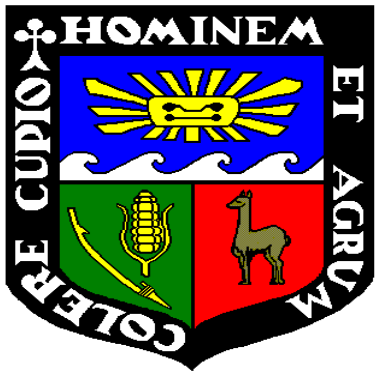
Técnicas Multivariadas Avanzadas

Regresión lineal

Ms Carlos López de Castilla Vásquez

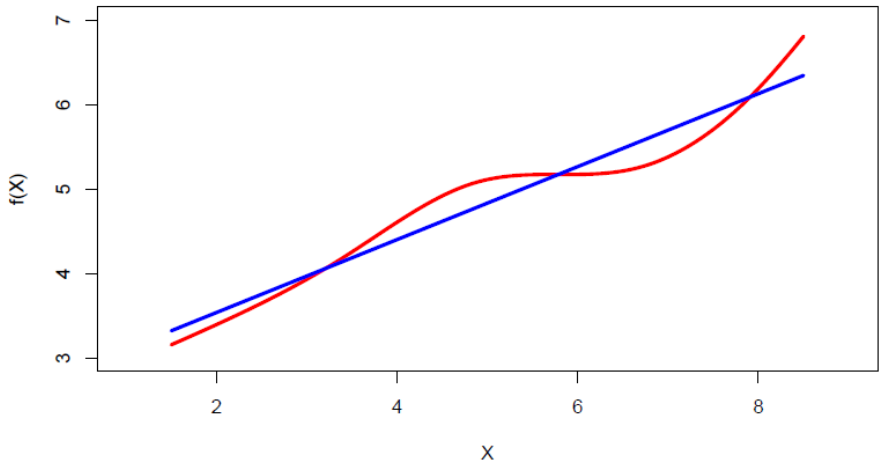
Universidad Nacional Agraria La Molina

2014-2

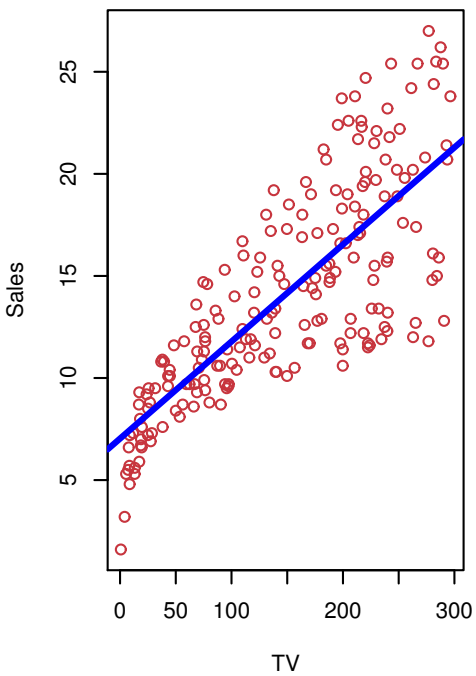


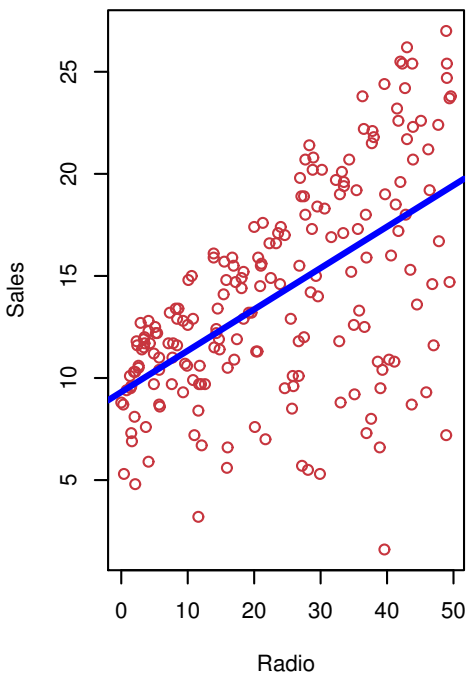
Regresión lineal

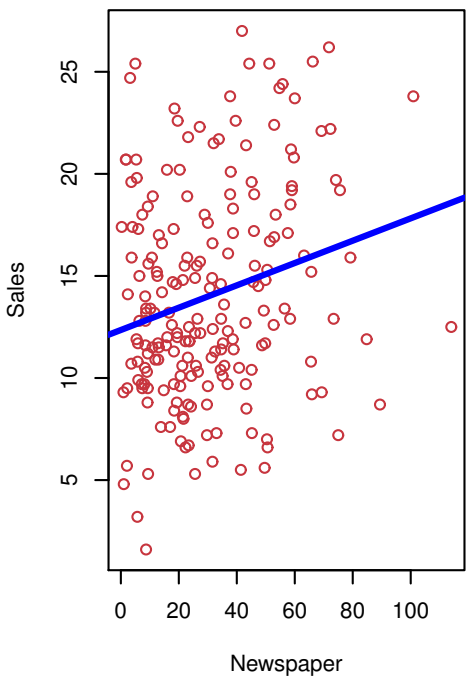
- La regresión lineal es una metodología simple para el aprendizaje supervizado. Asume que la dependencia de Y sobre $X = (X_1, X_2, \cdots, X_p)$ es lineal.
- Sin embargo las funciones de regresión *casi nunca* son lineales.
- Aunque pueda parecer muy simplista la regresión lineal es extremadamente útil en términos prácticos y conceptuales.



Regresión lineal para la data Advertising







Regresión lineal para la data Advertising

- ¿Existe relación entre el presupuesto en publicidad y las ventas?
- ¿Qué tan fuerte es la relación anterior?
- Qué medio de publicidad contribuye más en las ventas?
- ¿Con que precisión se pueden predecir las ventas?
- ¿La relación es lineal?
- ¿Existe sinergia entre los medios de publicidad?

Regresión lineal simple

- Se asume el modelo:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 y β_1 son constantes desconocidas llamadas *coeficientes* o *parámetros*.
- ϵ es un término de *error*.
- Si se tienen las *estimaciones* $\hat{\beta}_0$ y $\hat{\beta}_1$ se pueden predecir el valor de la variable respuesta:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

donde $\hat{y}$ indica la predicción de Y sobre x.

Método de mínimos cuadrados

- Sea $e_i = y_i - \hat{y}_i$ que representa el i-ésimo *residual*.
- Se define la *suma de cuadrados de los residuales* (RSS) por:

$$\text{RSS} = e_1^2 + \cdots + e_n^2$$

o equivalentemente:

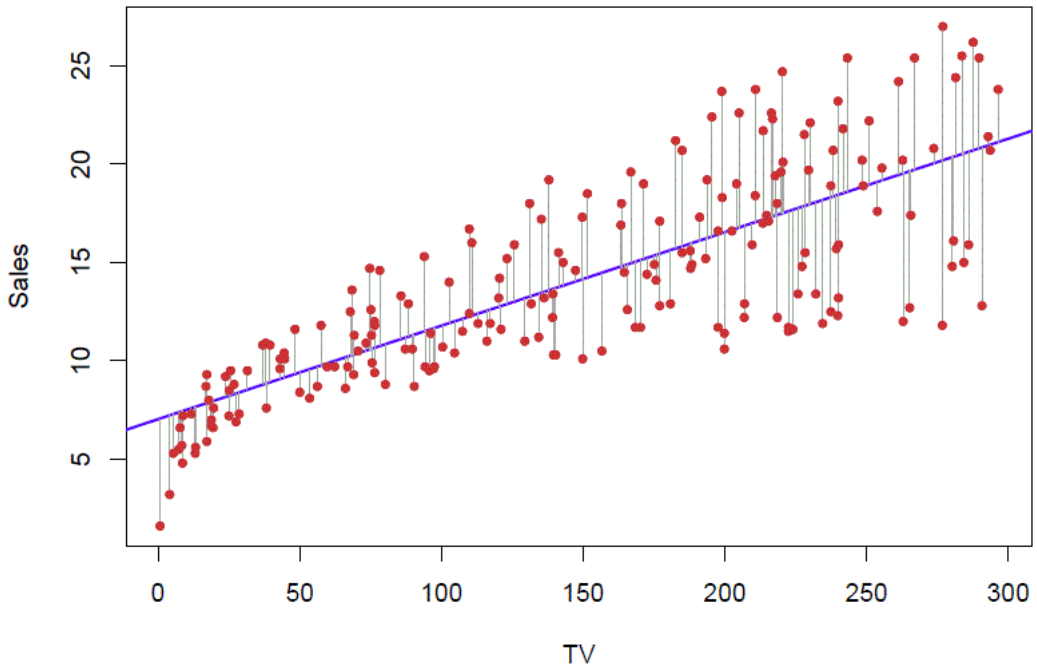
$$\text{RSS} = (y_1 - \hat{\beta}_0 + \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 + \hat{\beta}_1 x_n)^2$$

- El método de *mínimos cuadrados* escoge $\hat{\beta}_0$ y $\hat{\beta}_1$ de tal forma que se minimize RSS. Luego:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Ejemplo: Data Advertising

- La regresión de sales sobre TV captura la esencia de la relación a pesar de que resulta algo deficiente en la parte izquierda del gráfico.



Error estándar

- El *error estándar* de un estimador refleja su comportamiento bajo un muestreo repetido.
- Se tiene que:

$$\text{SE}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \qquad \text{SE}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

donde $\sigma^2 = \text{Var}(\epsilon)$.

- Un intervalo de confianza del $(1 - \alpha) \times 100\,\%$ para la pendiente tiene la forma:

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}\widehat{\text{SE}}(\hat{\beta}_1)$$

Pruebas de hipótesis

- Los errores estándar son usados para establecer *pruebas de hipótesis* sobre los coeficientes, por ejemplo:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

- Para evaluar la hipótesis nula se calcula:

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} \sim t_{n-2}$$

- Usando R se puede obtener la probabilidad de observar un valor mayor o igual a $|t|$ llamado *p-value*.

Error estándar residual y R^2

- Se usa el *error estándar residual*:

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}$$

- También se puede calcular la fracción de la varianza explicada:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

donde $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y}^2)$ es la *suma de cuadrados del total*.

- Se puede probar que $R^2 = r^2$ donde r es el *coeficiente de correlación* entre X y Y.

Regresión lineal múltiple

- El modelo de regresión lineal múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Se interpreta β_j como el *efecto promedio* que tiene cada incremento unitario en X_j sobre Y *manteniendo fijos* el resto de predictores.
- La interpretación anterior es posible solo cuando los predictores no se encuentran correlacionados.
- En la data Adversiting el modelo seria:

$$\text{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} + \beta_3 \text{newspaper} + \epsilon$$

Regresión lineal múltiple

- "Data Analysis and Regression" Mosteller and Tukey 1977

  Un coeficiente de regresión estima el cambio esperado en Y por unidad de cambio en X_j cuando todos los otros predictores permanecen fijos. Sin embargo los predictores usualmente cambian juntos.

- George Box

  Escencialmente todos los modelos son incorrectos, pero algunos son útiles.

Estimación y predicción

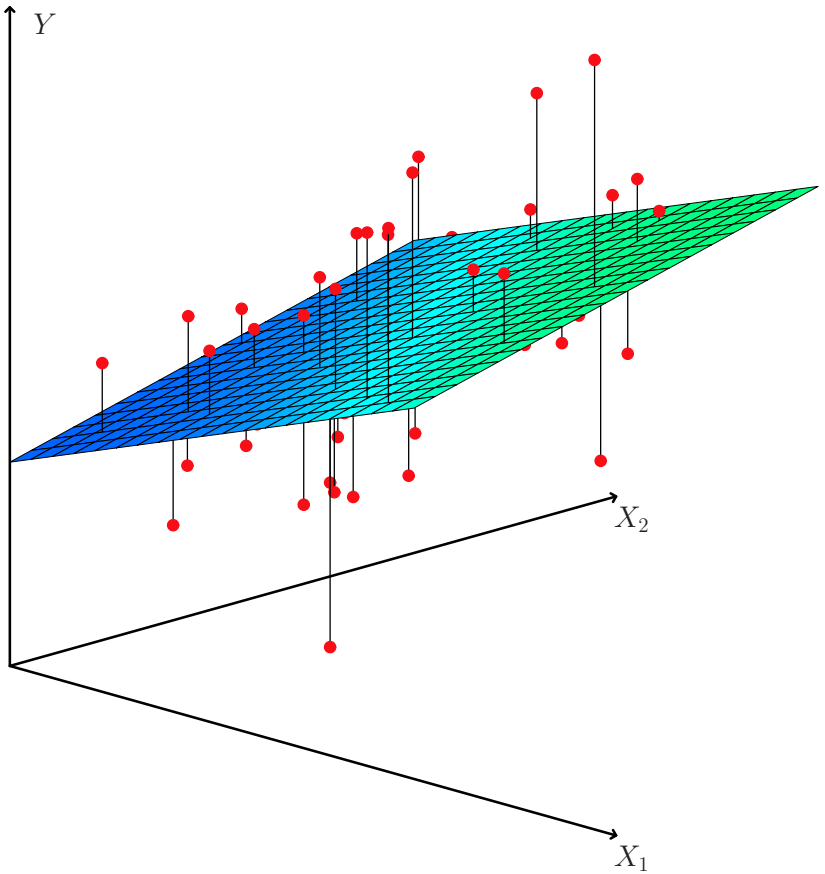
- Dadas las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ se pueden realizar predicciones usando:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- Las estimaciones de los parámetros se obtienen minimizando nuevamente la suma de cuadrados del residual RSS:

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Estimación y predicción



Algunas preguntas

- ¿Es al menos uno de los predictores útil para predecir la variable respuesta?
- ¿Ayudan todos los predictores a explicar Y o solo un subconjunto de ellos?
- ¿Que tan bien el modelo estima la data?
- ¿Con que precisión es posible predecir un valor para y dado un conjunto de valores para los predictores?
- Para la primera pregunta se puede usar:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

Identificando las variables importantes

- La metodología directa es llamada *all subsets* o *best subsets regression*.
- Se obtienen todos los posibles modelos de regresión y luego se eligen algunos usando algún criterio que permita un balance entre el *error de estimación* y el *tamaño del modelo*.
- Sin embargo no es posible examinar todos los posibles modelos ya que existen 2^p (si $p = 40$ existen más de un billón de modelos)
- Se requiere una método que busque a través de un subconjunto de estos modelos. Se presentan a continuación dos de estas metodologías.

Forward selection

- Se empieza con un *modelo nulo*, es decir aquel que contiene un intercepto pero ninguna variable predictora.
- Estimar las p regresiones lineales simples y agregar al modelo aquella variable con la que se obtenga el menor RSS.
- Agregar al modelo la variable que resulta en el menor RSS entre todos los modelo con dos variables que incluya la variable incuida en el paso anterior.
- Continuar hasta que se satisfaga algún criterio de parada, por ejemplo cuando todas las variables restantes tengan un p-value por encima de cierto umbral.

Backward selection

- Se empiezan con todas las variables en el modelo.
- Se elimina la variable con el mayor p-value.
- Se estima el modelo con $p-1$ variables y nuevamente se elimina aquella variable con el mayor p-value.
- Se continua hasta que se alcanza algún criterio de parada.
- Por ejemplo el proceso se detiene cuando todas las variables en el modelo tienen un p-value por debajo de un nivel de significación definido como *umbral*.

Criterios para seleccionar modelos

- Más adelante se discutirán algunos *criterios* para escoger un modelo óptimo dentro de aquellos obtenidos por la selección stepwise o backward.
- Estos criterios incluyen: C_p de Mallows, *Criterio de Información de Akaike* (AIC), *Criterio de Información Bayesiano* (BIC), R^2 *ajustado* y *Validación Cruzada* (CV).

Predictores cualitativos

- Muchas veces se tienen predictores cualitativos que toman valores dentro de un conjunto discreto.
- Estos predictores son también llamados variables categóricos o factores.
- En la data Credit se tienen cuatro variables cualitativas: gender, student (student status), status (marital status) y ethnicity (Caucasian, African American (AA) o Asian).
- Suponga que se desea estudiar las diferencias en el balance de la tarjeta de crédito entre los hombres y mujeres ignorando las otras variables.

Predictores cualitativos con dos niveles

- Se crea la siguiente variable:

$$x_i = \begin{cases} 1 & \text{si la } i-\text{esima persona es mujer} \\ 0 & \text{si la } i-\text{esima persona es hombre} \end{cases}$$

- El modelo resultante es:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \\ \beta_0 + \epsilon_i \end{cases}$$

Predictores cualitativos con más de dos niveles

- Para ethnicity se crean dos variables dummy. La primera podria ser:

$$x_{i1} = \begin{cases} 1 & \text{si la } i-\text{esima persona es Asiatica} \\ 0 & \text{si la } i-\text{esima persona no es Asiatica} \end{cases}$$

y la segunda:

$$x_{i2} = \begin{cases} 1 & \text{si la } i-\text{esima persona es Caucasica} \\ 0 & \text{si la } i-\text{esima persona no es Caucasica} \end{cases}$$

Predictores cualitativos con más de dos niveles

- Las variables anteriores pueden ser usadas en la ecuación de regresión para obtener el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \\ \beta_0 + \beta_2 + \epsilon_i \\ \beta_0 + \epsilon_i \end{cases}$$

- Se requiere una cantidad de variables dummy igual al número de niveles menos uno.
- El nivel que no tiene asociado ninguna variable dummy (AA) es conocida como *baseline*.

Extenciones del modelo lineal

- Se puede considerar en el modelo interacciones y efectos no lineles.
- En el análisis para la data Adversiting se asume que el efecto de los medios de publicidad sobre sales es independiente de la cantidad gastada en los otros.
- Por ejemplo, el modelo lineal:

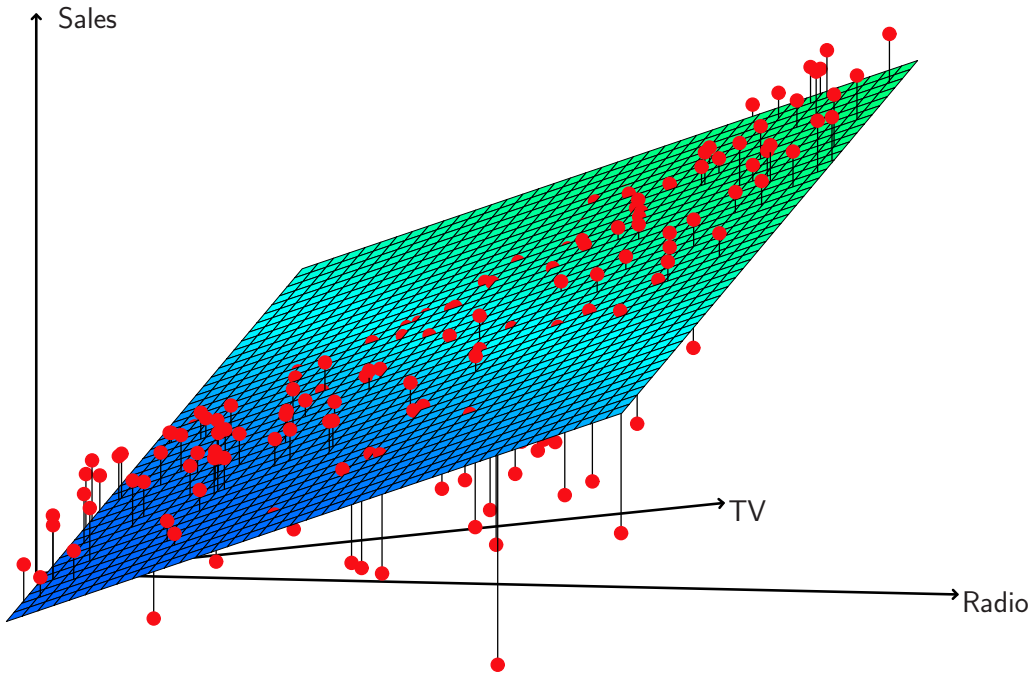
$$\text{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} + \beta_3 \text{newspaper}$$

considera que el efecto de incrementar TV en una unidad sobre sales siempre es β_1 independientemente de la cantidad gastada en radio.

Interacciones

- Suponga que gastar dinero en publicidad por radio incrementa la efectividad de la publicidad en TV, de tal forma que la pendiente para esta variable debería incrementarse conforme radio se incrementa.
- En esta situación, con un presupuesto fijo de 100000 dolares, gastando la mitad en radio y la otra mitad en TV podriamos incrementar sales más que si gastaramos todo el presupuesto solo en TV o solo en radio.
- En marketing, esto es conocido como efecto de *sinergia* y llamado efecto de *interacción* en estadística.

Interacción en la data adversiting



Interacción en la data adversiting

- Se observa que para valores bajos de TV o radio los valores de sales son menores que los predecidos por el modelo lineal.
- Sin embargo cuando se divide la publicidad entre dos medios se tiende a subestimar sales.
- El modelo tiene la forma:

$$\text{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} + \beta_3 \text{TV} \times \text{radio}$$

- Los resultados en R sugieren que la interacción es importante.
- El R^2 para el modelo anterior es 96,8 % comparado con un 89,7 % para el modelo que no considera interacción.

Interacción en la data adversiting

- Esto significa que $(96{,}8 - 89{,}7)/(100 - 89{,}7) = 69\,\%$ de la variabilidad en sales que queda luego de estimar el modelo aditivo es explicado por el término de interacción.
- Los coeficientes estimados sugieren que un incremento en 1000 dólares en la publicidad para TV esta asociado a un incremento en sales de:

$$(\hat{\beta}_1 + \hat{\beta}_3 \text{radio}) \times 1000 = 19 + 1{,}1 \text{radio unidades}$$

- Un incremento en 1000 dólares en la publicidad en radio esta asociado con un incremento en sales de:

$$(\hat{\beta}_2 + \hat{\beta}_3 \text{TV}) \times 1000 = 29 + 1{,}1 \text{TV unidades}$$

Jerarquía

- Muchas veces el efecto de interacción tiene un p-value pequeño pero los efectos principales asociados (en este caso TV y radio) no son importantes.
- El principio de jerarquía:

  Si se incluye un efecto de interacción en un modelo, también se debe incluir los efectos principlaes aún cuando su p-values asociados no sean significativos.

- La razón para este principio esta en que la interacción es difícil de interpretar en un modelo sin los efectos principales ya que el significado de los coeficientes estimados cambiaría.

Interacciones entre variables cualitativas y cuantitativas

- Considere la data Credit y suponga que se desea predecir balance usando income (cuantitativa) y student (cualitativa).
- Sin considerar un témino de interacción el modelo tiene la forma:

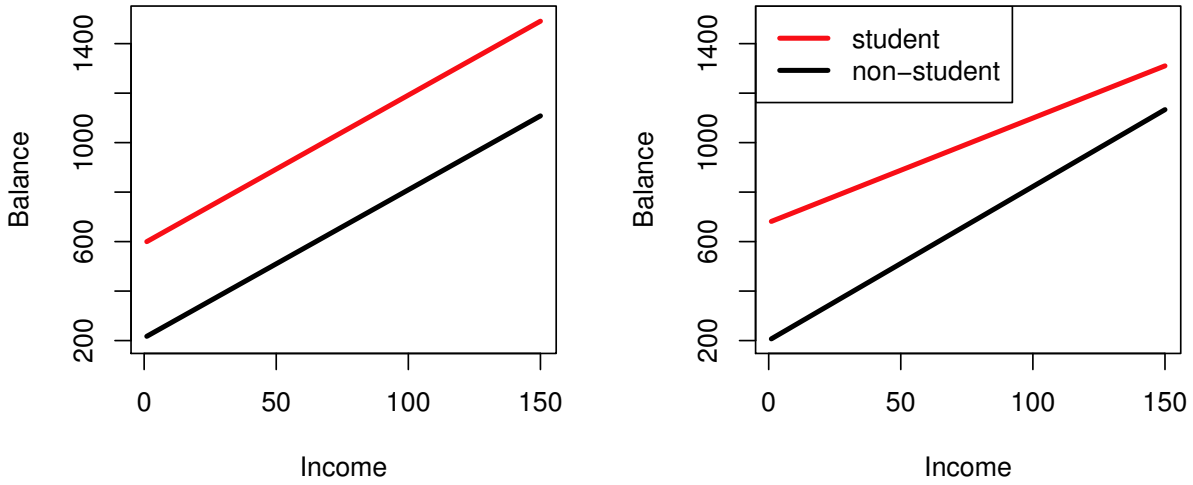
$$\text{balance}_i = \beta_0 + \beta_1 \text{income} + \beta_2 x_i = \begin{cases} \beta_0 + \beta_2 + \beta_1 \text{income} \\ \beta_0 + \beta_1 \text{income} \end{cases}$$

- Considerando interacción:

$$\text{balance}_i = \beta_0 + \beta_1 \text{income} + \beta_2 x_i + \beta_3 \text{income} x_i$$

Gráfico data Credit

- El gráfico de la izquierda considera el modelo sin interacción entre income y student.
- El gráfico de la derecha considera el modelo con interacción entre income y student.

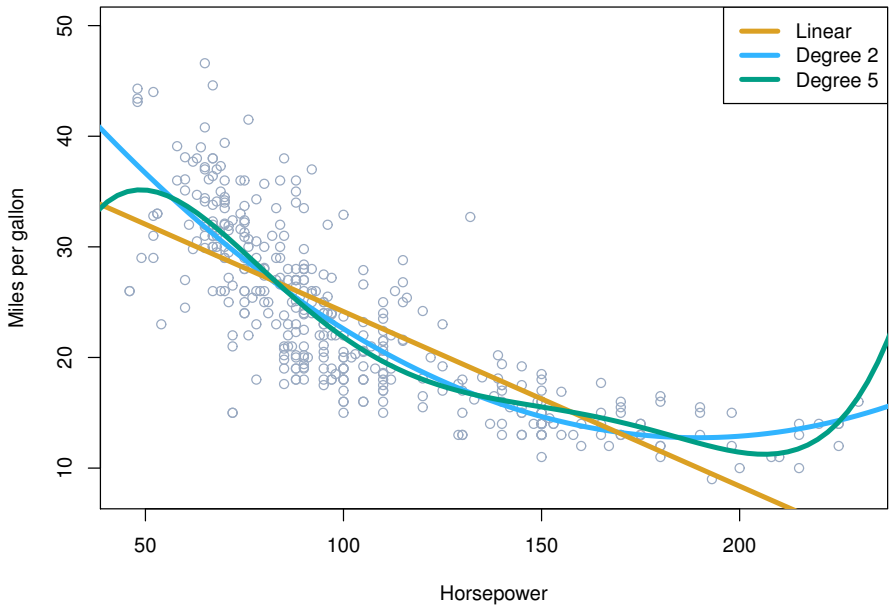


Efectos no lineales

- El gráfico sugiere que:

$$\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{horsepower}^2 + \epsilon$$

proporciona un mejor ajuste.



Generalizaciones

- En el resto del curso se discuten métodos que expanden el alncance de los modelos lineales.
- *Problemas de clasificación*: regresión logística, support vector machines.
- *No linealidad*: Suavización por kernel, splines, modelos aditivos generalizados, vecinos más cercanos.
- *Interacciones*: Arboles, bagging, random forest y boosting.
- *Estimación regularizada*: regresión ridge y lasso.