Ejemplos de Aplicación - Semana 4

Cadena de comidas rápidas

En una cadena de comidas rápidas se dispone de información de los m^2 y facturación anual para sus 27 sucursales. Se desea establecer la superficie es un buen predictor de la facturación anual. Los datos recolectados son los siguientes:

Datos: *Ejemplo_Aplicacion_Semana_4.xlsx* Hoja: *Problema1*

Cargamos las librerías

```
# Para importar datos de excel
library(readxl)
# Para graficos más profesionales
library(ggplot2)
# Librerias

library(dplyr)
```

Cargamos los datos del archivo

```
# IMPORTAR DATOS
datos <- read_excel("Ejemplo_Aplicacion_Semana_4.xlsx",
                    sheet = "Problema1")

# Le cambio el nombre a las columnas/variables para que sean mas sencillos
colnames(datos) <- c("superficie", "ventas")
```

a) Graficar las variables registradas. Plantear la relación causal.

Generamos el gráfico

```r
# Gráfico de dispersion entre la variable explicativa y la de respuesta (mas profesional)
ggplot(datos, aes(x = superficie, y = ventas)) +
  geom_point() +
  # fijamos los límites inferiores de los ejes en 0.
  # NA = Not avaiable. Dejamos que R decida la mejor opción
  xlim(0, NA)+
  ylim(0, NA)+
  # Establecemos tema bw (estilo de gráfico mas sobrio)
  theme_bw() +
  # Agregamos en gris ejes cartesianos
  geom_vline(xintercept = 0, color = "grey")+
  geom_hline(yintercept = 0, color = "grey")
```

b) Estimar y evaluar el modelo $Ventas = \beta_0 + \beta_1 Superficie + \epsilon$. Grafique las observaciones junto con el modelo propuesto. Es válido el modelo desde el punto de vista estadístico. ¿Que opina de este modelo? Grafique el diagrama de dispersión juntamente con la recta de regresión y los límetes del intervalo de prediccion al 95%.

```r
modelo1 <- lm(formula = ventas ~ superficie, data = datos)
resumen1 <- summary(modelo1)
resumen1
```

```r
# (5:90)/10 genera un vector con todos los numeros de 5 a 90 divididos por 10. Es decir 0.
datos_grafico1 <- data.frame(superficie = (5:90)/10)
predict1 <- predict(modelo1, newdata = datos_grafico1,
                    interval = "prediction")
datos_grafico1 <- cbind(datos_grafico1, predict1)

# Vemos en pantalla el dataset generado (5 filas)
head(datos_grafico1, n = 5)

# Gráfico de dispersion entre la variable explicativa y la de respuesta
# No indicamos el dataset en ggplot() porque tenemos que tomar información de distintos da
ggplot() +
  geom_point(aes(x = superficie, y = ventas), data = datos) +

  # fijamos los límites inferiores de los ejes en 0.
  # NA = Not avaiable. Dejamos que R decida la mejor opción
```

```
  xlim(0, NA)+
  ylim(0, NA)+
  # Agregamos en gris ejes cartesianos
  geom_vline(xintercept = 0, color = "grey")+
  geom_hline(yintercept = 0, color = "grey")+

  geom_line(aes(superficie, lwr), data = datos_grafico1, color = 'red')+
  geom_line(aes(superficie, upr), data = datos_grafico1, color = 'red')+

  # Agregamos el gráfico del modelo lineal
  geom_smooth(aes(x = superficie, y = ventas), data = datos,
              se = FALSE, method = lm)+
  # Establecemos tema bw (estilo de gráfico mas sobrio)
  theme_bw()
```

c) Calcule los valores de h_{ii} y los residuos estudentizados r_i para los valores de la muestra original (sin excluir la observacion sospechosa). ¿Observa algún valor elevado para alguna observación?

```
library(MASS)

datos$residuos_estudentizados <- studres(modelo1)

datos$hii <- hatvalues(modelo1)

head(datos, 27)
```

d) La observación ubicada en el extremo derecho genera sospechas. Extraiga el punto de la muestra. Reestime el modelo. ¿Es este modelo mejor que el anterior? ¿Por qué?

```
# Revisando los datos observamos que la observación e cuetionada es la que está en la fila
datos_reducidos <- datos[-27, ]
modelo2 <- lm(formula = ventas ~ superficie, data = datos_reducidos)
resumen2 <- summary(modelo2)
resumen2
```

e) Grafique ambos modelos de regresión en el mismo gráfico. ¿Considera la observación en el extremo derecho es un punto influyente? ¿Por qué? Observando los resultados del los modelos estimados, ¿Cuanto se modifica la pendiente de la recta si excluimos la observación? Estime los $DFBETAS$ utilizando la funciones de R *dfbeta* (devuelve las diferencias absolutas de los coeficientes estimados) y *dfbetas* (devuelve las diferencias estandarizadas de los coeficientes estimados). Interprete la diferencias absoluta en términos del problemas. Comparar el valor estandarizado con el límite para deerminar si se trata de in puntoo influyente.

```
# Gráfico de dispersion entre la variable explicativa y la de respuesta
# No indicamos el dataset en ggplot() porque tenemos que tomar información de distintos da
ggplot() +
  geom_point(aes(x = superficie, y = ventas), data = datos) +

  # fijamos los límites inferiores de los ejes en 0.
  # NA = Not avaiable. Dejamos que R decida la mejor opción
  xlim(0, NA)+
  ylim(0, NA)+
  # Agregamos en gris ejes cartesianos
  geom_vline(xintercept = 0, color = "grey")+
  geom_hline(yintercept = 0, color = "grey")+

  # Agregamos el gráfico del modelo lineal  ORIGINAL en rojo (nótese que data = datos)
  geom_smooth(aes(x = superficie, y = ventas), data = datos,
              se = FALSE, method = lm, color = 'red')+
    # Agregamos el gráfico del modelo lineal  ORIGINAL en rojo (nótese que data = datos_re
  geom_smooth(aes(x = superficie, y = ventas), data = datos_reducidos,
              se = FALSE, method = lm, color = 'blue')+

  # Establecemos tema bw (estilo de gráfico mas sobrio)
  theme_bw()

# Nos interesa la observación 27
dfbeta(modelo1)
dfbetas(modelo1)
```

f) Para determinar si una observación es compatible con un determinado modelo, es frecuente seguir el siguiente procedimiento. Indique si la observación de la derecha es compatible con este modelo.

- Extraer de la muestra la observación cuestionada
- Estimar el modelo
- Calcular el Intervalo de prediccion con los datos restantes
- Si la observacion cuestionada cae dentro del intervalo de predicción, se dice que no hay indicios estadísticos de que la observación sea incompatible con el modelo estimado. Caso contrario, nos indica que estadísticamente la observación no es compatible con el modelo.

g) Para quitar observaciones del dataset utilizado para estimar un modelo, no alcanza con que estadísticamente sea sospechoso. Se rquiere además indicios estraestadísticos que permitar determinar el curso de acción a seguir. Para la observación de la derecha, usted consultó en la empresa y le indicaron que esa observación corresponde a la casa central y que posee las mesas en un parque abierto a diferencia del resto de las sucursales. ¿Qué debería hacer? ¿Quitaría la observación del modelo? Dicuta, la respuesta no necesariamente es única.

Problema del catalizador

En un proceso de químico se busca hallar la relación existente entre la fracción remanente de un catalizador en un producto químico [*masadecatalizador/masadelproductoqumico*] en función del tiempo de reacción. Para ello se realizaron mediciones para distintos tiempos [*minutos*] de reacción.

Datos: *Ejemplo_Aplicacion_Semana_4.xlsx* Hoja: *Problema12*

a) Estimar y validar el modelo lineal

```
# IMPORTAR DATOS
datos <- read_excel("Ejemplo_Aplicacion_Semana_4.xlsx",
                    sheet = "Problema2")
# datos[27, 2] <- 550
# datos[27, 1] <- 7.5
# Le cambio el nombre a las columnas/variables para que sean mas sencillos
colnames(datos) <- c("tiempo", "fraccion")

modelo1 <- lm(formula = fraccion ~ tiempo, data = datos)
summary(modelo1)
```

b) Realice un análisis de los residuos y determine si existen outliers y/o puntos influyentes. En caso de encontrar alguno, de una posible razón para que esto ocurre y evalúe quitar el/los punto/s del modelo. Luego compare ambos modelos, determine cual es mejor e interprete los parámetros del mismo.

```
# Agregamos a los datos una comuna con los y_pico y otra con los residuos estudentizados
datos$y_pico = predict(modelo1)
datos$resid_stud = studres(modelo1)

ggplot(datos) +
  geom_point(aes(y_pico, resid_stud))+

  # Agregamos lineas horizontales para determinar limites +-2
  geom_hline(yintercept = 2, color = 'red')+
  geom_hline(yintercept = -2, color = 'red')+

  geom_hline(yintercept = 0, color = 'black')+
  theme_bw()
```

c) Estudie la linealidad del modelo. Si encuentra algún problema, proponga una solución y compare el resultados. Observe el valor de R^2 del nuevo modelo propuesto.

```
# Agregamos a los datos una comuna con los y_pico y otra con los residuos estudentizados
ggplot(datos) +
  geom_point(aes(tiempo, fraccion))+
  xlim(0, NA)+
  ylim(0, NA)+
  theme_bw()

ggplot(datos) +
  geom_point(aes(tiempo, log(fraccion)))+
  xlim(0, NA)+
  # ylim(0, NA)+
  theme_bw()

ggplot(datos) +
  geom_point(aes(log(tiempo), fraccion))+
  xlim(0, NA)+
  ylim(0, NA)+
```

```
  theme_bw()

ggplot(datos) +
  geom_point(aes(log(tiempo), log(fraccion)))+
  xlim(0, NA)+
  # ylim(0, NA)+
  theme_bw()
```

```
modelo2 <- lm(formula = log(fraccion) ~ log(tiempo), data = datos)
summary(modelo2)
```

```
# Agregamos a los datos una comuna con los y_pico y otra con los residuos estudentizados
datos2 <- datos
datos2$y_pico = predict(modelo2)
datos2$resid_stud = studres(modelo2)

ggplot(datos2) +
  geom_point(aes(y_pico, resid_stud))+

  # Agregamos lineas horizontales para determinar limites +-2
  geom_hline(yintercept = 2, color = 'red')+
  geom_hline(yintercept = -2, color = 'red')+

  geom_hline(yintercept = 0, color = 'black')+
  theme_bw()
```

d) Suponga que detecta un error de medición en la observación número 6. Quite la observación y vuelva a correr el modelo.

```
datos <- datos[-6, ]
modelo2 <- lm(formula = log(fraccion) ~ log(tiempo), data = datos)
summary(modelo2)
```

```
# Agregamos a los datos una comuna con los y_pico y otra con los residuos estudentizados
datos2 <- datos
datos2$y_pico = predict(modelo2)
datos2$resid_stud = studres(modelo2)
```

```r
ggplot(datos2) +
  geom_point(aes(y_pico, resid_stud))+

  # Agregamos lineas horizontales para determinar limites +-2
  geom_hline(yintercept = 2, color = 'red')+
  geom_hline(yintercept = -2, color = 'red')+

  geom_hline(yintercept = 0, color = 'black')+
  theme_bw()
```

e) De acuerdo con la especificación del producto, el contenido final de catalizador, en ningún caso debe superar el 4% en masa. Estime en que proporción de casos no se cumpliría la norma si se interrumpe el proceso en 25 minutos

```r
new_data <- data.frame(tiempo = 25)

predict(modelo2, newdata = new_data, interval = "prediction")

# A la semiamplitud del intervalo la divido por el valor de la t-student y obtengo el desv
desvio_estimado <- (-3.0823 - (-4.149751)) / 2 / qt(.975, 43-2)

# Ahora estimo la probalidad solicitada
dt((log(.04) - (-3.616025))/desvio_estimado, 43-2)
```