Estadística II - Taller 02 Semestre: 2024-01

Profesores: Johnatan Cardona Jimenez, Freddy Hernández Barajas, Raul Alberto Perez

Monitor: Ronald Palencia

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

Solución parte teorica

Punto 1

1.1 *Falso.* Un alto valor de R^2 no garantiza que el modelo pueda hacer predicciones útiles. Un modelo podría tener un alto R^2 debido al sobreajuste, lo cual significa que se ajusta muy bien a los datos de la muestra pero no necesariamente a nuevos datos

1.2 *Falso.* Un alto valor de R^2 sugiere que la recta de regresión se ajusta bien a los datos observados, pero no prueba que el ajuste sea el mejor posible. Es necesario realizar pruebas de ajuste adicionales, como la prueba F para regresión, para confirmar la bondad del ajuste.

1.3 *Falso.* Un R^2 cercano a cero sugiere que el modelo no explica bien la variabilidad de la respuesta en torno a su media. Sin embargo, esto no implica directamente que X y Y no estén relacionadas. Podría ser que la relación entre ellas no sea lineal o que el modelo no haya capturado todas las variables relevantes.

1.4

$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-2} \times \sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$

1.5 Estimación puntual y por intervalo de la respuesta media:

La estimación puntual de la media condicional de Y dado un valor específico de X, denotado x_0, se calcula como

$$\hat{E}[Y|x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

. El intervalo de confianza para esta estimación puntual se calcula como:

$$\hat{E}[Y|x_0] \pm t_{\frac{\alpha}{2},n-2} \times \text{SE}(\hat{E}[Y|x_0])$$

donde $\text{SE}(\hat{E}[Y|x_0])$ es el error estándar de la estimación.

1.6 Predicción de valores futuros: Para predecir un nuevo valor de Y, denotado y_0, para un valor específico de X, denotado x_0, se utiliza la ecuación de la recta de regresión estimada:

Nota, poner la grafica

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Y el intervalo de predicción para este valor futuro se calcula como:

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-2} \times \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

donde MSE es el error cuadrático medio, n es el tamaño de la muestra, $\bar{x}$ es la media de los valores de X, y S_{xx} es la suma de los cuadrados de las diferencias de los valores de X respecto a su media.

1.7

Table 1: Tabla ANOVA completada

Fuente de Variación	SS (Suma de Cuadrados)	df (Grados de Libertad)	MS (Cuadrado Medio)	F
Entre Grupos	10.25	1	10.25	27.33
Dentro de Grupos	20.50	28	0.375	
Total	30.75	29		

Parte practica

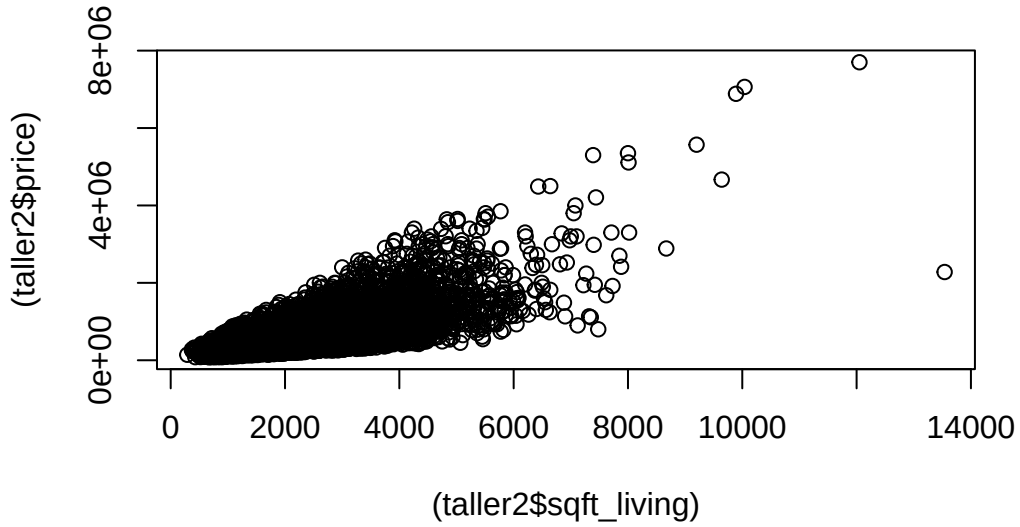
Literal 1

```
# Carga el paquete 'dplyr' si no ha sido cargado previamente.
# Es útil para la manipulación de datos y se asume que está instalado.
library(dplyr)

# 'data2' es un nuevo objeto de datos que se crea al leer el archivo CSV 'kc_house_data.csv'
# Este archivo CSV debe estar en el directorio de trabajo actual de R o se debe proporcionar
data2 = read.csv("kc_house_data.csv")

# 'taller2' es un nuevo objeto de datos que se crea a partir del objeto 'data2'.
```

```r
# Se utiliza la función 'select' del paquete 'dplyr' para seleccionar solo las columnas 'pri
# Esto crea un nuevo dataframe con solo estas dos columnas de interés.
taller2 = data2 %>%
  select(price, sqft_living)
```

Literal 2

```r
plot((taller2$sqft_living), (taller2$price))
```



Relación Positiva: Se observa una tendencia positiva en los datos, lo que indica que, en general, a medida que el tamaño de la vivienda en pies cuadrados aumenta, también lo hace el precio de la propiedad. Esto sugiere que sqft_living podría ser un buen predictor del price.

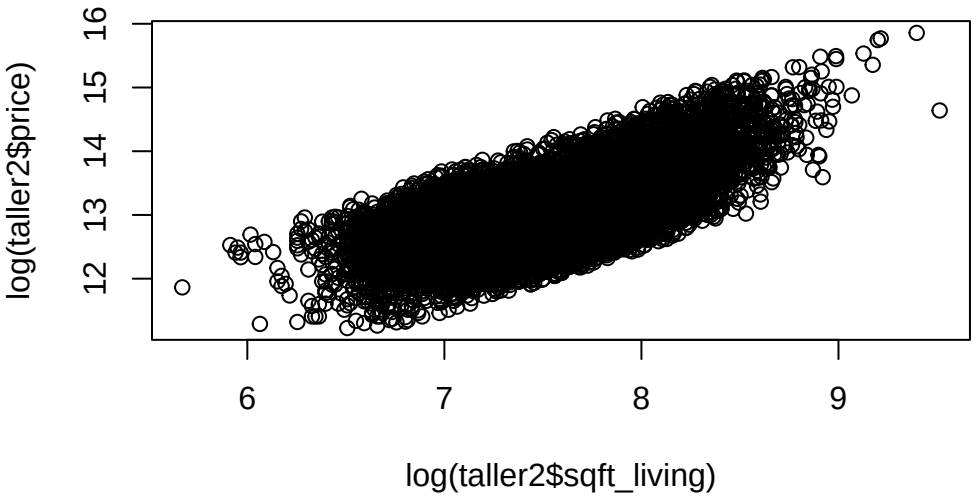
Concentración de Datos: La mayoría de los datos están concentrados en el rango inferior de sqft_living y price, lo que indica que la mayoría de las viviendas tienen tamaños y precios menores en comparación con unas pocas propiedades de mayor tamaño y precio.

Posibles Outliers: Hay algunos puntos que se desvían significativamente de la tendencia general y se sitúan lejos de la concentración principal de datos. Estos puntos podrían considerarse valores atípicos y podrían tener un impacto significativo en la regresión si no se manejan adecuadamente.

Heterocedasticidad: Parece que la varianza de los precios aumenta a medida que aumenta sqft_living. Esto podría ser un indicativo de heterocedasticidad, lo que significa que el uso de modelos de regresión lineal estándar podría no ser adecuado sin transformaciones o métodos estadísticos que tomen en cuenta la heterocedasticidad.

Transformación de Datos: Dada la forma del gráfico, podría ser útil transformar las variables antes de realizar un análisis de regresión, tal como aplicar logaritmo a ambas variables para normalizar la distribución y estabilizar la varianza de los errores.

```
plot(log(taller2$sqft_living), log(taller2$price))
```



Literal 3

Modelo general

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Para seleccionar el 80 de los tados usaremos el siguiente código y luego se usara la función lm para ajustar el modelo:

```
# Establece una semilla para hacer reproducible la aleatorización.
# Esto es útil para garantizar que los resultados sean consistentes en múltiples ejecuciones
set.seed(123)

# 'indices' es un vector que contiene una muestra aleatoria del 80% de los índices de las fi
# 'nrow(taller2)' calcula el número total de filas en 'taller2'.
# 'sample' toma una muestra aleatoria de estos índices.
indices = sample(1:nrow(taller2), size = 0.8*nrow(taller2))

# 'train_data' es un nuevo conjunto de datos que contiene solo las filas de 'taller2' indexa
# Este será el conjunto de datos de entrenamiento usado para ajustar el modelo.
train_data = taller2[indices, ]

# 'test_data' es otro nuevo conjunto de datos que contiene las filas restantes de 'taller2' c
```

```r
# Este será el conjunto de datos de prueba utilizado para validar el modelo.
test_data = taller2[-indices, ]

# 'modelo2' es un modelo de regresión lineal ajustado utilizando la función 'lm' en R.
# Este modelo predice 'price' como una función lineal de 'sqft_living' usando solo el conjunt
modelo2 = lm(train_data$price~train_data$sqft_living)
```

```r
summary(modelo2)
```

```
Call:
lm(formula = train_data$price ~ train_data$sqft_living)

Residuals:
     Min       1Q   Median       3Q      Max
-1465693  -147174   -24984   106153  4371080

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -41627.883   4889.428  -8.514   <2e-16 ***
train_data$sqft_living    279.714      2.151 130.055   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 259900 on 17288 degrees of freedom
Multiple R-squared:  0.4945,    Adjusted R-squared:  0.4945
F-statistic: 1.691e+04 on 1 and 17288 DF,  p-value: < 2.2e-16
```

La ecuación del modelo de regresión lineal simple ajustado es:

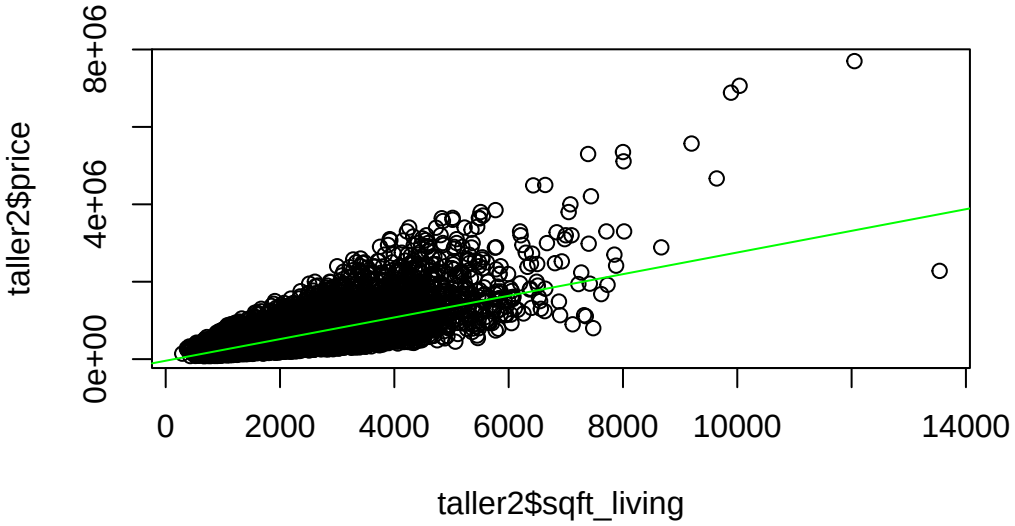
$$\text{price} = -41627.883 + 279.714 \times \text{sqft\_living}$$

Donde: - El intercepto (o término constante) es -41627.883. - La pendiente (o coeficiente para la variable `sqft_living`) es 279.714.

Para poner la liena ajustada usaremos la funcion abline del R.

```r
# Crea un diagrama de dispersión con 'sqft_living' en el eje x y 'price' en el eje y.
plot(taller2$sqft_living, taller2$price)

# Añade una línea de regresión al diagrama de dispersión existente.
```

```r
# 'modelo2' contiene el modelo de regresión lineal que se ajustó previamente.
# El color de la línea de regresión se establece en verde.
abline(modelo2, col="green")
```



Tendencia Lineal: La recta de regresión indica una tendencia lineal positiva entre las dos variables. Esto sugiere que, en promedio, a medida que aumenta el área habitable de una propiedad, también lo hace su precio.

Densidad de Datos: La mayoría de los puntos de datos están agrupados en el extremo inferior del eje x (área habitable) y del eje y (precio), lo que sugiere que la mayoría de las propiedades en la muestra tienen un área habitable menor y son menos costosas.

Valores Atípicos (Outliers): Hay varios puntos que se encuentran lejos de la concentración principal de datos y de la línea de regresión. Estos valores atípicos podrían ser propiedades que son inusualmente caras para su tamaño o que tienen características únicas que aumentan su precio.

Ajuste del Modelo: La línea de regresión parece capturar la tendencia general de los datos, pero debido a la dispersión de los puntos y la presencia de valores atípicos, es posible que el modelo no capture todas las complejidades de la relación entre el área habitable y el precio. Esto puede ser un indicativo de heterocedasticidad, donde la varianza de los errores del modelo no es constante.

Literal 4

Las hipótesis para la prueba de significancia de la pendiente y el análisis de varianza son:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Donde β_1 es el coeficiente de la variable independiente `sqft_living`.

Prueba de Significancia para la Pendiente (Prueba t)

La prueba t para la pendiente se realiza para determinar si hay suficiente evidencia para rechazar la hipótesis nula de que la pendiente es igual a cero. Los resultados de la prueba t para la pendiente se presentan de la siguiente manera:

```
summary(modelo2)
```

```
Call:
lm(formula = train_data$price ~ train_data$sqft_living)

Residuals:
     Min       1Q   Median       3Q      Max
-1465693  -147174   -24984   106153  4371080

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -41627.883   4889.428  -8.514   <2e-16 ***
train_data$sqft_living    279.714      2.151 130.055   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 259900 on 17288 degrees of freedom
Multiple R-squared:  0.4945,    Adjusted R-squared:  0.4945
F-statistic: 1.691e+04 on 1 and 17288 DF,  p-value: < 2.2e-16
```

Valor de la pendiente estimada $\hat{\beta}_1 = 279.714$

Error estándar de la pendiente (SE): $= 2.151$

Valor t: $= 130.055$

Valor p: $< 2 \times 10^{-16}$

Prueba de Significancia de la Regresión (Análisis de Varianza - Prueba F)

El análisis de varianza (ANOVA) se utiliza para determinar si la variabilidad entre grupos (en este caso, debido a la regresión) es mayor que la variabilidad dentro de los grupos (residuos). Los resultados de ANOVA se presentan así:

```
anova(modelo2)
```

```
Analysis of Variance Table

Response: train_data$price
                          Df     Sum Sq    Mean Sq F value    Pr(>F)
train_data$sqft_living     1 1.1427e+15 1.1427e+15   16914 < 2.2e-16 ***
Residuals              17288 1.1680e+15 6.7559e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$ Suma de cuadrados debido a la regresión (SSR): = 1.1427 ×10^{15}\ Suma de cuadrados de los residuos (SS
= 1.1680 ×10^{15} Grados de libertad asociados con la regresión (df): = 1 \ Grados de libertad de los residuos:
= 17288 \ Valor F: = 16914 \ Valor p: < 2 ×10^{-16}

$$

Ambas pruebas, la prueba t para la pendiente y la prueba F en el análisis de varianza, llevan a la misma conclusión: rechazar la hipótesis nula y concluir que hay una relación lineal estadísticamente significativa entre sqft_living y price.

Relación entre la Prueba t y la Prueba F

Ambos enfoques, la prueba t para la pendiente y la prueba F en el análisis de varianza, se utilizan para determinar la significancia estadística de la relación entre las variables independientes y dependientes. En el contexto de la regresión lineal simple, donde solo hay una variable independiente, ambos llegan a la misma conclusión y son matemáticamente relacionados. El cuadrado del valor t para la pendiente es igual al valor F cuando hay una sola variable predictora:

$t^2 = F$

Por lo tanto, si la pendiente es significativa en la prueba t, la regresión será significativa en la prueba F. En este caso, como ambas pruebas dan un valor p muy bajo, podemos concluir con confianza que la relación entre el área habitable y el precio es estadísticamente significativa.