Estadística II - Taller 02 Semestre: 2024-01

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

Monitor: Ronald Palencia

El conjunto de datos (kc_house_data.csv) utilizado en este análisis contiene información sobre los precios de los casas y sus características asociadas. Aquí hay un breve resumen del conjunto de datos:

- id: Un identificador único para cada propiedad listada.
- date: La fecha en la que la casa fue vendida.
- price: El precio de venta de la casa.
- bedrooms: El número de dormitorios en la casa.
- bathrooms: El número de baños en la casa, a menudo en formato decimal para representar baños parciales (por ejemplo, 1.5 para un baño completo y un aseo).
- sqft_living: El área habitable en pies cuadrados de la casa.
- sqft_lot: El tamaño total del terreno en pies cuadrados.
- floors: El número de pisos en la casa.

Para el desarrollo de este taller trabajaremos con las variables Price y sqft_living.

A continuación se muestran las primeras 6 observaciones o registros del conjunto de datos

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
7129300520	2014-10-13	221900	3	1.00	1180	5650	1
6414100192	2014-12-09	538000	3	2.25	2570	7242	2
5631500400	2015-02-25	180000	2	1.00	770	10000	1
2487200875	2014-12-09	604000	4	3.00	1960	5000	1
1954400510	2015-02-18	510000	3	2.00	1680	8080	1
7237550310	2014-05-12	1225000	4	4.50	5420	101930	1

Parte teorica

1 Respoder falso o verdadero y argumentar en caso de ser verdadero o dar un contra ejemplo en caso de ser falso

1.1 Un R^2 alto indica que el modelo puede hacer predicciones útiles.

1.2 Un R^2 alto indica que la recta de regresión tiene buen ajuste.

1.3 Un R^2 cercano a cero indica que X y Y no están relacionados.

1.4 La formula del intervalo de predicción es:

a) $\hat{y}_0 \pm t_{\alpha/2,n-2} \times \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$
b) $\hat{y}_0 \pm t_{\alpha/2,n-5} \times \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$
c) $\hat{y}_0 \pm t_{\alpha/2,n-2} \times \sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$
d) $\hat{y}_0 \pm t_{\alpha,n-3} \times \sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$

1.5 Estimación puntual y por intervalo de la respuesta media $E[Y|x_0] + \varepsilon_0$

1.6 Predicción de valores futuros $y_0 = \beta_0 + \beta_1 x_0 = E[Y|x_0]$

1.7 Sólo se podrán hacer inferencias sobre la respuesta cuando $X = x_0 \in [X_{\min}, X_{\max}]$, donde $X_{\min}$ y $X_{\max}$ son los valores mínimo y máximo de la variable predictora, que fueron fijados en la muestra.

2 Completar la siguiente tabla anova.

Table 2: Tabla ANOVA parcialmente completada

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

Parte Practica

Considere el area cuadrada habitable de las propiedades como la covariable (X) y el precio como la variable respuesta (Y) para responder las preguntas de la 1 a la 7.

1. Realice la lectura de la base de datos, seleccione únicamente las variables numéricas.
2. Elabore un gráfico de dispersión de las variables para encontrar aquella que presente una mejor relación lineal con respecto a la variable respuesta.

3. Escriba la ecuación del modelo de regresión, junto con sus supuestos. Ajuste un modelo de regresión lineal simple y añada la recta de regresión a la gráfica generada anteriormente. **Nota**: seleccione aleatoriamente el 80% de los datos para ajustar el modelo.

4. Realice la prueba de significancia para la pendiente, luego realice la prueba de significancia de la regresión usando análisis de varianza. ¿Ambos enfoques permiten llegar a la misma conclusión? ¿Qué relación existe entre una prueba y la otra?

5. De una interpretación de los parámetros β_0 y β_1 del modelo, claro está, si es posible hacerlo.

6. Calcule el R^2 usando el coeficiente de correlación y usando sumas de cuadrados, compare estos entre sí y compárelos con las salidas de R. Realice una interpretación de este.

7. Use el modelo para predecir los precios de las viviendas del 20% de los datos que NO usó para ajustar el modelo. Calcule los respectivos intervalos de confianza y de predicción. ¿Cuáles intervalos son más anchos? ¿Por qué cree usted que esto sucede?

8. **Tarea:** Repetir los literales del 1 al 7 con las variables Price y sqft_lot (Cambiar sqft_living por sqft_lot)