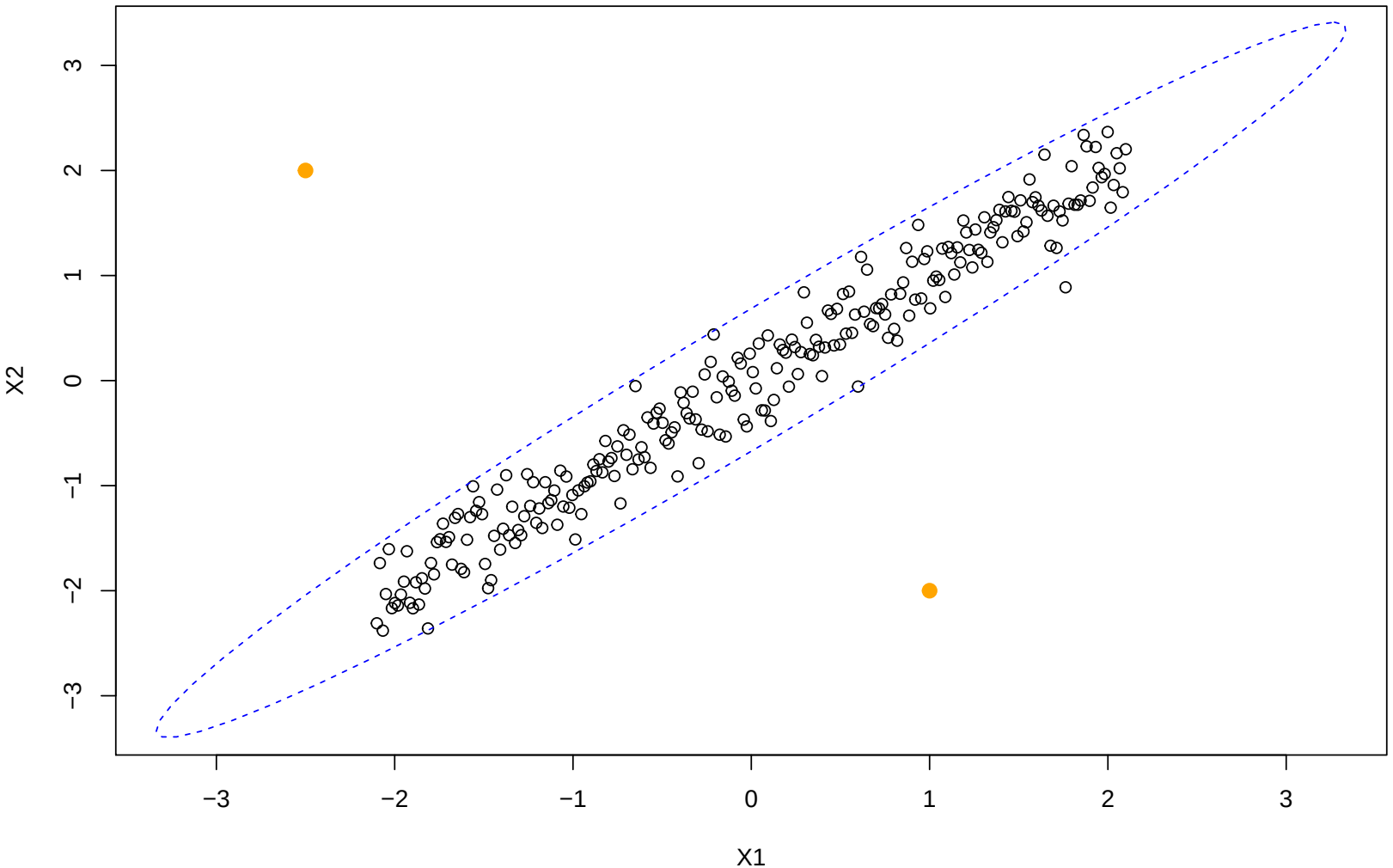
Introducción a la analítica

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

REGRESIÓN LINEAL MÚLTIPLE



Este problema se acentúa a medida que aumenta el número de predictores en el modelo ya que no hay una forma sencilla de considerar en un gráfico todas las dimensiones simultáneamente. Sin embargo, es posible cuantificar el leverage de una observación usando el Estadístico Leverage. Para el MRLS, el estadístico leverage se define como:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}$$

En esta expresión se observa que h_i se incrementa con la distancia de x_i a $\bar{x}$.

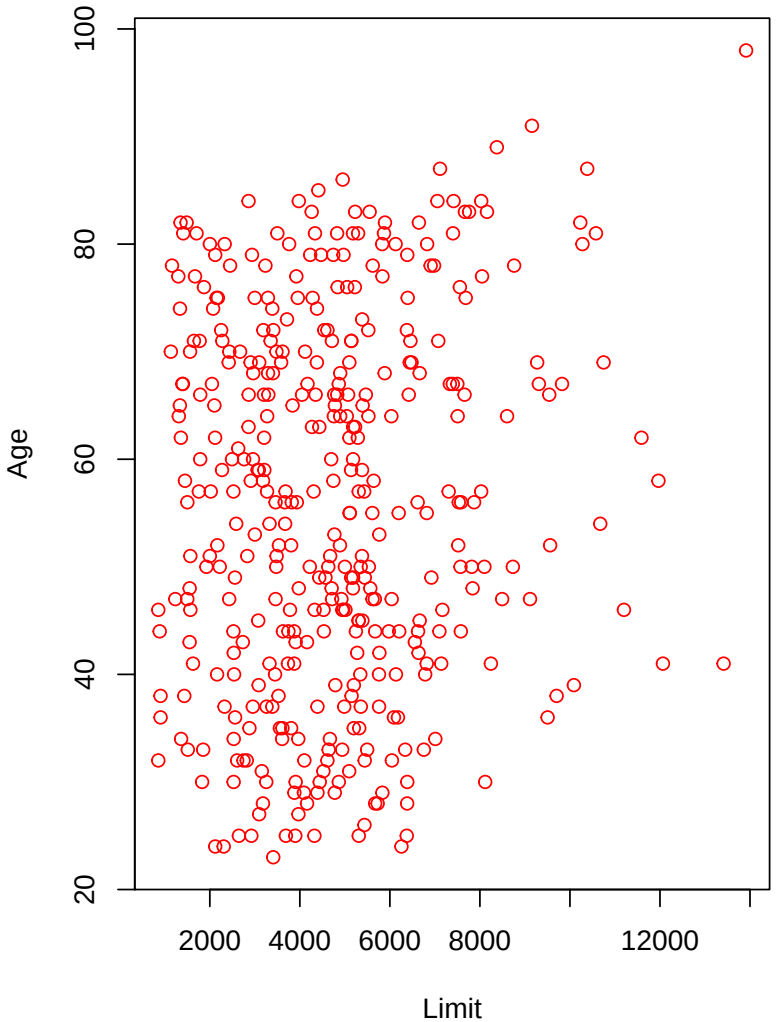
Para el caso de RLM el i-ésimo leverage es el elemento de la diagonal de la matriz hat $h_{ii} = Diag(\mathbf{H})$:

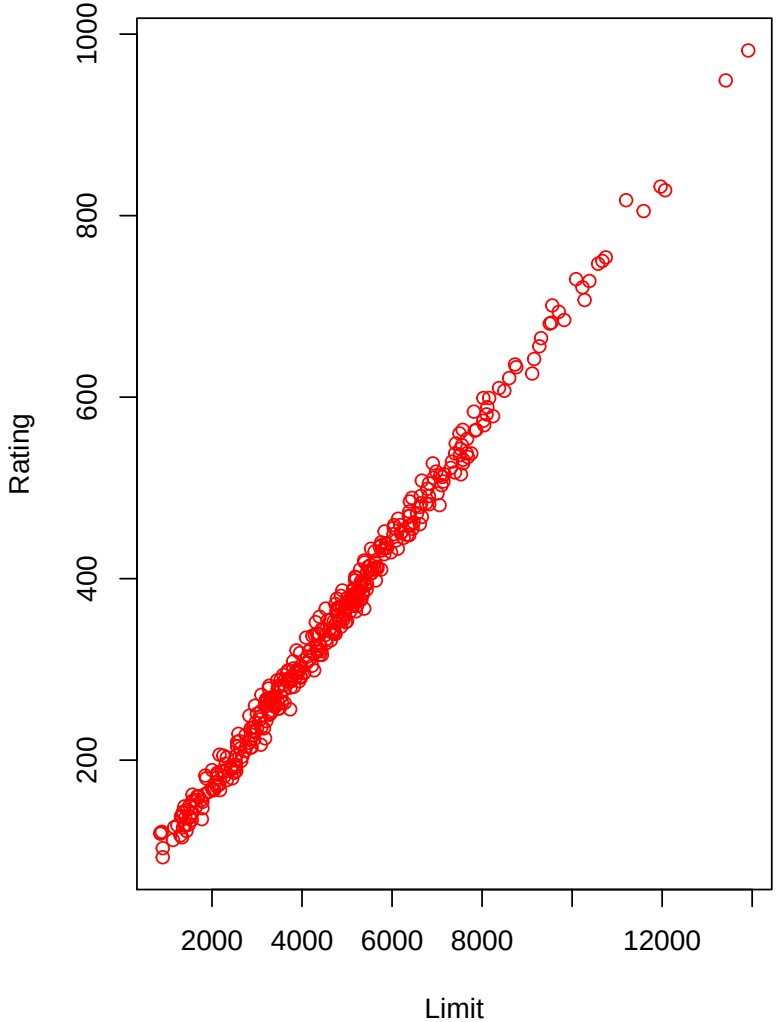
$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}$$

donde $\mathbf{X}$, es la matriz de diseño (cuyas filas corresponden a las observaciones y cuyas columnas corresponden a las variables independientes o explicativas). El estadístico leverage siempre está entre $1/n$ y 1 y el leverage promedio para todas la observaciones siempre es igual a $(p+1)/n$ (Ejercicio). Entonces, si una observación específica tiene un estadístico leverage que excede por mucho a $(p+1)/n$ entonces se puede pensar que tiene un leverage alto. Combinación de outlier y leverage alto no es conveniente para ajustar MLR.

- **Colinealidad**. La *colinealidad* se refiere a la situación en la cual dos o más predictores están muy relacionados el uno con el otro. Considere el siguiente gráfico obtenido usando los datos de Credit. Limit y Age no parecen estar relacionados mientras que Limit y Rating si están muy relacionados entre sí.

REGRESIÓN LINEAL MÚLTIPLE

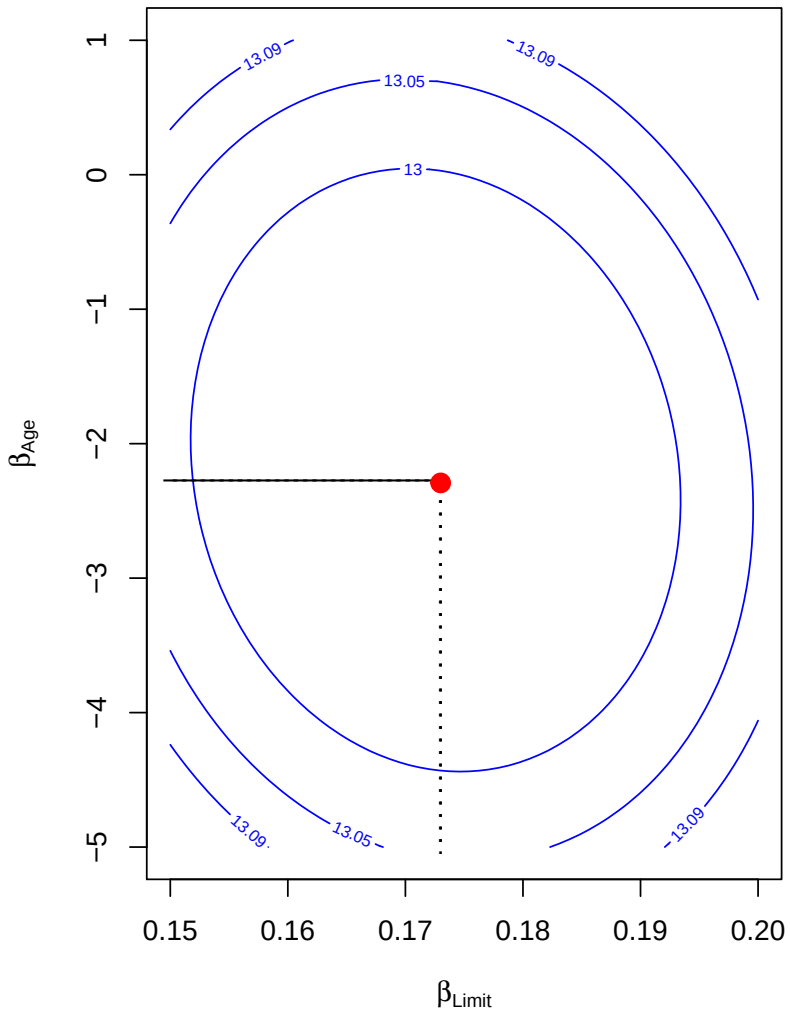


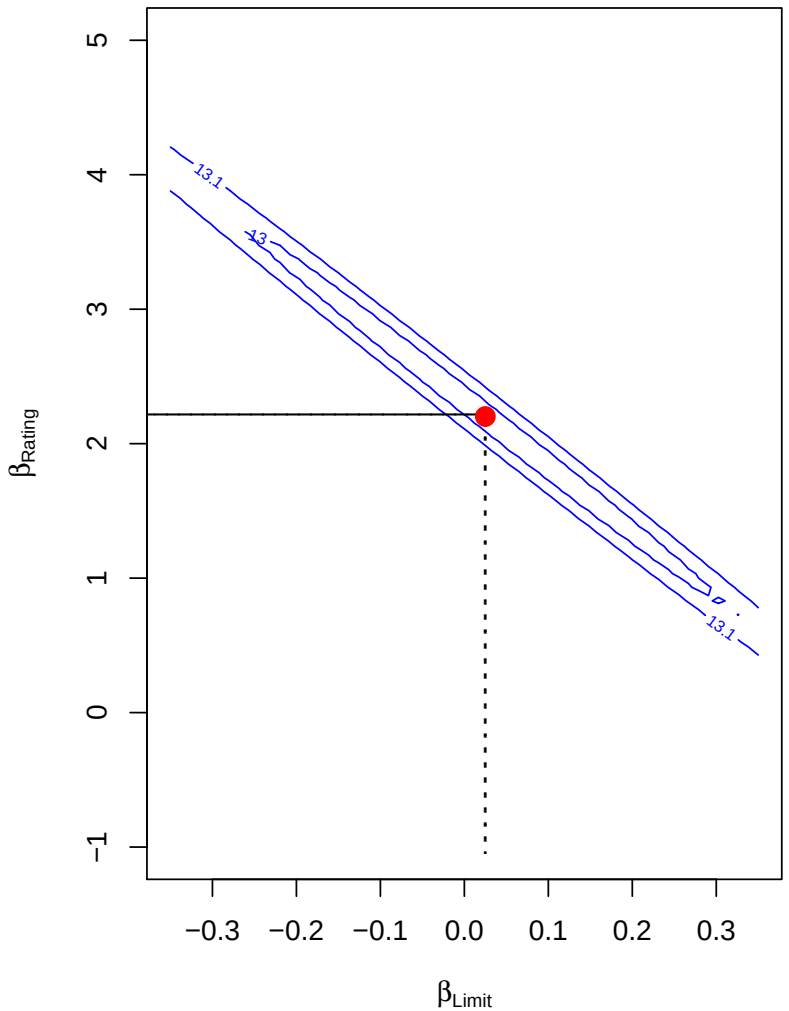


La colinealidad presenta problemas en RLM puesto que puede ser difícil separar los efectos individuales de variables colineales sobre la respuesta. Es decir, puesto que Limit y Rating tienden a crecer o decrecer juntas puede ser difícil medir cómo cada una de ellas se asocia por separado con la respuesta Balance.

```
## Warning: package 'ISLR' was built under R version 3.6.3

## Warning: package 'latex2exp' was built under R version 3.6.3
```





El panel izquierdo muestra la solución OLS en rojo. Por ejemplo, el verdadero valor para el coeficiente de Limit es casi fijo que estará entre 0.15 y 0.20. En contraste, en el panel derecho, el coeficiente para Limit estará entre -0.3 y 0.3 el cual es un rango de valores mucho más amplio que el del panel izquierdo. Esta incertidumbre es causada por la colinealidad entre Limit y Rating.

Otro efecto de la colinealidad se observa en la siguiente tabla, donde en ausencia de Rating, Limit es altamente importante (valor-p$<$0.0001) pero en presencia de Rating, Limit ya no es importante (valor-p=0.7012). La importancia de la variable Limit está enmascarada u oculta por la colinearidad con Rating (efecto confusor de la variable Rating).

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -173.411     43.828  -3.957    0.000
## Age           -2.291      0.672  -3.407    0.001
## Limit          0.173      0.005  34.496    0.000

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -377.537     45.254  -8.343    0.000
## Rating         2.202      0.952   2.312    0.021
## Limit          0.025      0.064   0.384    0.701
```

¿Cómo detectar colinealidad para evitar el reporte de asociaciones espúreas? Una forma simple de detectar colinealidad es a través de la matriz de correlación de los predictores. Un elemento de esta matriz que sea grande en valor absoluto indica un par de predictores altamente correlacionados y por ende un potencial problema de colinealidad en los datos.

Pero, infortunadamente, no todos los problemas de colinealidad se pueden detectar a través de la inspección de la matriz de correlación, pues es posible que haya correlación entre tres o más variables, aún si ningún par de predictores tenga, particularmente, una alta correlación. Este fenómeno se conoce como *Multicolinealidad*. En vez de inspeccionar la matriz de correlación de los predictores, una mejor forma de evaluar la multicolinealidad es calculando el **Factor de Inflación de la Varianza (conocido como el VIF)**.

El VIF es la razón de la varianza de β_j, cuando se ajusta el modelo completo (o full) dividida por la varianza de β_j cuando se ajusta el modelo solo con la variable X_j. El menor valor posible del VIF es 1, el cual indica la completa ausencia de colinealidad. Típicamente en la práctica hay un monto pequeño de colinealidad entre los predictores. Como una regla de dedo, un valor VIF que va de 5 a 10 indica un monto problemático de colinealidad.

El VIF se define formalmente como:

$$VIF(\beta_j) = \frac{1}{1 - R^2_{X_j \mid X_{-j}}}$$

donde $R^2_{X_j \mid X_{-j}}$ es el R^2 de una regresión de X_j en todos los otros predictores; por ejemplo, si $j = 1$, esta regresión sería:

$$X_1 = \alpha_0 + \alpha_2 X_2 + \cdots + \alpha_p X_p + \varepsilon$$

y si $j \neq 1$:

$$X_j = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \cdots + \alpha_p X_p + \varepsilon$$

Si $R^2_{X_j \mid X_{-j}}$ está cerca de 1, habrá muy posiblemente colinealidad y por lo tanto el VIF será grande.

Si se detecta colinealidad ¿qué se puede hacer? Hay dos soluciones simples:

1. Elimine una de las dos variables problemáticas de la regresión. Esto usualmente se puede hacer sin comprometer seriamente el ajuste puesto que la presencia de colinealidad implica que la información que esta variable proporciona acerca de la respuesta es redundante en presencia de las otras variables.
2. Combine las variables colineales en un solo predictor, por ejemplo un promedio estandarizado puede generar un nuevo predictor.

(Ejercicio: Leer seccion 3.4 del texto ISLR)

Comparación entre regresión lineal y Knn. - RL es un ejemplo de una aproximación paramétrica ya que asume una forma lineal funcional para $f(X)$.

- Los métodos paramétricos tienen algunas ventajas: fáciles de ajustar, sus coeficientes son relativamente fáciles de interpretar y permite implementar test estadísticos de manera sencilla
- Los métodos paramétricos tienen algunas desventajas: supuestos muy fuertes en relación a la forma de $f(X)$ y tienen en general poca flexibilidad.

REGRESIÓN LINEAL MÚLTIPLE VS REGRESIÓN KNN

- Por su parte, los métodos no paramétricos no asumen de manera explicita una forma paramétrica para $f(X)$ y por lo tanto son más flexibles en regresión.
- Sin embargo, son poco interpretables pero excelentes para predecir en muchos casos.
- En este módulo se verá un método no paramétrico conocido como **Regresión de k vecinos más cercanos (Knn Regression)**, el cual está muy relacionado con el clasificador Knn que ya se discutió.

- Dado un valor para K y un punto x_0 para predecir, la regresión Knn primero identifica las K observaciones de entrenamiento que están más cercanas a x_0 (conjunto $\mathcal{N}_0$).
- Luego, Knn estima $f(x_0)$ usando el promedio de todas **las respuestas** de entrenamiento en $\mathcal{N}_0$. En otras palabras:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

La siguiente figura ilustra dos ajustes con regresión Knn a un conjunto de datos ficticio con dos predictores X_1 y X_2. Con $K = 7$ se obtiene un ajuste más suave. En general el valor óptimo de K depende del bias-variance tradeoff.

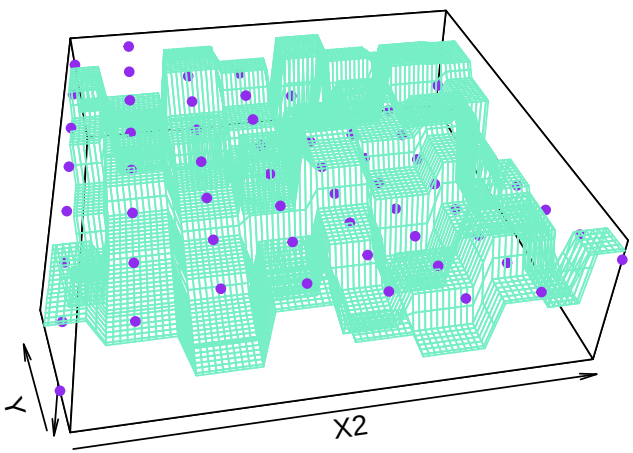
REGRESIÓN LINEAL MÚLTIPLE VS REGRESIÓN KNN

```
## Warning: package 'plot3D' was built under R version 3.6.3
```

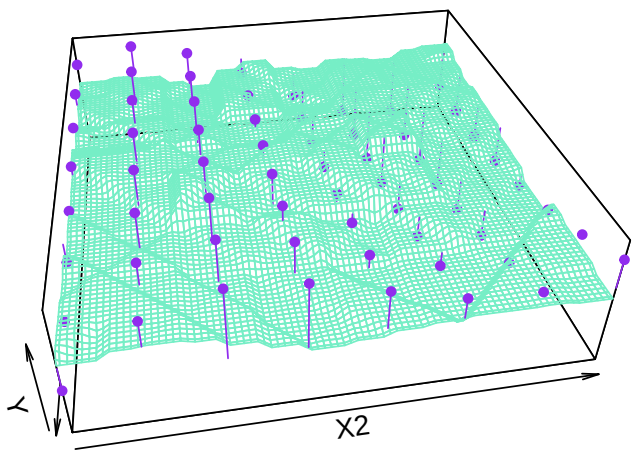
```
## Warning: package 'FNN' was built under R version 3.6.3
```

```
## Warning: package 'wesanderson' was built under R version 3.6.3
```

Knn Regression K=1.



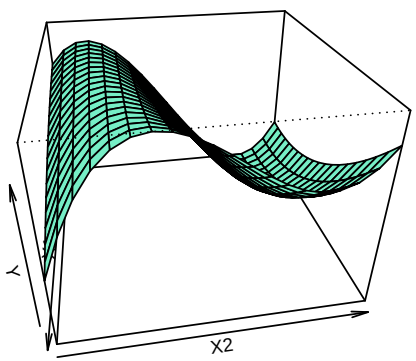
Knn Regression K=7.



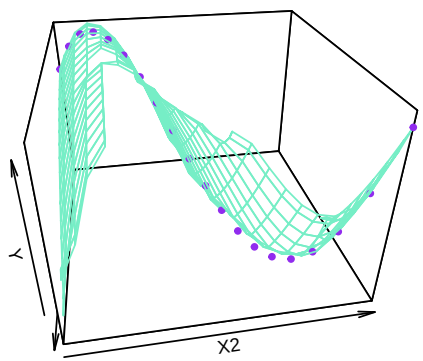
Un valor pequeño de K proporciona un ajuste muy flexible el cual tendrá poco sesgo pero mucha varianza. Esta varianza es debida al hecho de que la predicción en una región dada depende totalmente de una sola observación. En contraste, valores grandes de K proporcionan un ajuste menos variable y más suave, pues la predición en una región es un promedio de algunos puntos y por lo tanto, cambiar una observación tiene un efecto pequeño. Más adelante se verán métodos para seleccionar un valor óptimo para K en regresión Knn.

Sin embargo, el suavizamiento puede incrementar el sesgo al enmascarar algo de la estructura de $f(X)$. Compare el ajuste con K=12 y la $f(X)$ real de los datos.

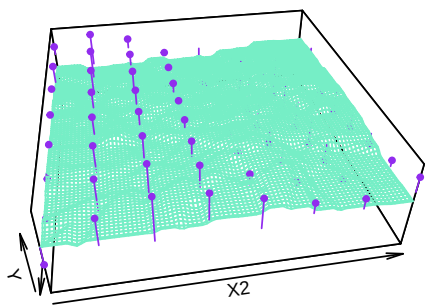
True data



Scatter 3D

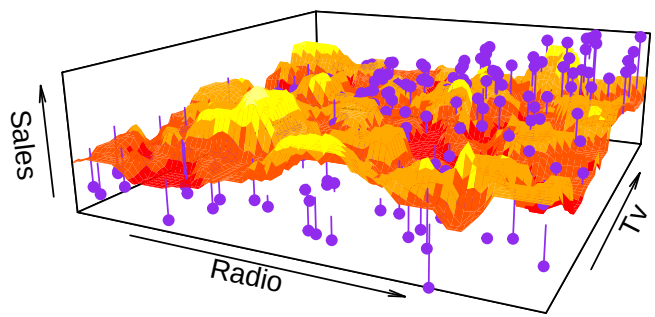


Knn Regression K=12.

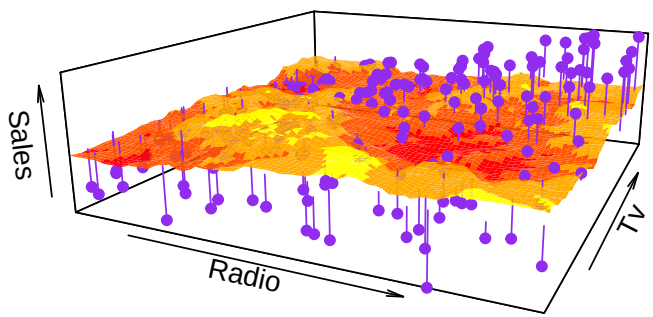


Aplicación Advertising data con TV y Radio:

Knn Regression. k=3

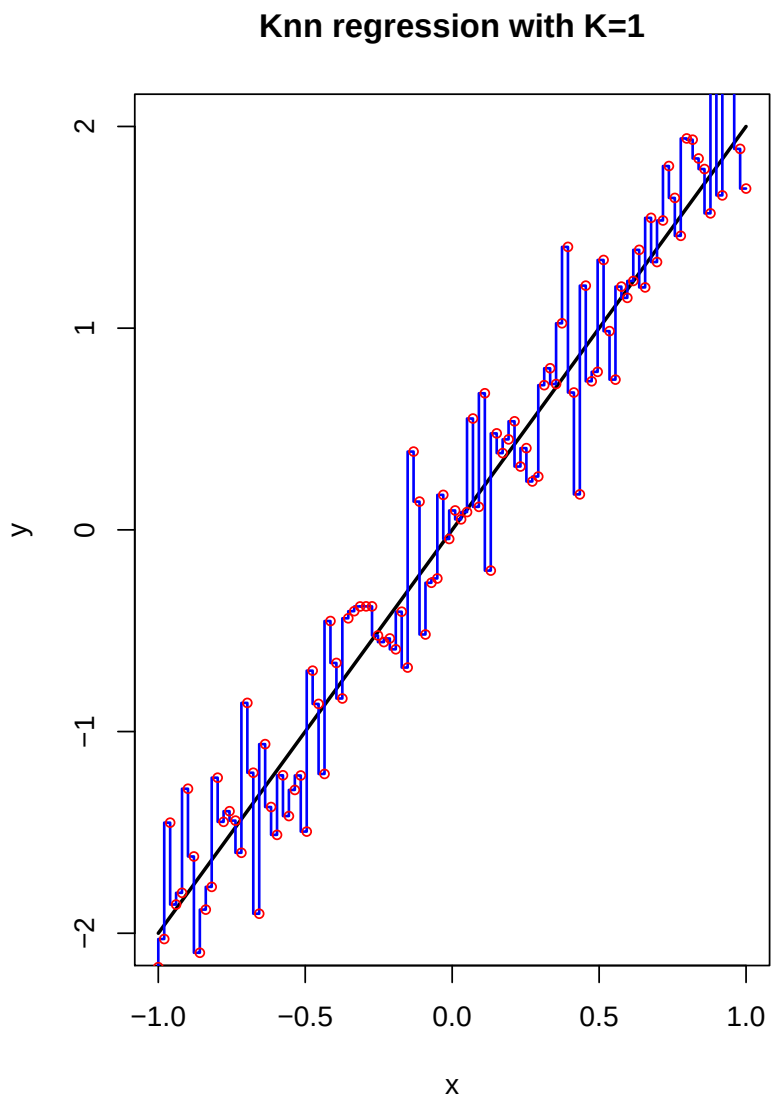


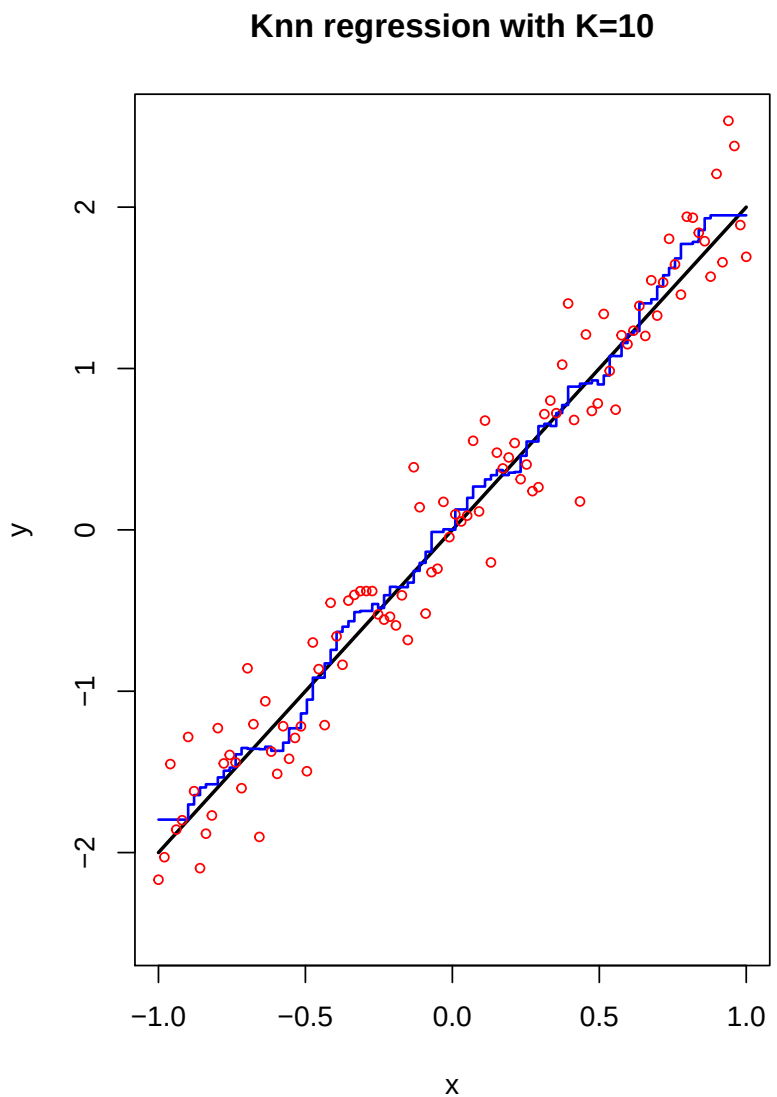
Knn Regression. k=12



¿En que situaciones un método paramétrico (por ejemplo OLS) superará a uno no paramétrico (por ejemplo regresión Knn)? OLS superará a Knn si la forma paramétrica que se ha seleccionado está cercana a la forma real de $f(X)$. Considere los siguiente datos simulados

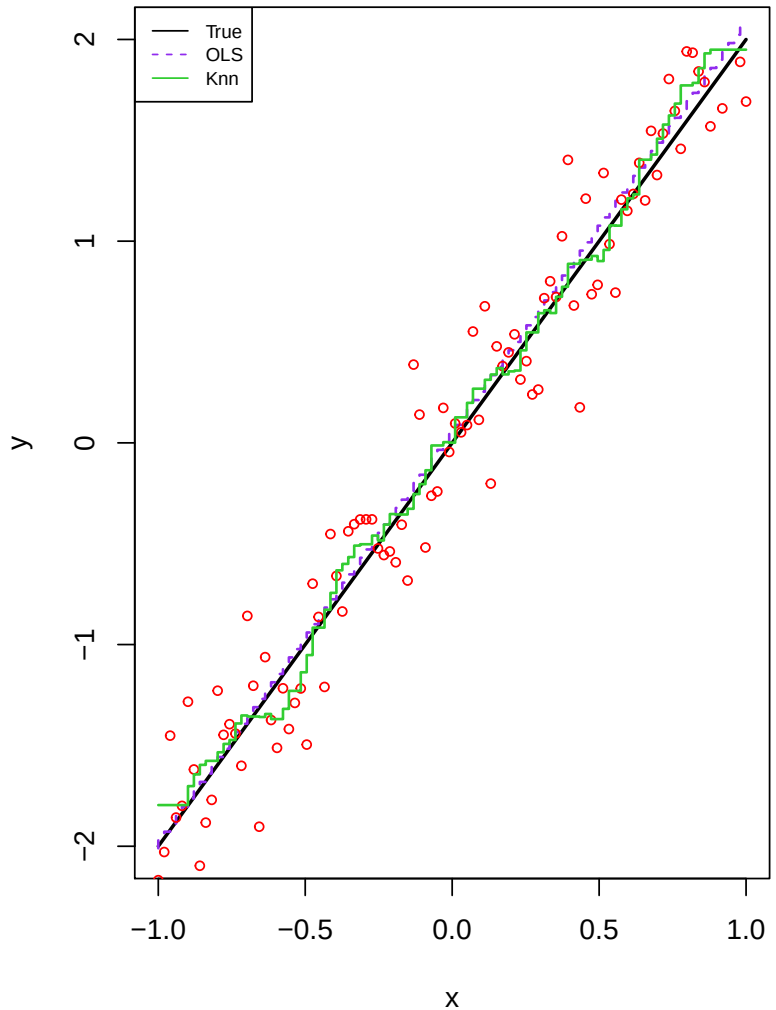
REGRESIÓN LINEAL MÚLTIPLE VS REGRESIÓN KNN

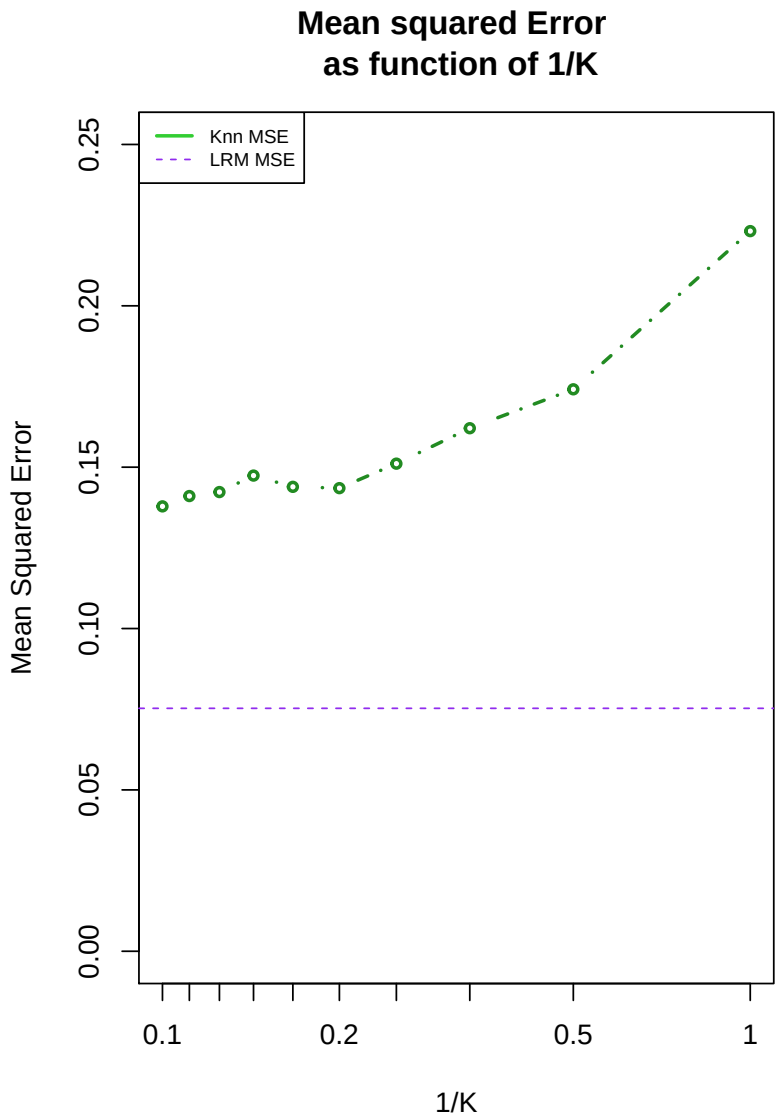




Para $K = 1$ las predicciones son demasiado variables mientras que para $K = 10$ el ajuste es mucho más cercano a la verdadera $f(X)$. Sin embargo, puesto que $f(X)$ es lineal es difícil para Knn (no paramétrico) competir con OLS, pues un método no paramétrico incurrirá en un costo en varianza que no se compensa (offset) con una reducción en el sesgo. Observe el siguiente gráfico. El ajuste OLS es superior al Knn con K=1 pero casi el mismo con K=10:

REGRESIÓN LINEAL MÚLTIPLE VS REGRESIÓN KNN

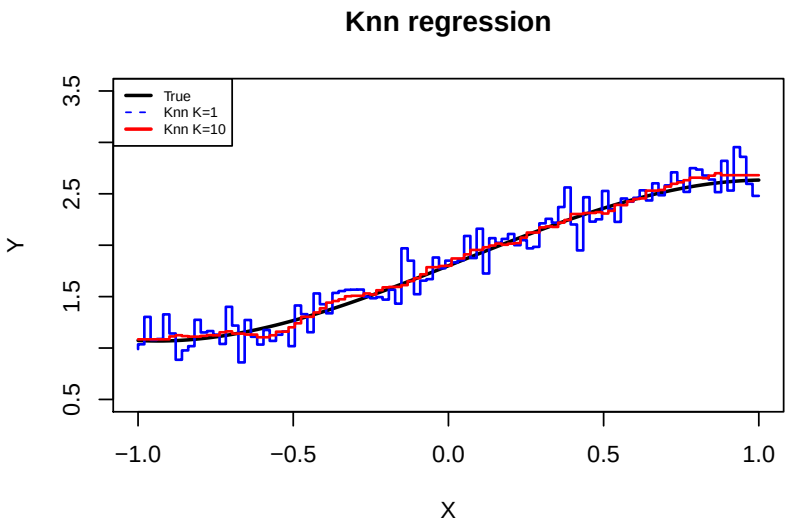


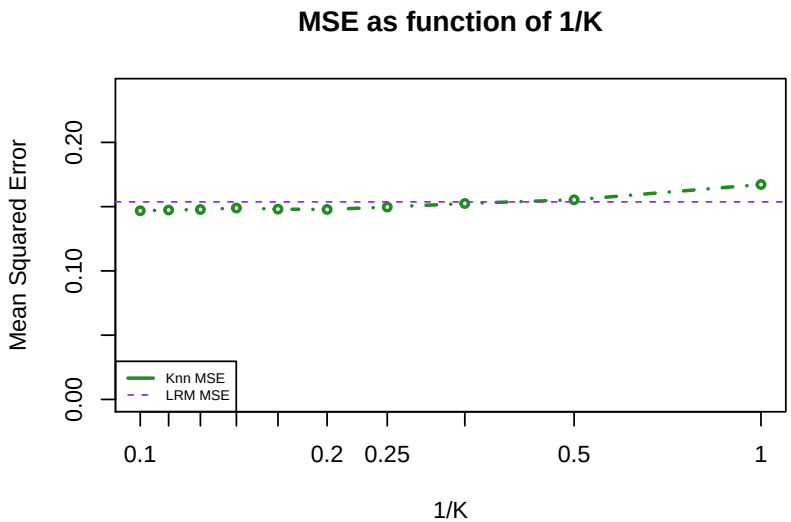


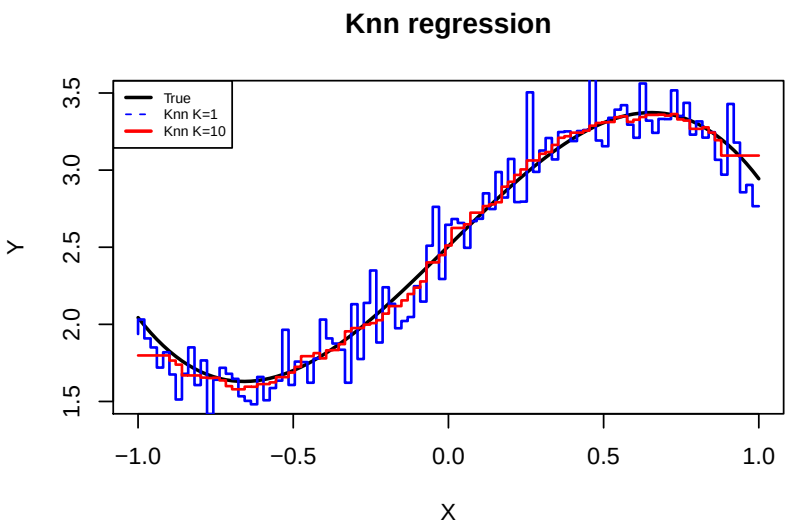
En el panel derecho se observa que la línea verde (MSE para varios valores de K en Knn) está por encima de la línea púrpura (MSE para RLS) con lo cual se demuestra que RLS domina a la regresión Knn para valores pequeños de K, pero a medida que K aumenta su desempeño es similar.

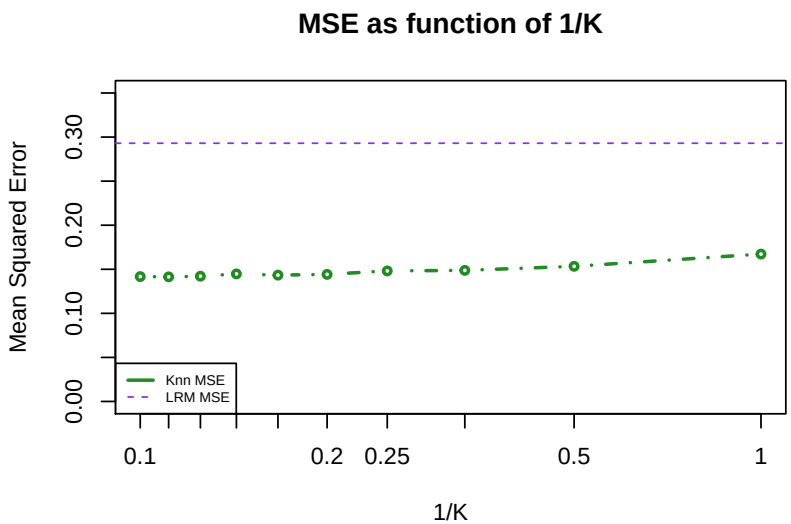
En la práctica, la relación entre X y Y es raramente conocida. En los siguientes gráficos se ilustran los desempeños de OLS y Knn para distintos niveles de no linearidad en la relación entre X y Y.

REGRESIÓN LINEAL MÚLTIPLE VS REGRESIÓN KNN







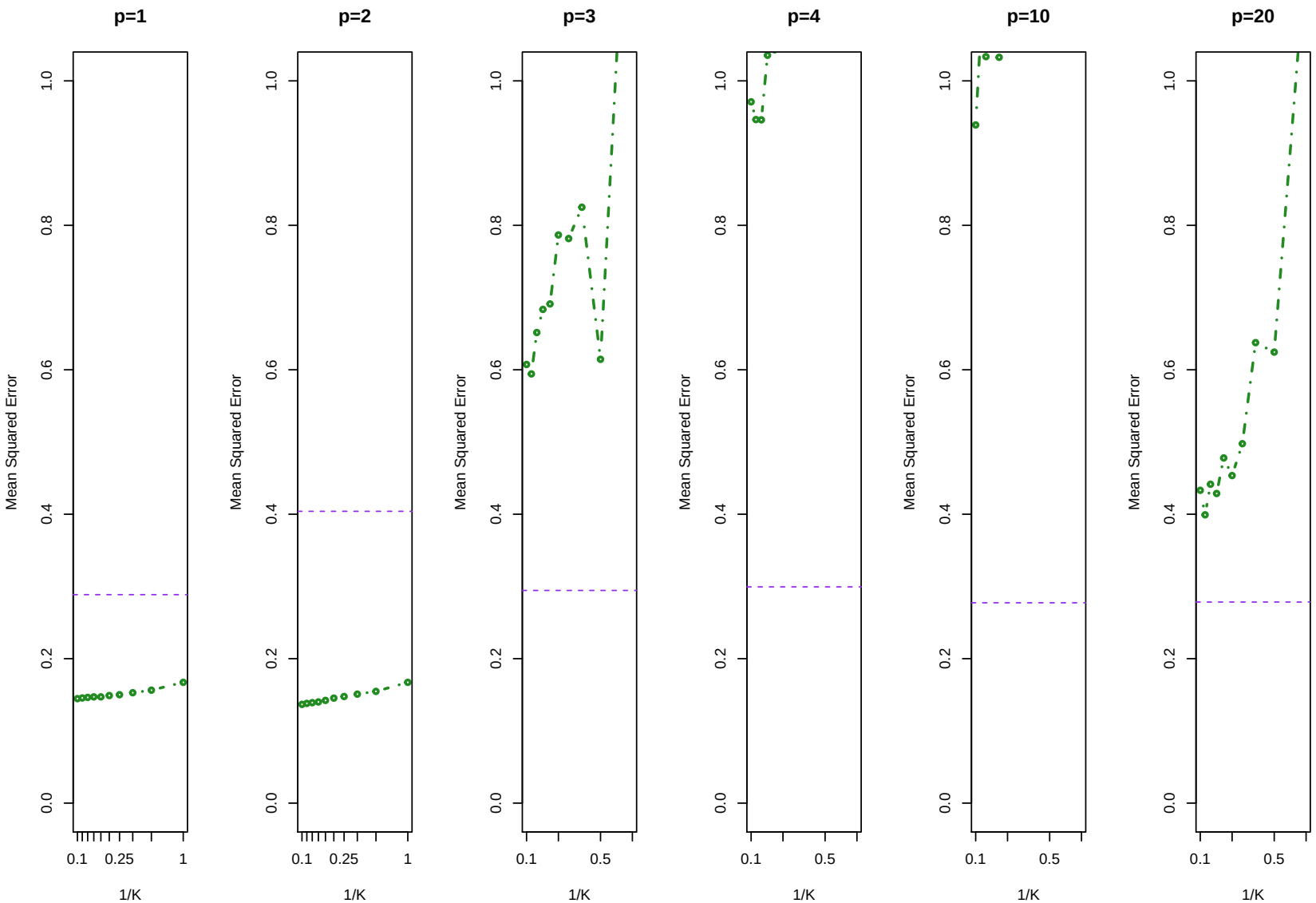


En la fila de arriba la relación es cercana a lineal y OLS es superior para $K < 3$ (MSE, línea púrpura debajo de línea verde), pero para $K \geq 3$ Knn es ligeramente mejor (MSE, línea púrpura arriba de línea verde). La segunda fila, ilustra un caso de mayor nolinealidad y Knn le gana a OLS para todos los valores de K (MSE, línea púrpura arriba de línea verde). También, el MSE para Knn con alta linealidad prácticamente no cambia para valores K de 1 a 10.

Los tres grupos de gráficos anteriores, ilustran situaciones en las cuales Knn tiene un desempeño ligeramente inferior a OLS cuando la relación es lineal, pero notable y consistente mejor cuando la relación es altamente no lineal. Se podría entonces concluir, que en la práctica (donde se desconoce a $f(X)$) Knn es mejor que OLS ya que si $f(X)$ es lineal Knn será casi igual a OLS para valores de K grandes y si $f(X)$ es nolineal, Knn tiene un mejor desempeño. Pero en la realidad, aún cuando $Y = f(X)$ es altamente nolineal, en presencia de muchos predictores (p), OLS domina en general, en términos del MSE, a Knn.

A continuación se ilustra esta situación, donde se considera la misma función de la segunda fila del último gráfico, pero a la cual se le van agregando secuencialmente predictores adicionales que no están asociados con Y (noisy predictors). En los casos de $p = 1$ o $p = 2$ Knn domina a OLS, pero para $p \geq 3$ OLS tiende a dominar a Knn.

REGRESIÓN LINEAL MÚLTIPLE VS REGRESIÓN KNN



Este deterioro en el desempeño de Knn a medida que se aumenta la dimensión p, esta asociado a una reducción en el tamaño muestral, ya que a medida que p se incrementa, el problema es el de distribuir n observaciones en p dimensiones, lo cual ocasiona que una observación tenga pocos o ningún vecino cercano (es como distribuir poca personas en un área de muchas hectáreas). Este fenómeno se conoce como La Maldición de la Dimensionalidad (The Curse of Dimensionality).

Como regla general, los métodos paramétricos tienden a tener un mejor desempeño que los no paramétricos cuando hay un número pequeño de observaciones por predictor. Aún en problemas donde la dimensionalidad es baja, se recomienda usar RLM en vez de Knn ya que se gana en interpretabilidad.

CLASIFICACIÓN

En muchas situaciones de la vida real, la variable de respuesta es cualitativa o categórica y no cuantitativa. Por ejemplo, el color de los ojos es cualitativa (verde, café, gis, azul). El proceso de predecir respuesta categóricas se conoce como CLASIFICACIÓN. El proceso de clasificación consiste en asignar una observación a una categoría o clase. Puesto que algunos de estos métodos antes de clasificar predicen la probabilidad de cada una de las categorías o niveles de la variable cualitativa, ellos se comportan de alguna manera como modelos de regresión.

Por el momento se discuten solamente tres clasificadores ampliamente usados: Regresión logística (LR), Discriminante lineal (LDA) y K vecinos más cercanos (Knn). Más adelante en el curso, se presentan otros métodos alternativos y modernos de clasificación tales como modelos aditivos generalizados (GAM), árboles (Trees), Bosques aleatorios (Random forest) y máquinas de soporte vectorial o de vectores (SVM)

CLASIFICACIÓN

Los problemas de clasificación surgen con frecuencia (tal vez más que los problemas de regresión con respuesta cuantitativa). Algunos ejemplos incluyen:

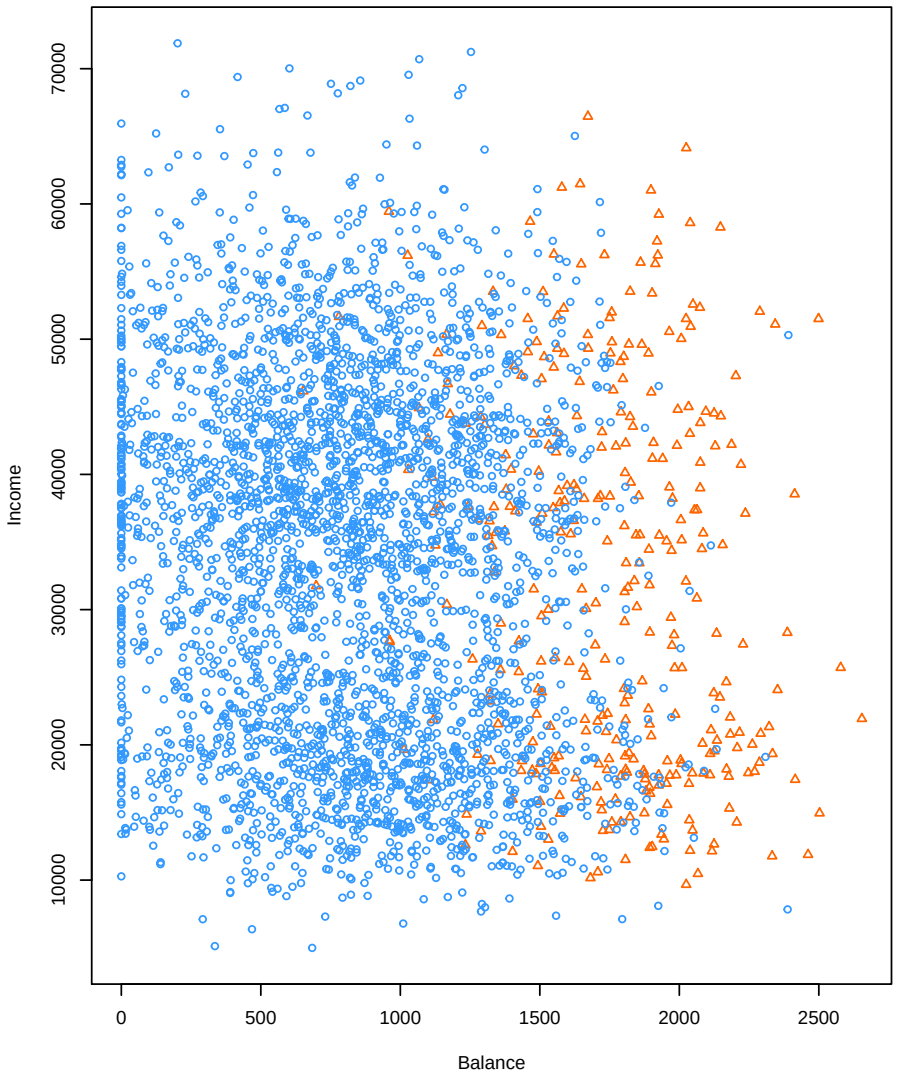
- Una persona llega a emergencias con unos síntomas que podrían atribuirse a una de tres condiciones médicas ¿En cuál de las tres podría clasificarse?
- Un servicio de banco online debe ser capaz de determinar si una transacción en un sitio es o no fraudulenta (Si=1, No=0) y solo se cuenta con la dirección IP del usuario, historia previa de transacciones y así sucesivamente (un conjunto limitado de predictores)
- Con base en datos de una secuencia de ADN para un cierto número de pacientes con y sin una enfermedad específica, un biólogo quiere averiguar cuál mutación se asocia a la enfermedad y cuál no (respuesta y predictor dicótomos)

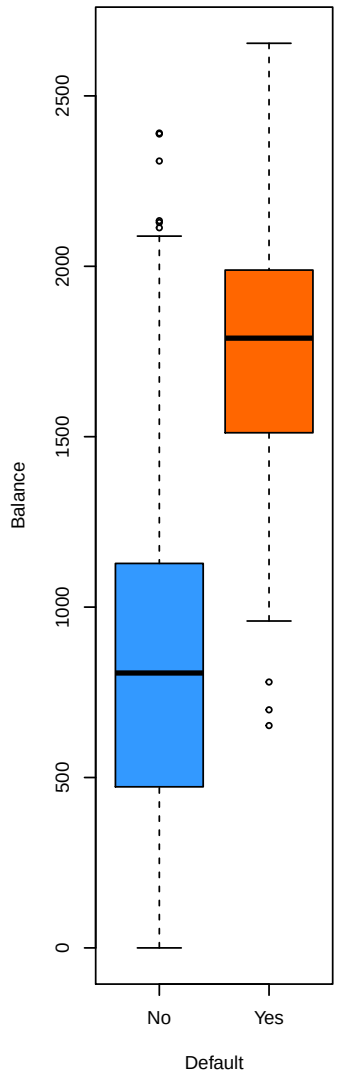
En el escenario de clasificación se tiene un conjunto de entrenamiento (Training Set) $(x_1, y_1), \ldots, (x_n, y_n)$ que se puede usar para construir un clasificador. Se desea que dicho clasificador tenga un buen desempeño no solo en el conjunto de entrenamiento sino también en un conjunto de validación o de prueba (Test Set). Para ilustrar, se usará el Default dataset de ISLR [1].

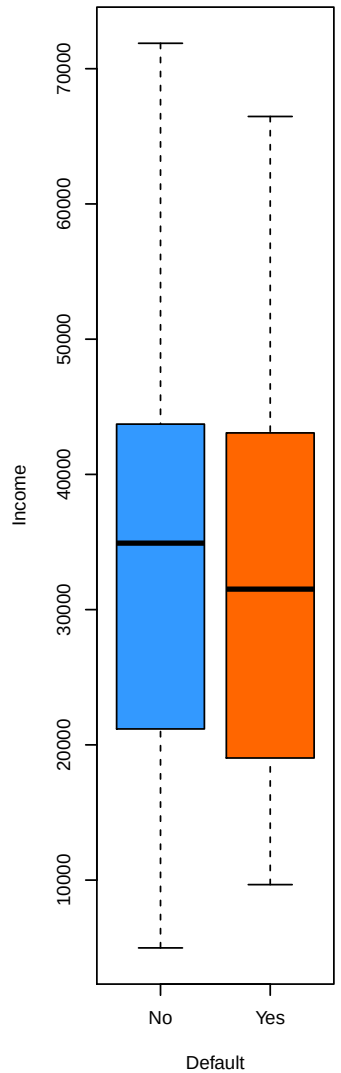
Con estos datos, interesa predecir si un individuo incumplirá con su obligación de pagar su tarjeta de crédito, con base en su ingreso anual y el balance mensual de su tarjeta de crédito. El conjunto de datos se muestra a continuación:

[1] http://faculty.marshall.usc.edu/gareth-james/ISL/data.html

CLASIFICACIÓN







La tasa global de default (incumplimiento en obligación) es del 3 % por lo que se muestra solo una fracción de las personas que no incumplieron (Default="No"). De los boxplots, parece que los individuos que incumplieron (Default="Yes") tienden a tener mayores balances en sus tarjetas de crédito en comparación a los que no incumplen. El ingreso (income) de los que incumplen parece ser similar al de los que cumplen. Pero ¿Cómo se puede evaluar el efecto conjunto de los predictores X_1=Balance y X_2=Income sobre la respuesta cualitativa Y =Default?

De hecho, el MRL no es apropiado para este tipo de respuestas binarias ya que podría generar valores de probabilidad de pertenecia a una de las categorás que podrían ser o negativas o mayores a uno, lo cual no es admisible para una probabilidad. Esto se ilustrará a continuación. Suponga que se está tratando de predecir la condición médica de un paciente en urgencias con base en sus síntomas.

Hay tres posibles diagnósticos: derrame cerebral (stroke), sobredosis por droga (drug overdose), y ataque epiléptico (epileptic seizure), por lo que

$$Y = \begin{cases} 1 & \text{If stroke;} \\ 2 & \text{If drug overdose;} \\ 3 & \text{If epileptic seizure.} \end{cases}$$

Usando esta codificación, se pueden usar OLS para ajustar un modelo de regresión lineal para predecir a Y con base en un conjunto de predictores $X_1,\ X_2, \ldots, X_p$. Un problema con esta codificación, es que si por ejemplo se considera esta otra (igualmente válida):

$$Y = \begin{cases} 1 & \text{If If epileptic seizure;} \\ 2 & \text{stroke;} \\ 3 & \text{If drug overdose.} \end{cases}$$

Estas dos codificaciones podrían producir modelos lineales distintos que podrían conducir a distintos conjuntos de predicciones de observaciones de prueba. En el caso binario la situación es más favorable. Por ejemplo, si solo se consideran dos condiciones para el paciente: stroke y drug overdose, la respuesta sería

$$Y = \begin{cases} 0 & \text{If stroke;} \\ 1 & \text{If drug overdose.} \end{cases}$$

y entonces ajustar un modelo lineal OLS para esta respuesta binaria y predecir drug overdose si $\hat{Y} > 0.5$ y stroke si $\hat{Y} \leq 0.5$. Aún si se considera la codificación

$$Y = \begin{cases} 1 & \text{If stroke;} \\ 0 & \text{If drug overdose.} \end{cases}$$

la RL producirá las mismas predicciones finales solo que con el β para Balance con signos contrarios.

```
## Default
##    0    1
## 9667  333
```

```
##                   Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -0.0751919588 3.354360e-03 -22.41618 1.262551e-108
## Balance      0.0001298722 3.474933e-06  37.37401 2.774969e-286
```

```
## Default
##    0    1
##  333 9667
```

```
##                   Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)  1.0751919588 3.354360e-03 320.53563  0.000000e+00
## Balance     -0.0001298722 3.474933e-06 -37.37401 2.774969e-286
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

En este caso binario, OLS tiene sentido ya que le predictor lineal estimado por este método, $X\hat{\beta}$ es de hecho una estimación de

$$Pr\,(\text{drug overdose} \mid X)$$

pero algunas estimaciones podrían estar fuera del intervalo $[0,\ 1]$. Sin embargo, las predicciones proporcionan un ordenamiento y pueden interpretarse como estimaciones de las probabilidades crudas. Como dato curioso, sucede que las clasificaciones que se obtienen con RL para predecir una respuesta binaria serán iguales a las obtenidas vía análisis discriminante lineal (LDA[2])

[2] LDA se discute más adelante