Introducción a la analítica

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

Si los supuestos del modelo lineal son correctos:

$$E\left(\frac{RSS}{(n-p-1)}\right) = \sigma^2$$

y si H_0 es cierta,

$$E\left(\frac{(TSS-RSS)}{p}\right) = \sigma^2$$

Entonces, cuando no hay relación entre Y y los predictores, se espera que el estadístico F tome un valor cercano a 1. Por otra parte, si H_1 es cierta, entonces

$$E\left(\frac{(TSS-RSS)}{p}\right) > \sigma^2$$

por lo que se espera que F sea mayor a 1.

Refiérase al ejemplo de Advertising:

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: Sales ~ 1
## Model 2: Sales ~ TV + Radio + Newspaper
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    199 5417.1
## 2    196  556.8  3    4860.3 570.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De la salida anterior, se observa el estadístico F de la regresión $Sales \sim TV + Radio + Newspaper$ el cual tiene un valor de 570.3. Puesto que este valor es mucho más grande que 1, él proporciona evidencia contundente contra la hipótesis nula H_0, es decir el estadístico F sugiere que al menos uno de los medios publicitarios debe estar relacionado con las ventas. Sin embargo, ¿Qué hubiera pasado si el estadístico F estuviera cerca de 1? ¿Qué tan grande debe ser este estadístico antes de que se pueda rechazar a H_0 y concluir que hay una relación importante entre predictores y respuesta? La respuesta depende de los valores de n y p.

Cuando n es grande, un estadístico F que es solo ligeramente más grande que 1 podría aún proporcionar evidencia contra H_0. En contraste, se necesita un valor de F grande para rechazar a H_0 si n es pequeño. Cuando H_0 es cierta y los errores ε_i se distribuyen de acuerdo a una normal , $F \sim F(p, n-p-1)$ y esto permite obtener un valor-p con el cual se rechaza o no a H_0. De la salida de R anterior, se observa un valor-p muchísimo menor a 0.05 por lo que se puede argumentar que hay mucha evidencia, evidencia fuerte, de que al menos uno de los medios publicitarios está asociado con las ventas (Sales).

Algunas veces es de interés probar si un conjunto particular de q coeficientes son cero:

$$H_0 : \ \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p$$

En este caso, se ajusta un segundo modelo que usa todas las variables excepto esas últimas q. Si la suma de cuadrados residual para ese modelo es RSS_0 entonces el estadístico F apropiado es:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} \sim F(q, \ n-p-1) \quad \text{, Bajo } H_0$$

Note que en la tabla de estimaciones anterior se reporta un estadístico t y un valor-p. Estos porporcionan información de si cada predictor individual se relaciona con la respuesta, después de ajustar por los otros predictores. Sucede que cada uno de estos son exactamente equivalentes al test F que omite cada una de las variables de ese modelo, dejando todas las otras. Es decir, reporta el efecto parcial de adicionar esa variable en el modelo.

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: Sales ~ TV + Radio
## Model 2: Sales ~ TV + Radio + Newspaper
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    197 556.91
## 2    196 556.83  1  0.088717 0.0312 0.8599
```

Se verificará esto en R: note que $t^2 = (-0.177)^2 = 0.0312 = F$ para Newspaper.

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: Sales ~ TV + Newspaper
## Model 2: Sales ~ TV + Radio + Newspaper
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    197 1918.56
## 2    196  556.83  1    1361.7 479.33 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se verificará esto en R: note que $t^2 = (21.893)^2 = 479.3 = F$ para Radio.

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: Sales ~ Radio + Newspaper
## Model 2: Sales ~ TV + Radio + Newspaper
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 3614.8
## 2    196  556.8  1      3058 1076.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se verificará esto en R: note que $t^2 = (32.809)^2 = 1076.4 = F$ para TV.

Pero si se tienen los valores p individuales para cada variable, ¿por-qué se quiere mirar el estadístico global F? Después de todo parece probable que si cualquiera de los valores-p para las variables indivi-duales es muy pequeño entonces al menos uno de los predictores está relacionado con la respuesta. Esta lógica no es del todo precisa especialmente cuando el número de predictores p es grande.

Por ejemplo, considere una situación con $p = 100$ y que la hipótesis nula $H_0 : \beta_1 = \cdots = \beta_{100} = 0$ es verdadera (ausencia de asociación). En esta situación, cerca del 5 % de los valores-p asociados con cada variable estarán abajo de 0.05 por el azar. En otras palabras, se espera ver aproximadamente 5 valores-p pequeños aún en ausencia de asociación verdadera entre los predictores y la respuesta. Entonces, si se usan los estadísticos t y sus valores-p para decidir si hay o no una asociación entre la respuesta y los predictores, hay una alta posibilidad de que se concluya erróneamente que hay una relación cuando no la hay.

Pero el estadístico F no sufre de este problema ya que ajusta por el número de predictores p. Entonces, si H_0 es cierta, solo hay un chance de 5 % de que el estadístico F produzca un valor-p abajo de 0.05 sin importar el número de predictores (p) o el número de observaciones (n). Este F funciona bien en la medida que p sea pequeño en comparación a n ¿Qué hacer si $p > n$? En este caso, OLS no funciona adecuadamente y una respuesta plausible a esta pregunta la proporcionan los autores del texto ISLR y se expone más adelante en el curso.

- **¿Todos los predictores ayudan a explicar Y o solamente un subconjunto de los predictores es útil?** Si el estadístico F detecta asociación entre la respuesta y los predictores, el siguiente paso es identificar cuáles son esos predictores. Se pueden examinar los valores-p individuales pero si p es grande, esto podría conducir a identificar erróneamente como asociados con la respuesta a algunos predictores. Es posible que todos los predictores estén asociados con la respuesta pero es más común la situación en la cual solo un subconjunto de los predictores está asociado con la respuesta. El proceso de identificar dicho subconjunto, se conoce como **selección de variables** (variable selection).

Idealmente, se debería llevar a cabo el proceso de selección de variables con base en una gran cantidad de modelos, cada uno conteniendo un subconjunto distinto de predictores. Por ejemplo, si $p = 2$ se deben considerar 4 modelos y seleccionar el mejor de acuerdo a algún criterio: C_p de Mallow, AIC, BIC, y el R^2 ajustado. la mala noticia, es que si hay p predictores habrá un total de 2^p modelos distintos con subconjuntos de las p variables. Por ejemplo, si $p = 30$ hay $2^{30} = 1073741824$ modelos distintos lo cual no es práctico y en algunos casos no es viable computacionalmente.

A menos que p sea pequeño, no se deben considerar todos los posibles modelos, se necesita un método automático y eficiente para seleccionar un conjunto pequeño de modelos. Hay tres aproximaciones clásicas: Selección hacia adelante (Forward selection), Selección hacia atrás (Backward selection) y Selección mixta (Mixed selection o stepwise selection).

- **Selección hacia adelante (Forward selection)**. Se empieza con el modelo nulo (modelo de solo intercepto). Luego se ajustan p modelos de regresión lineal simple y se agrega, al modelo nulo, la variable que produzca el menor RSS. Luego se corren todos los modelos con dos variables (incluyendo siempre la del paso anterior) y se adiciona al modelo la variable que resulte con el menor RSS. Se continua de esta manera hasta que algún criterio de parada se satisfaga.

- **Selección hacia atrás (Backward selection)**. Se empieza con el modelo full (el que incluye todos los predictores), se ajusta y se remueve la variable (solo una) con el valor-p más grande (la menos significativa). Las nuevas $p - 1$ variables se usan para ajustar de nuevo un modelo y de este se remueve el predictor menos significativo. Se continua de esta manera hasta que se satisfaga alguna regla de parada. Por ejemplo, se puede parar cuando los predictores que quedan después de un cierto número de pasos son todos significativos (es decir, tienen un valor-p<0.05).

- **Selección mixta (Mixed selection)**. Esta es una combinación de forward y backward. Se empieza sin variables en el modelo similar a Forward selection y se adiciona la variable con mejor ajuste de acuerdo al RSS (o a algún otro criterio válido). Se continua adicionando variables una por una. Si en algún punto el valor-p para una de las variables en el modelo supera un cierto valor (por ejemplo 0.05), entonces esta variable se remueve del modelo (Backward). Se continua de esta manera (alternando Forward y Backward) hasta que solo queden, si es que quedan, variables significativas (y todas las que queden fuera del modelo tengan un valor-p grande si se adicionaran al modelo).

Ejemplo de Backward, Forward y Stepwise en R. Para implementar estos procedimientos en R, se debe instalar la librería *olsrr*, la cual cuenta con las funciones adecuadas. Considere los datos de Advertising una vez más. Note que los tres procedimientos concuerdan en la selección de variables importantes, aunque en general este no es siempre el caso y lo común es que los tres métodos seleccionen variables diferentes. En esos casos, se debe seleccionar uno de ellos con base en, por ejemplo, su plausibilidad. También es posible que los procedimientos dejen por fuera variables importantes desde el punto de vista de la investigación.

```r
#Example of Backwards, Forward, and Stepwise selection procedures. Class 7
library(ISLR)
library(MASS)
library(olsrr)#Require for selection of variables procedures
```

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

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
Advertising<-read.csv(file="Advertising.csv",
                      header=T,sep=',',dec='.')
Advertising=Advertising[2:5] #To eliminate ID variable
```

```r
#Backward selection

ols_step_backward_aic(lm(sales~., data=Advertising))
```

```
##
##
##                       Backward Elimination Summary
## ---------------------------------------------------------------------
## Variable         AIC        RSS       Sum Sq      R-Sq      Adj. R-Sq
## ---------------------------------------------------------------------
## Full Model    782.362    556.825    4860.323    0.89721       0.89564
## newspaper     780.394    556.914    4860.235    0.89719       0.89615
## ---------------------------------------------------------------------
```

```r
#Forward selection

ols_step_forward_aic(lm(sales~., data=Advertising))
```

```
##
##                          Selection Summary
## ------------------------------------------------------------------------
## Variable       AIC        Sum Sq       RSS        R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------
## TV           1044.091    3314.618    2102.531    0.61188       0.60991
## radio         780.394    4860.235     556.914    0.89719       0.89615
## ------------------------------------------------------------------------
```

```r
#Stepwise selection

ols_step_both_aic(lm(sales~., data=Advertising))
```

```
##
##
##                                Stepwise Summary
## ------------------------------------------------------------------------
## Variable     Method        AIC         RSS         Sum Sq      R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------
## TV          addition    1044.091    2102.531    3314.618    0.61188       0.60991
## radio       addition     780.394     556.914    4860.235    0.89719       0.89615
## ------------------------------------------------------------------------
```

Nota: La selección hacia atrás no se puede usar si $p > n$ mientras que la selección hacia adelante siempre se puede usar. Más adelante en el curso se discute con mucha más profundidad estos métodos.

- **¿Qué tan bien se ajusta el modelo a los datos?** Dos de las medidas numéricas de ajuste de modelos más comunes son el RSE y el R^2, que es la fracción de varianza explicada. Estas cantidades se calculan e interpretan de la misma forma que en Regresión Lineal Simple (RLS). Recuerde que en RLS, el R^2 es el cuadrado de la *correlación* entre la respuesta y la variable. En RLM, sucede que el R^2 es igual a $Cor\left(Y,\ \hat{Y}\right)^2$, el cuadrado de la correlación entre la respuesta y el modelo ajustado; de hecho una propiedad del modelo lineal ajustado es que él maximiza esta correlación entre todos los posibles modelos lineales.

Un R^2 cercano a 1 indica que el modelo explica una gran porción de la varianza en la variable respuesta. Por ejemplo, considere de nuevo los datos de Advertising. Calcule el R^2 con TV, Radio y Newspaper y luego calcúlelo de nuevo pero esta vez sin Newspaper:

Refiérase al ejemplo de Advertising:

```
library(ISLR)
library(MASS)
Advertising<-read.csv(file="Advertising.csv",
                      header=T,sep=',',dec='.')
Sales=Advertising$sales
TV=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
fit1 <- lm(Sales ~ TV+Radio+Newspaper)
fit2<-lm(Sales~TV+Radio)
list(R2_Modelo1=summary(fit1)$r.squared,R2_Modelo2=summary(fit2)$r.squared)
```

```
## $R2_Modelo1
## [1] 0.8972106
##
## $R2_Modelo2
## [1] 0.8971943
```

Note que solo hay un pequeño incremento en el R^2 si se incluye Newspaper en un modelo que ya incluye TV y Radio. El R^2 siempre se incrementa cuando se incluyen más variables en el modelo, aún si las otras variables están asociadas debilmente con la respuesta.

Esto es debido al hecho de que adicionar otra variable a la ecuación de OLS debe permitir ajustar de manera más precisa el conjunto de datos de entrenamiento. De esta manera, el estadístico R^2, que también se calcula con los datos de entrenamiento, se debe incrementar. En el ejemplo anterior, el hecho de adicionar Newspaper al modelo que ya contenía TV y Radio, lleva a solo un incremento muy pequeño en el R^2 proporciona evidencia adicional de que Newspaper se debe eliminar del modelo.

Los tres métodos estuvieron de acuerdo en hacer esto. Newspaper no proporciona una mejora real al ajuste a la muestra de entrenamiento y su inclusión probablemente llevará a resultados pobres en muestras de prueba (test samples) debido a sobreajuste. En contraste, el modelo que contiene solo TV como predictor tiene un R^2 de 0.6119:

```
library(ISLR)
library(MASS)
Advertising<-read.csv(file="Advertising.csv",
                      header=T,sep=',',dec='.')
Sales=Advertising$sales
TV=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
fit1 <- lm(Sales ~ TV)
list(R2_Modelo1=summary(fit1)$r.squared)
```

```
## $R2_Modelo1
## [1] 0.6118751
```

Pero si se adiciona Radio, hay un incremento sustancial en el R^2 estimado:

```
## [1] 1.681361
```

```
## $R2_Modelo2
## [1] 0.8971943
```

Esto implica que un modelo que use TV y Radio para predecir las ventas es mejor que un modelo que use solo TV. Ahora se verá qué pasa con el R^2 cuando se considera solo Radio:

```
## $R2_Modelo3
## [1] 0.3320325
```

Radio por si sola explica una proporción de la varianza muy baja (33.2 % solamente), lo cual resalta la importancia de incluir TV en el modelo para predecir las ventas.

El modelo que contiene solamente TV y Radio como predictores tiene un RSE=1.681361, el modelo con TV, Radio y Newspaper tiene un RSE=1.68551, y el modelo que solo tiene TV un RSE=3.258656:

```
library(ISLR)
library(MASS)
Advertising<-read.csv(file="Advertising.csv",
                      header=T,sep=',',dec='.')
Sales=Advertising$sales
TV=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
fit2 <- lm(Sales ~ TV+Radio)
fit3 <- lm(Sales ~ TV+Radio+Newspaper)
fit4 <- lm(Sales ~ TV)
list(RSE_Modelo2=summary(fit2)$sigma,RSE_Modelo3=summary(fit3)$sigma,RSE_Modelo4=summary(fit4)$sigma)
```

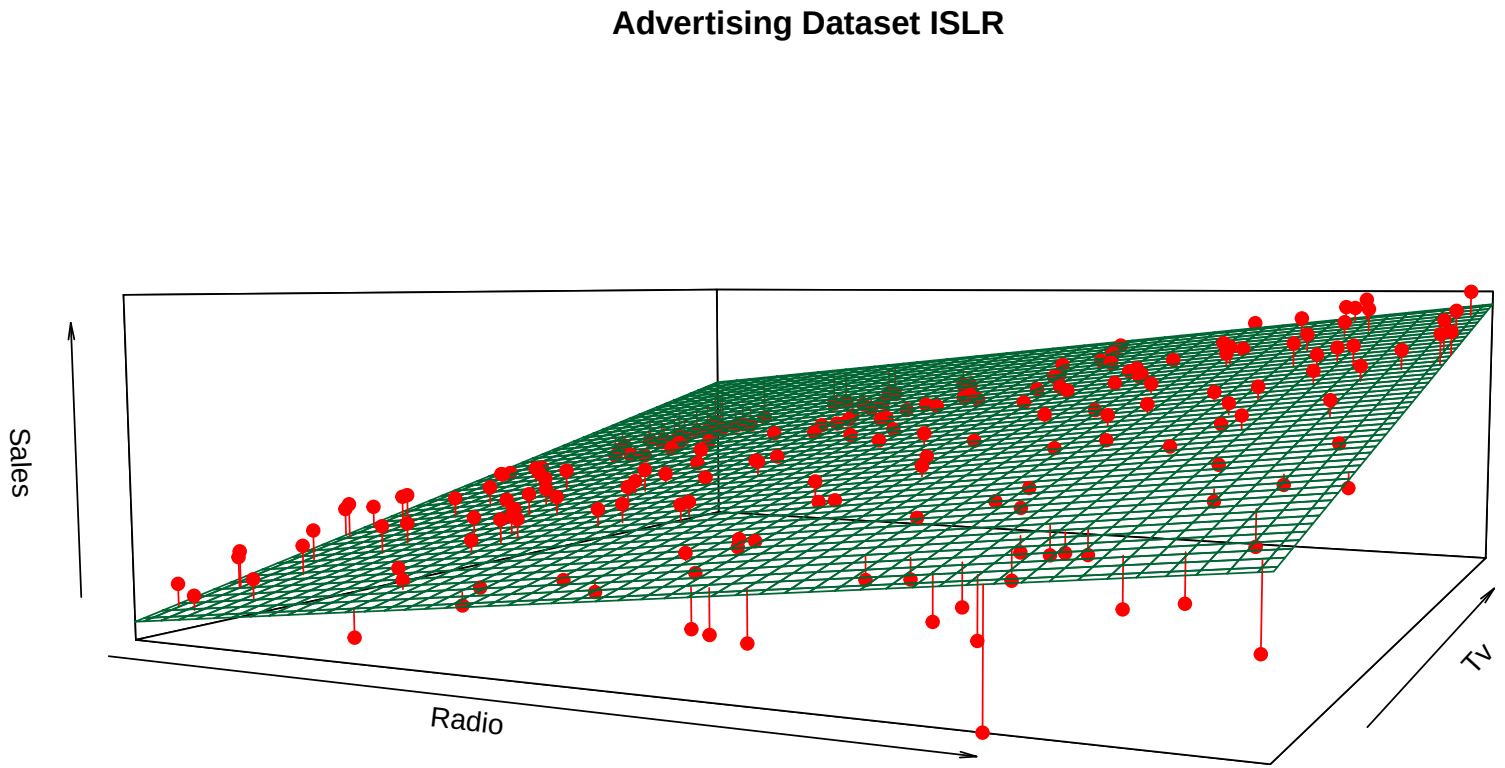
```
## $RSE_Modelo2
## [1] 1.681361
##
## $RSE_Modelo3
## [1] 1.68551
##
## $RSE_Modelo4
## [1] 3.258656
```

Estos resultados corroboran que un modelo con TV y Radio predicen las ventas mucho mejor que aquel que usa solo TV. En general, el RSE se define:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

Por lo tanto, modelos con más variables pueden tener un RSE más grande si la reducción en el RSS es pequeña en relación al incremento en p. Esto se observa cuando se compara 1.681361 con 1.68551.

Otra forma de mirar el ajuste de los datos, adicional a mirar el R^2 y el RSE, es por medio de un gráfico de los datos y el modelo ajustado. Refiérase al ejemplo de Advertising (Sales~Tv+Radio):



- **Dado un conjunto de predictores ¿qué valor de Y se debería predecir y qué tan precisa es esta predicción?** Una vez que se ha ajustado un modelo de regresión múltiple, no es difícil predecir una respuesta Y: se usa $\hat{Y}$ con base en un conjunto de predictores $X_1, X_2, \ldots, X_p$. Sin embargo hay tres fuentes de incertidumbre asociadas a esta predicción:

1. Los coeficientes estimados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ son estimaciones de $\beta_0, \beta_1, \ldots, \beta_p$, respectivamente. Esto es, el plano OLS:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

es **solo una** estimación para el plano poblacional verdadero:

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

La imprecisión en los coeficientes estimados está relacionada con el **error irreducible**. Se puede calcular un IC a fin de evaluar qué tan cerca está $\hat{Y}$ de $f(X)$.

2. Por supuesto, en la práctica asumir un modelo lineal para $f(X)$ es casi siempre una aproximación a la realidad, por lo que hay una fuente adicional potencial de error reducible el cual se denomina **Sesgo**. Entonces, cuando se usa un modelo lineal, de hecho se está estimando la mejor aproximación lineal a la verdadera superficie. Sin embargo, aquí se está ignorando esta discrepancia y se procede como si el modelo lineal fuera correcto.

3. Aún si se conociera a $f(x)$ (es decir, aún si se conocieran los verdaderos valores $\beta_0, \beta_1, \ldots, \beta_p$) la respuesta no podría ser predicha perfectamente debido al error aleatorio ε (o error irreducible). Se usan intervalos de predicción para evaluar qué tanto varía Y con respecto a $\hat{Y}$. Estos intervalos de predicción siempre son más anchos que los IC ya que ellos incorporan el error reducible[1] y el error irreducible[2].

[1] Error en la estimación de $f(X)$

[2] La incertidumbre de qué tanto un punto individual difiere del plano poblacional verdadero.

Ejemplo con datos de Advertising. Dado que se gastan 100000US en TV y 20000US en Radio en cada ciudad y también en una sola ciudad, el IC y el IP serán:

```
library(ISLR)
library(MASS)
Advertising<-read.csv(file="Advertising.csv",
                      header=T,sep=',',dec='.')
Sales=Advertising$sales
TV=Advertising$TV
Radio=Advertising$radio
fit <- lm(Sales ~ TV+Radio)
newdata<-data.frame(TV=100,Radio=20)#Los valores con los que se evaluan los intervalos
CI<-predict(fit,newdata,interval="confidence");length_CI=CI[3]-CI[2]
PI<-predict(fit,newdata,interval="predict");length_PI=PI[3]-PI[2]
list(Confidence_Interval=CI,length_CI=length_CI,Prediction_Interval=PI,length_PI=length_PI)
```

```
## $Confidence_Interval
##        fit      lwr      upr
## 1 11.25647 10.98525 11.52768
##
## $length_CI
## [1] 0.542423
##
## $Prediction_Interval
##        fit      lwr      upr
## 1 11.25647 7.929616 14.58332
##
## $length_PI
## [1] 6.6537
```

Se usa un IC para cuantificar la incertidumbre alrededor del promedio de ventas en un gran número de ciudades. En este caso el IC=[10.985, 11.528] que significa que aproximadamente 95 % de intervalos de esta forma contendrán el verdadero valor de $f(X)$. Por otra parte, un IP se puede usar para cuantificar la incertidumbre alrededor de las ventas para una ciudad particular. En este caso el IP=[7.93, 14.58] que significa que 95 % de intervalos de esta forma contendrán el verdadero valor de Y para esa ciudad donde se gastan esos montos en TV y Radio, respectivamente. El IP es más ancho que el IC lo cual refleja la mayor incertidumbre acerca de las ventas para una ciudad dada en comparación a las ventas promedio en muchas ciudades.

A manera de información, se presentan las fórmulas para ambos tipos de intervalos:

1. **Intervalo de confianza para la respuesta promedio**:

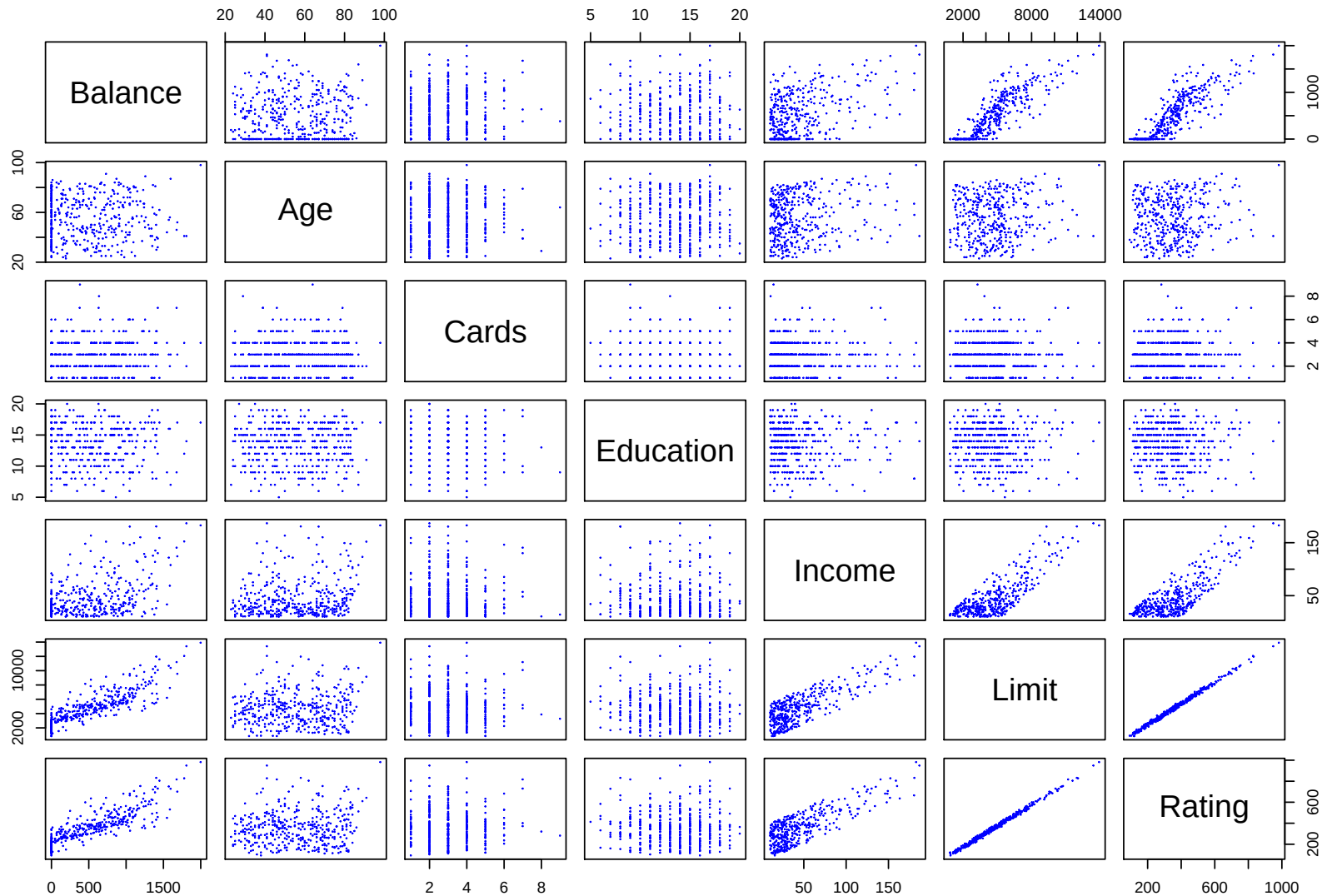
   $\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^T \left(X^T X\right) x_0}$, $\hat{y}_0$ significa que $\hat{y}$ se evalua en x_0

2. **Intervalo de predicción para una nueva observación**:

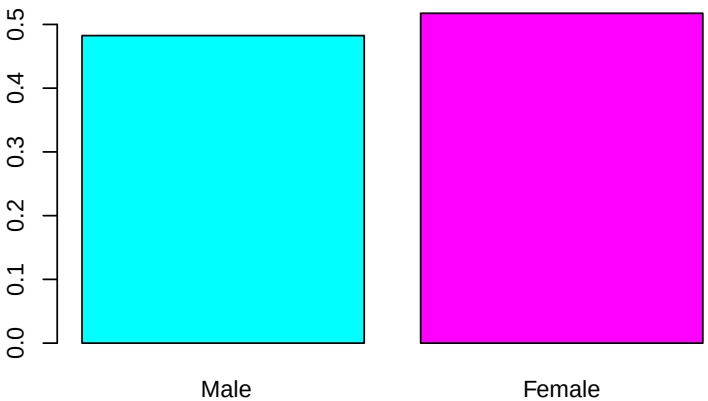
   $\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \left(1 + x_0^T \left(X^T X\right) x_0\right)}$ donde $\hat{y}$ se evalua en x_0

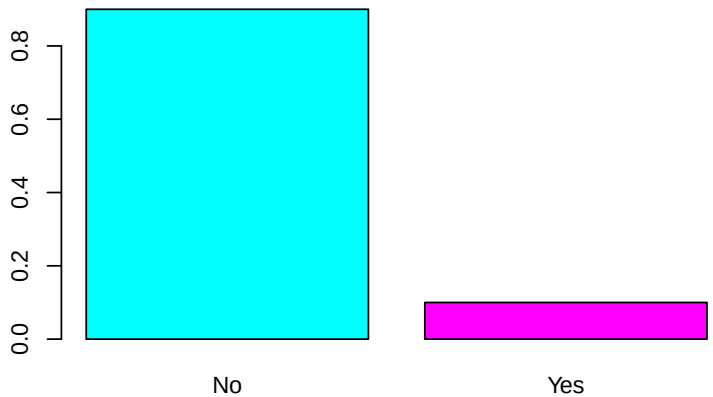
Caso de predictores categóricos o cualitativos. Hasta ahora se ha asumido que todos los predictores son cuantitativos, pero en la práctica este no es necesariamente el caso y frecuentemente aparecen *predictores cualitativos*. Considere unos nuevos datos, los datos Credit (de ISLR) que registra como respuesta $y =$ Balance, y como predictores: Age, Cards (número de tarjetas de crédito), Education (en años), Income (en miles de dolares), Limit (crédito límite), y Rating (tasa de crédito).

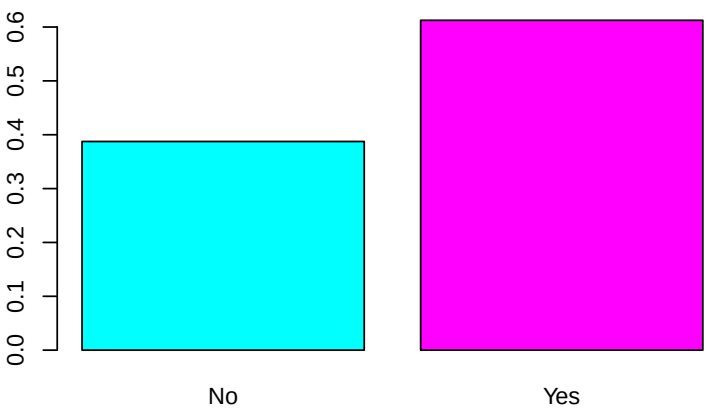
En el siguiente gráfico se muestra un scatterplot para todos los pares de predictores:

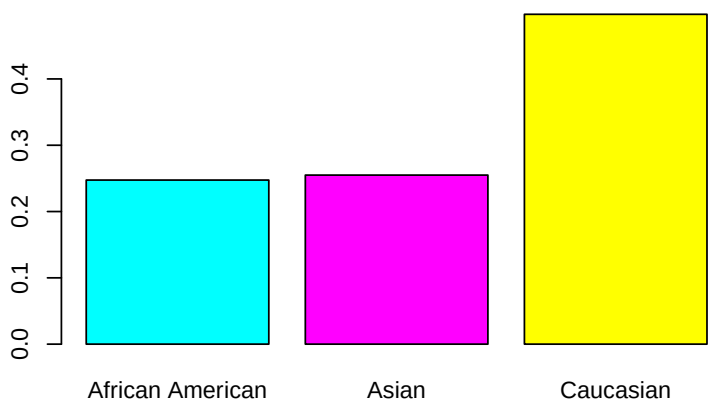


Adicional a estos 6 predictores cuantitativos (pues Balance es la respuesta), se cuenta en la base con 4 variables cualitativas: Gender, Student (Student status), Status (marital status), y Ethnicity (Caucasian, African American or Asian):









RLM con predictores con solo dos niveles. Suponga que se quiere explorar diferencias en balance en tarjetas de crédito entre mujeres y hombres, ignorando las otras variables por el momento. Una variable cualitativa con dos niveles se puede incorporar en un modelo de regresión por medio de una variable indicadora o dummy. Por ejemplo:

$$x_i = \begin{cases} 1 & \text{Si la persona } i \text{ es una mujer} \\ 0 & \text{Si la persona } i \text{ es un hombre} \end{cases}$$

Esta variable genera dos modelos:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{Si la persona } i \text{ es una mujer} \\ \beta_0 + \varepsilon_i & \text{Si la persona } i \text{ es un hombre} \end{cases}$$

β_0 se puede interpretar como el balance promedio de la tarjeta de crédito entre los hombres mientras que $\beta_0 + \beta_1$ es el promedio de la tarjeta de crédito entre las mujeres y β_1 es la diferencia promedio en el balance de la tarjeta de crédito entre mujeres y hombres. La siguiente tabla muestra los coeficientes estimados junto con otra informacion:

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  509.80311   33.12808 15.3888531 2.908941e-42
## GenderFemale  19.73312   46.05121  0.4285039 6.685161e-01
```

De la tabla anterior, el balance promedio para hombres es de 509.80US mientras que para mujeres es $509.8+19.73 = 529.53$US. Sin embargo, note que el valor-p para la variable dummy es muy alto, indicando que no hay diferencia en los balances promedio entre hombres y mujeres.