R 3.2. Estimación de regresión lineal simple Cheat Sheet

by julenx via cheatography.com/168626/cs/35527/

Información de lm

ols = lm(formula,data=..., subset=...)

> **y_x** (dependienteindependiente) para línea recta con intercepto ($\beta 0$)
> **y ~ -1 + x** para línea recta sin interc-epción
> **data:** es un parámetro opcional que sirve para especificar el dataframe al que nos referimos
> **subset:** sirve para especificar que la regresión sólo tenga en cuenta un subconjunto de las observaciones, definido mediante alguna condición lógica.

+ Cálculo manual de los coeficientes de regresión $\hat{\beta}=(\hat{\beta}_{0},\hat{-\beta}_{1})'$ a partir de la **ecuación formal**

...

3. Valores ajustados de la variable dependiente

ols$fitted

ols$fitted.values

fitted(ols)

> Extracción directa

6. Suma de cuadrados de los residuos

Denotado por SSR o sigma hat. cuantifica cuánta información de la variable dependiente se pierde con el modelo empleado.

output$sigma

> Extracción directa

ssr = sqrt(deviance(ols)/gdl)

> Extracción manual

resi2 = (residuos)^2

sigma = sqrt(sum(resi2)/gdl)

> Extracción manual

9. Intervalos de confianza en coefs. de regresión

ci.bhat <- confint(ols)

> Extracción manual. Nivel de confianza por defecto del 5%.

ci.bhat2 <- confint(ols, level = 0.9)

> Ejemplo para un nivel de confianza del 90%

Selección de modelos de regresión lineal simple

Selección del mejor modelo de regresión lineal simple que se ajusta a un dataset

Se cambia la variable independiente entre todas las que se consideren para ver con cuál obtenemos un mejor ajuste. Dos métodos para decidirlo

1. Comparación del **coeficiente de determ-inación ajustado**: el modelo con un mayor R^2_{adj} será el mejor de los compar-ados. Este coeficiente ya hemos visto como obtenerlo.

2. Comparación de los **criterios de inform-ación** de Akaike (AIC) y Bayesiano (BIC): cuanto menor sea el valor, mejor será el ajuste realizado.

Medir ajuste

require(knitr)

R2=c(summary(mod1)$adj.r.squared, summary(mod2)$adj.r.squared)

AIC=c(extractAIC(mod1)[2],extractAIC(mo-d2)[2])

BIC=c(extractAIC(mod1,k=log(nrow(wag-e1)))[2],extractAIC(mod2,k=log(nrow(wag-e1)))[2])

Medidas = data.frame(R2,AIC,BIC,row.n-ames=c("modelo1","modelo2"))

names(Medidas) = c("Coef.R2adj","AIC","B-IC")

knitr::kable(Medidas)

Análisis de la existencia de relación lineal

1. Regression output

output = summary(ols)

> **Columna Std Error:** Errores típicos de los parámetros estimados β 0 y β 1.
> **Columna t value:** Resultado de dividir cada estimado (estimate) entre su error típico. Es la base para llevar a cabo los contrastes de significatividad sobre los parámetros estimados $H0:\beta$ 0=0 y ☐ ☐0:β 1=0.
> **Columna Pr(>|t|)** p-valores. Si son menores que el nivel de signifcación por defecto (5%), se rechaza la hipótesis nula y por tanto, es estadísticamente significativo el coeficiente estimado para β1 o β0 .
> **Residual standard error:** (estimador de la desviación típica de los errores σ). Los degrees of freedom se calculan como el número total de obervaciones - número de parámetros estimados

4. Residuos estimados

Diferencia entre el valor ajustado y el valor real en cada individuo

uhat = ols$residuals

uhat = resid(ols)

> Extracción directa

residuos <- wdata$wage-yhat

> Extracción manual

$\hat{\epsilon}$

7. Errores de los parámetros estimados (b0 y b1)

Los errores estándar de los parámetros estimados (Std. Error) se obtienen a partir de la raíz cuadrada de la diagonal de la matriz de varianzas-covarianzas estimada

output$coef[,2]

> Extracción directa

varcov=vcov(ols) # Matriz

se = sqrt(diag(varcov)) # Errores estándar

> Cálculo manual

X = as.matrix(cbind(cte=1,wdata[,2]))

varcov2 = sigma^2 *solve(t(X) %*%* X)

se2 = sqrt(diag(varcov2))

> Cálculo manual

Par de números entre los cuales se estima que estará el valor de cada uno de los coeficientes estimados con un determinado nivel de confianza. El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más probabilidad de acierto (mayor nivel de confianza).

```
scatter.smooth(x = wdata$educ, y = wdata$wage, main = "Wage ~ Educ")
```

> Evidencia gráfica a través del gráfico de dispersión (librería graphics)

```
require(ggplot2)
```

```
ggplot(data = wdata, aes(x = educ, y = wage)) + geom_point(colour = "red4") + ggtitle("Gráfico de dispersion") + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```

> Evidencia gráfica a través del gráfico de dispersión (librería ggplot2)

```
cor(wdata$wage, wdata$educ)
```

> Evidencia numérica

```
cor.test(wdata$wage, wdata$educ)
```

> Evidencia contrastada (Pearson's product-moment correlation, más datos que cor, muestra p-valor)

10. Extraer el estadístico F y su p-valor

anova.ols = anova(ols)

Utilizamos la función anova para contrastar la significatividad global del conjunto de parámetros beta siendo la hipótesis nula H0: B0 = B1 = 0

Modificaciones modelo de regresión lineal

ols.nc <- lm(wage ~ 0 + educ, data = wdata)

output.nc <- summary(ols.nc)

> Modelo de regresión lineal simple sobre el origen: si queremos un modelo sin término constante que pase por el origen (0,0) (alternativa 1)

ols.cte <- lm(wage ~ 1, data = wdata)

output.cte <- summary(ols.cte)

> Modelo de regresión lineal simple sobre una constante: si queremos que la recta sea horizontal y la pendiente sea igual a 0. bhat = yhat (media)

plot(wdata$educ, wdata$wage)

abline(ols, lwd = 2, lty = 1)

abline(ols.nc, lwd = 2, lty = 2)

abline(ols.cte, lwd = 2, lty = 3)

legend("topleft", c("Completo", "Pasa por origen", "Solo constante"), lwd = 2, lty = 1:3)

> Representación gráfica de los tres modelos (el tercero es solo la constante)

Proceso

Estimación

> de los modelos de regresión lineal múltiple

Interpretación

> de la salida de los modelos de regresión lineal múltiple

Selección

> del mejor modelo de regresión lineal múltiple que se ajusta a un dataset

Diagnosis

> del modelo de regresión lineal múltiple elegido

2. Coeficientes de regresión estimados

5. Grados de libertad (degrees of freedom)

(número de observaciones - número de parámetros que hemos estimado, b0 y b1) $g.d.l.=n-k$ siendo n el número de obervaciones totales y k el número de parámetros estimados.

output$df[2]

df.residual(ols)

> Extracción directa

dof = nrow(wdata)- output$df[1]

> Cálculo manual

8. Coeficiente de determinación (1-(RSS/TSS))

R2=output$r.squared

R2adj=output$adj.r.squared

> Extracción directa

residuos2 = wdata$wage-mean(wdata$wage)

R2.1=1-(sum(resi2)/sum((residuos2)^2))

> Extracción manual 1

R2.2 = var(yhat)/var(wdata$wage)

> Extracción manual 2

R2.3 <- cor(wdata$wage,wdata$educ)^2

> Extracción manual 3

R2 mejorará, por el simple hecho de incluir más variables en el modelo. El coeficiente de determinación ajustado R2adj corrige esto ya que penaliza la incorporación de nuevas variables independiente. En este caso, no es una preocupación ya que tenemos una única variable independiente pero tendremos que tenerlo en cuenta en los modelos de regresión múltiple

11. Predicciones de nuevos datos

yhat1 = predict(ols, newdata = x)

> Si no incluimos nuevos datos en la variable independiente, el resultado de la función predict coincide con los resultados obtenidos con la función fitted.

Diagnosis modelo de regresión lineal simple

Diagnosis modelo de regresión lineal simple (cont)

plot(ols, which=2)

> gráfica de quantil-quantil normal. Los puntos deberían seguir la diagonal si los residuos están normalmente distribuidos. Si aparecen patrones tipo "S" o de ontra índole, posiblemente necesitemos ajustar otro modelo.

testSW = shapiro.test(residuals(ols))

> Para contrastar la normalidad de los residuos también podemos usar el test de Shapiro-Wilk o alguno de los vistos anteriormente. En este caso, tened en cuenta que lo que estamos contrastando es si los errores siguen una distribución normal no la variable. Podemos ver cómo obtenemos un p-valor < 5% , luego tendríamos que rechazar la hipótesis nula. Es decir, los residuos no seguirían una distribución normal.

plot(ols, which=3)

> La tercera gráfica es como la primera, pero usando una escala diferente, residuos estandarizados, y sirve para comprobar la homocedasticidad, la cual se cumple si los puntos forman una banda en torno a la horizontal.

plot(ols, which=5)

> La última gráfica trata sobre la identificación de puntos influyentes, aberrantes y con efecto palanca. Las observaciones influyentes son aquellas con un impacto desproporcionado en la determinación de los parámetros del modelo. Se identifican usando la distancia de Cook. Son de preocupar los puntos con valores superiores a la línea de Cook. En este caso, no tenemos ninguno. Un punto aberrante es una observación que tiene un valor muy alto del residuo asociado (con un valor muy negativo del residuo y muy a la derecha). En este caso, tampoco tenemos ninguno. Una observación con alto efecto palanca "-leverage" es una observación que no es predicha satisfactoriamente por el modelo de regresión (con un valor muy positivo del residuo y muy a la derecha). Tampoco tenemos ninguno.

output$coef

> Estimate, Std. Error, t value, Pr(>|t|)

bhat = coef(ols)

> Devuelve únicamente el valor estimado (estimate) para b0 y b1

+ Cálculo manual de los coeficientes de regresión $\hat{\beta}=(\hat{\beta}_{0},\hat{\beta}_{1})'$ a partir de la ecuación formal

$$
\begin{align*}
\hat{\beta}_{1}&=\frac{Cov(x,y)}{Var(x)}\\
\hat{\beta}_{0}&=\bar{y}-\hat{\beta}_{1}\bar{x}
\end{align*}
$$

plot(ols, which=1)

> nos ayuda a decidir si las variables están linealmente relacionadas. Si es así, no debería de existir una relación sistemática entre los resíduos (errores) y los valores predichos (ajustados) según los quantiles en z-score ((x-media)/sd)