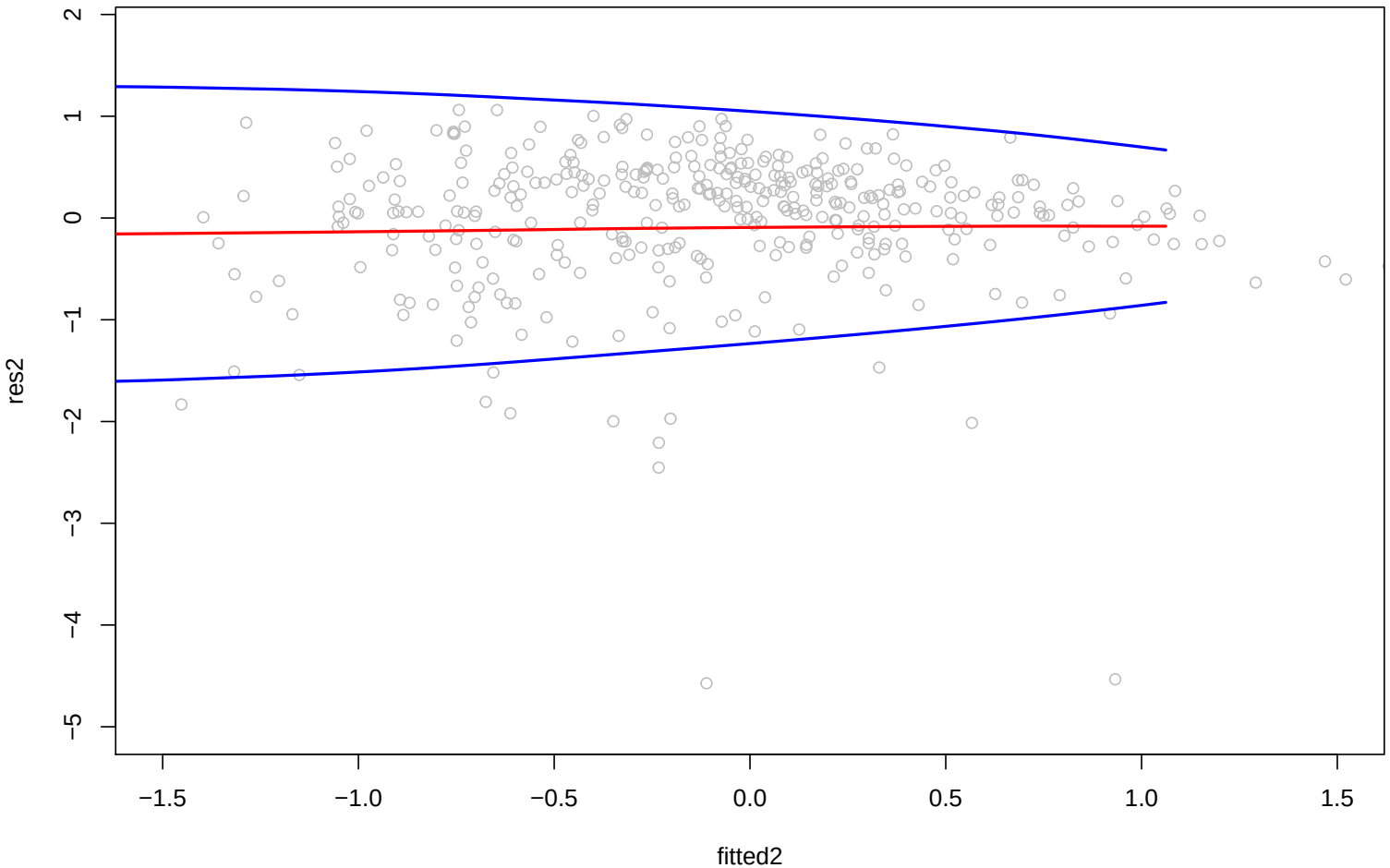
Para evaluar si los errores son o no incorrelacionados se pueden graficar los residuales del modelo en función del tiempo. Si son incorrelacionados entonces no debe haber un patrón discernible. Por otra parte, si los residuales están positivamente correlacionados, se observará que los residuales adyacentes tendrán valores similares (fenómeno conocido como Tracking). En algunas ocasiones, estas correlaciones se pueden controlar desde la etapa del diseño experimental. El siguiente gráfico proporciona una ilustración usando datos simulados:


ρ=0. 0
Residual
−0.6
−0.2
0.2
0.6
0
20
40
60
80
100
ρ=0. 5
Residual
−0.5
0.0
0.5
0
20
40
60
80
100
ρ=0. 9
Residual
−1.0
0.0
0.5
0
20
40
60
80
100
Observation

- **Varianza no constante de los errores**. En ocasiones, la varianza de los términos de error se puede incrementar con el valor de la respuesta (fenómeno conocido como heterocedasticidad) y producir un patrón en forma de embudo en el gráfico de residuales.

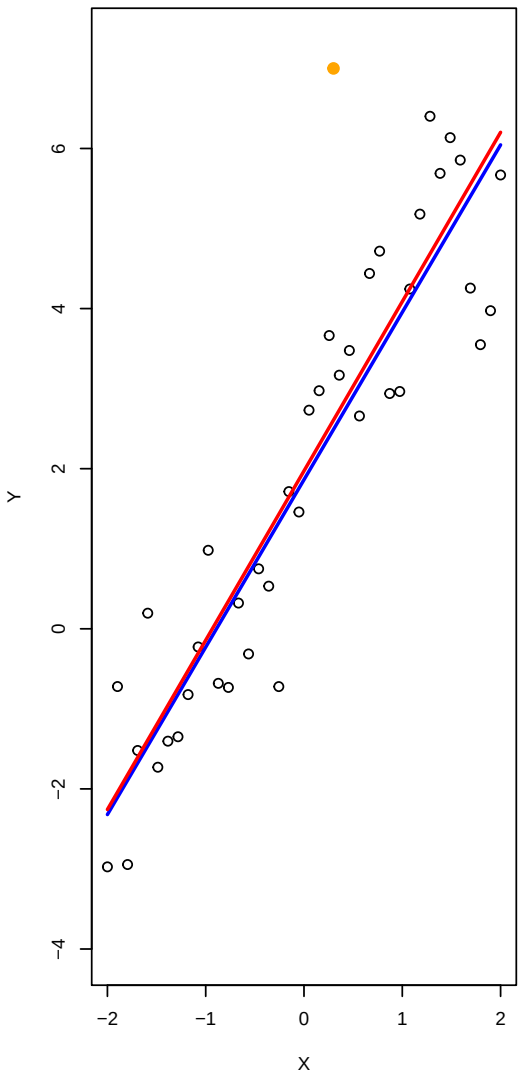


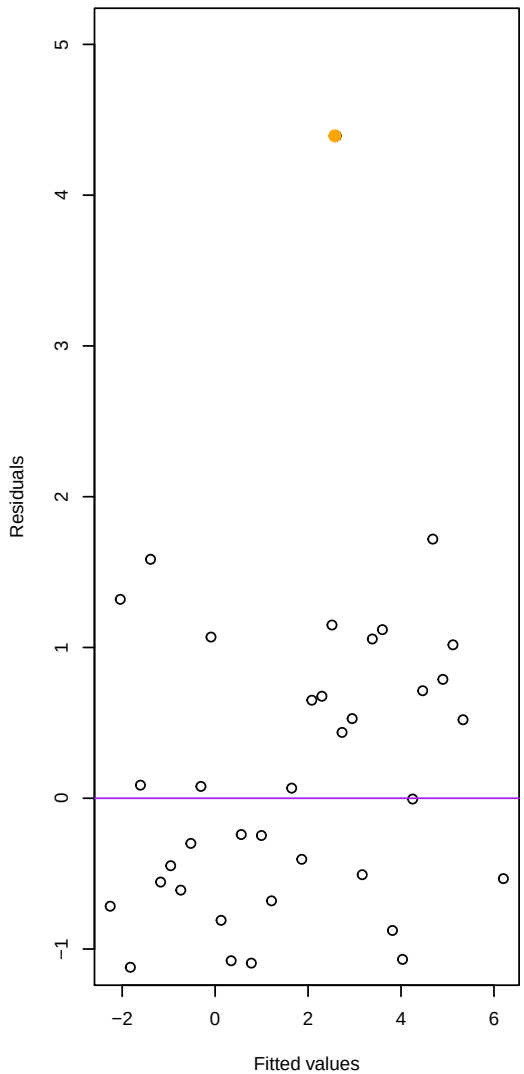
- **Varianza no constante de los errores**. Una transformación, por ejemplo con logaritmo, puede estabilizar la varianza

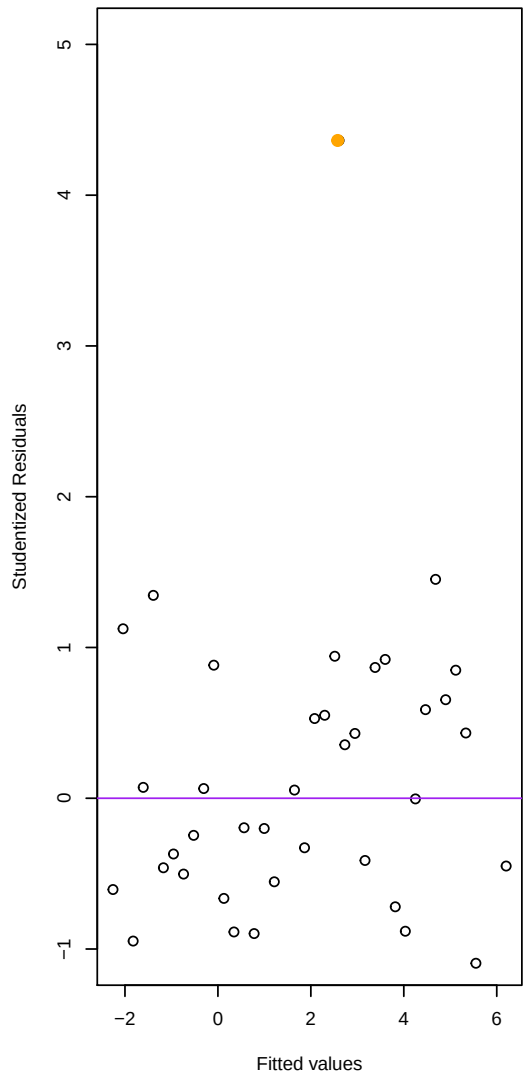


- **Outliers**. Un outlier es un punto para el cual y_i está lejos del valor predicho por el modelo. Un ejemplo de un outlier se presenta a continuación. Observe que el outlier no tiene casi ningún efecto en la rectas OLS, pero si se comparan los RSE del modelo sin outlier y el modelo con outlier, que son respectivamente 1.02 y 1.234 sí se observa un cambio importante.

```
##  RSE_SIN  RSE_CON
## 1.020301 1.233851
```







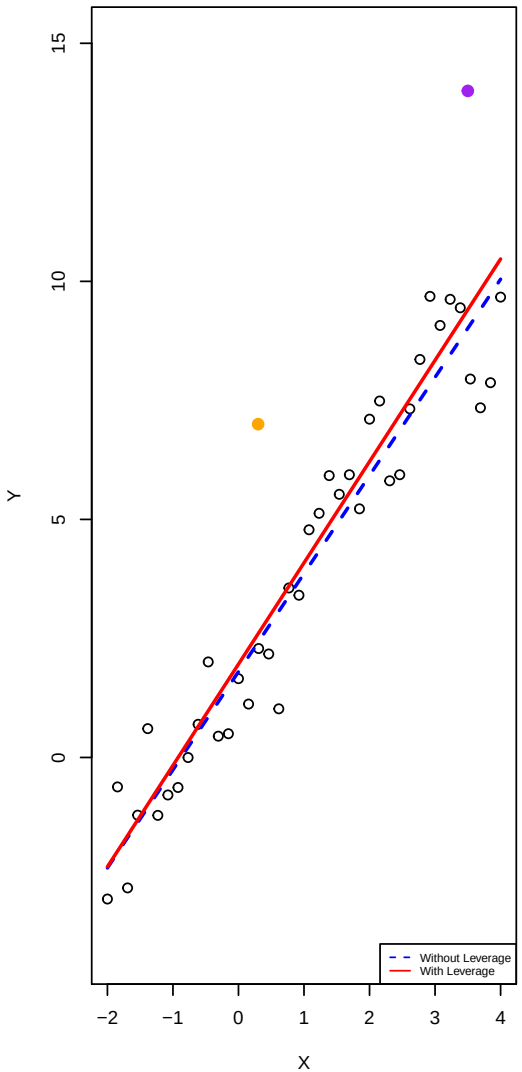
Puesto que el RSE se usa para calcular todos los IC y los valores-p, este cambio causado solo por un solo outlier puede comprometer seriamente la interpretación del ajuste. También, los R^2 se afectan, pues la inclusión del outlier hace que este decline de 0.8611 a 0.8088.

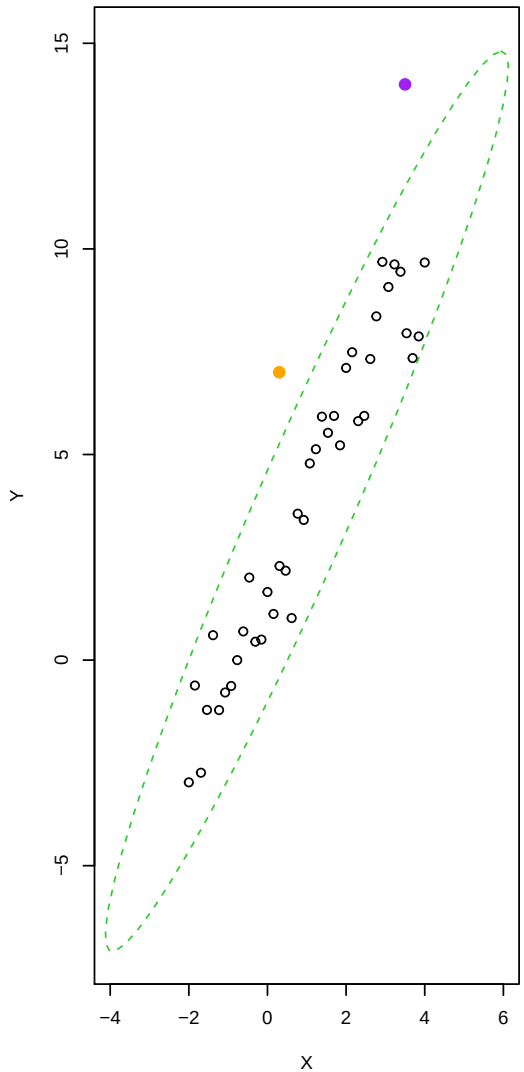
Los gráficos de residuales y residuales estudentizados versus valores ajustados permiten detectar el outlier. De hecho, valores cuyo residual estudentizado es mayor a 3 en valor absoluto son posibles outliers. En el gráfico anterior, el outlier tiene un residual mayor a 4 mientras los otros están dentro de la franja definida por -2 y 2.

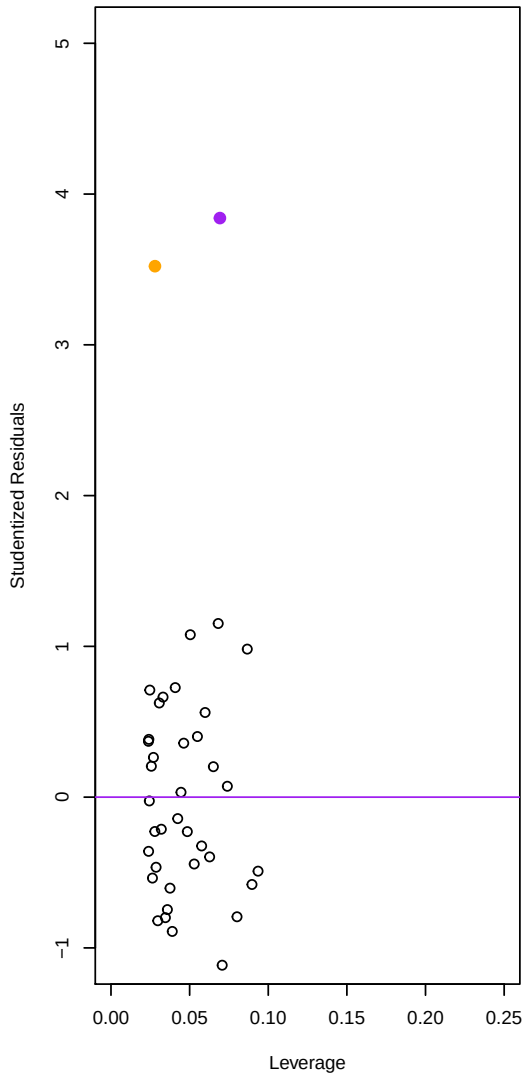
Si un outlier es debido a un error, se puede remover de la base de datos pero un outlier también puede indicar una deficiencia en el modelo, por ejemplo la falta de un predictor importante en el modelo. Pero también pueden ser manifestaciones importantes de una característica, por ejemplo podrían asociarse a personas muy fuertes o muy rápidas o muy inteligentes. Por esto, se deben manejar con sumo cuidado.

- **Puntos con 'Leverage' alto**. Outliers son observaciones con y_i inusual dado un predictor x_i. En contraste, observaciones con leverage alto tiene un valor inusual para x_i (muy alto o muy bajo). Considere los datos simulados anteriores pero con la adición de un punto con alto leverage $(x = 3.5, y = 14)$. Observe que en el panel del centro se observan el outlier y el punto con leverage alto. Claramente, el punto con leverage alto se aleja del conglomerado definido por el resto de datos. Se pudiera también argumentar que el outlier no es de alto leverage al estar su x_i cercano al resto de los x's.

REGRESIÓN LINEAL MÚLTIPLE

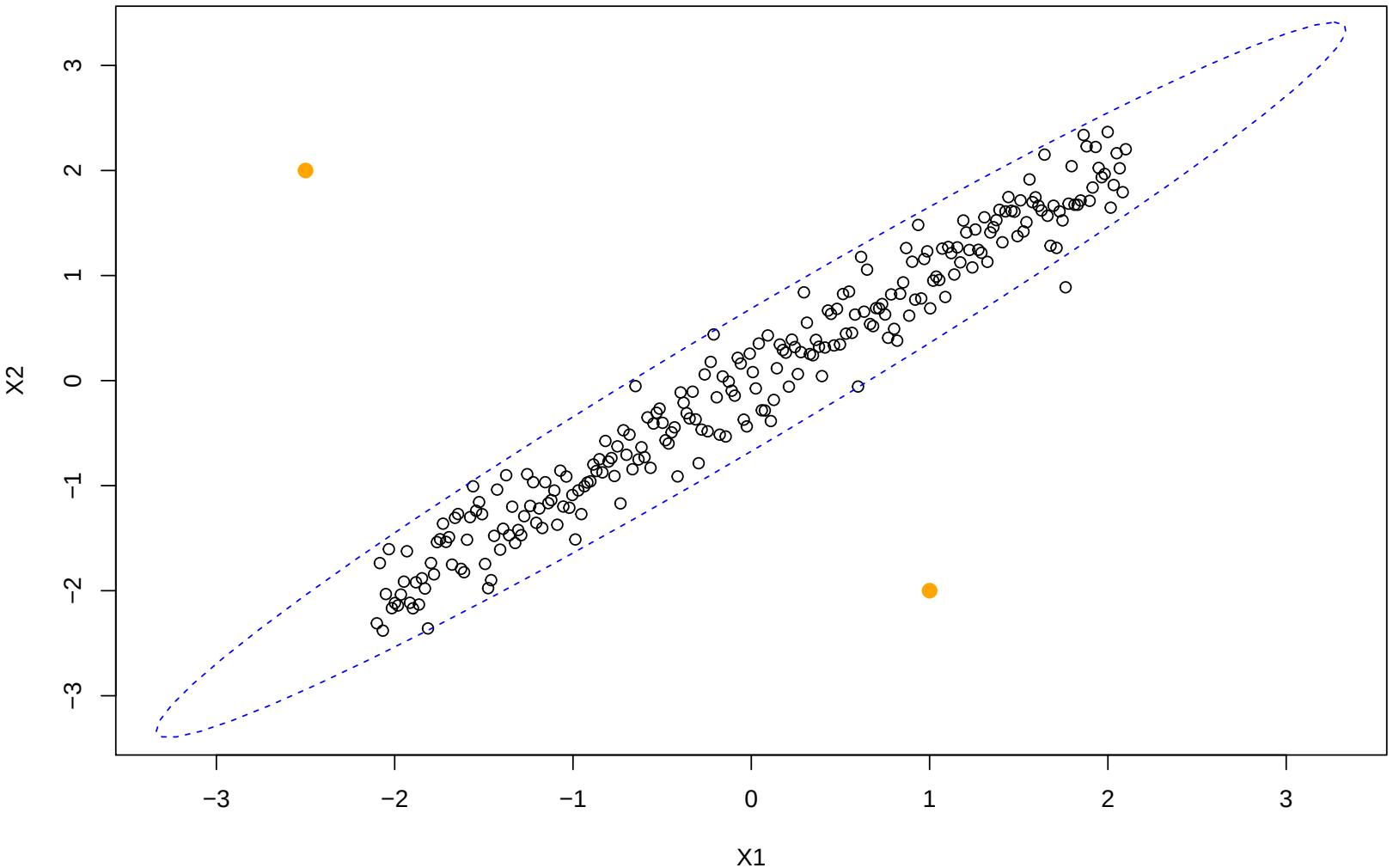






En general, observaciones con leverage alto tienden a tener un impacto más notable en la recta OLS. De hecho tiene más impacto que los outliers. También, observaciones con leverage alto, pueden llegar a invalidar la totalidad del ajuste. Por estas razones es importante identificar observaciones con leverage alto. En RLS es relativamente sencillo identificarlos: se buscan observaciones para las cuales el valor del predictor está fuera del rango normal de las observaciones.

Pero en el caso de RLM, es posible tener una observación que esté dentro del rango de observaciones individuales pero inusual en términos del conjunto total de predicciones. Considere un conjunto de datos con dos predictores X_1 y X_2. Observe en el siguiente gráfico que muchas de las observaciones caen dentro de la elipse azul punteada, mientras dos puntos (color naranja) están bien por fuera de este rango pero sus valores para X_1 y X_2 no son inusuales. Esto implica que si se examina solo a X_1 o solo a X_2 no se identificarán esos puntos con leverage alto.


X2
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
X1