Universidad de Costa Rica
Facultad de Ingeniería
Escuela de Ingeniería Industrial
Área de conocimiento: Ciencias básicas y de la ingeniería – Probabilidad y estadística

Ejercicio integrador resuelto – Regresión y ANOVA

Elaborado por: Ing. Steven García Goñi.

Tema:	**Modelo de regresión lineal múltiple y ANOVA**
Tipo de actividad:	Formativa
Modalidad:	Estudio individual
Fuente de datos:	Ficticia

Contexto:

Se realizó un estudio por parte de personas estudiantes de la carrera de Bach. en Ciencias del Movimiento Humano (CIMOHU), cuyo objetivo era examinar el efecto que tiene la posición de salida (+ y -) y el tipo de calentamiento (A, B y C) en el recorrido de 100 metros planos. Se trabajó con una población de 17 a 24 años, todos estudiantes masculinos de la sede Rodrigo Facio de la Universidad de Costa Rica.

Como variable de respuesta se mide el tiempo en segundos (s) que cada persona tarda en recorrer los 100 metros. Se asume que todas las mediciones son correctas y que los instrumentos de medición utilizados son adecuados.

En resumen, se puede decir que la fuente de datos es correcta y que los datos se recolectaron de forma adecuada. La hipótesis que se sigue, basada en la teoría de las CIMOHU indica que si debiera existir diferencia entre el tiempo de recorrido de los 100 m planos dado los regresores estudiados.

Solución No. 1:

Entre los pasos previos, relacionados con la teoría, se encuentra la determinación del nivel de confianza que se desea utilizar en los análisis. Este se debe JUSTIFICAR en función del contexto y NO se determina con base en conveniencia (0.1, 0.05, 0.01). Notará que en este ejemplo resuelto no escoge un nivel de confianza, esto se hace de forma intencional para que la persona estudiante reflexione y determine su proprio nivel de confianza.

1. *Especificación del modelo*

Con base en la teoría, se obtiene el siguiente modelo

$$y_{segundos} = \beta_0 + \beta_1 Salida + \beta_2 Calentamiento + \varepsilon_{ij}$$

2. *Estimación de los parámetros*

Una vez especificado el modelo, se debería proceder con la recolección de los datos (la muestra), que en este caso particular ya fue realizado.

No obstante, con objetivos didácticos vale la pena preguntarse ¿Cuántas muestras debo recoger? Depende. ¿De qué? Del mínimo efecto que se desea detectar como significativo, es decir, de la potencia de la prueba. Si necesita repasar este concepto proceda de forma individual.

Los parámetros estimados según la muestra provista son

term	estimate	std.error	statistic	p.value
(Intercept)	10.623	0.797	13.323	0.000
pos_salida+	-0.675	0.797	-0.847	0.400
calentamientoB	2.848	0.977	2.916	0.005
calentamientoC	0.865	0.977	0.886	0.379

¿Nota algo raro? Los términos estimados son la posición de salida (+) y el calentamiento (B y C), pero ¿qué es esto? Bueno, esto sucede cuando en la regresión se incluyen variables categóricas. Note también que no están todas las categorías ¿qué pasa con posición de salida (-) y calentamiento (A)? Pues se llaman categorías de referencia e intencionalmente no forman parte del modelo (porque si lo estuviesen serían colineales). Habiendo aprendido esto, hay que reescribir el modelo del punto anterior:

$$y_{segundos} = \beta_0 + \beta_1 Salida(+) + \beta_2 Calentamiento(B) + \beta_3 Calentamiento(C) + \varepsilon_{ij}$$

Con objetivos didácticos, para introducir a las variables categóricas, se adelanta la interpretación de los coeficientes (recuerde que este es un paso final). Estos se interpretan de esta manera:

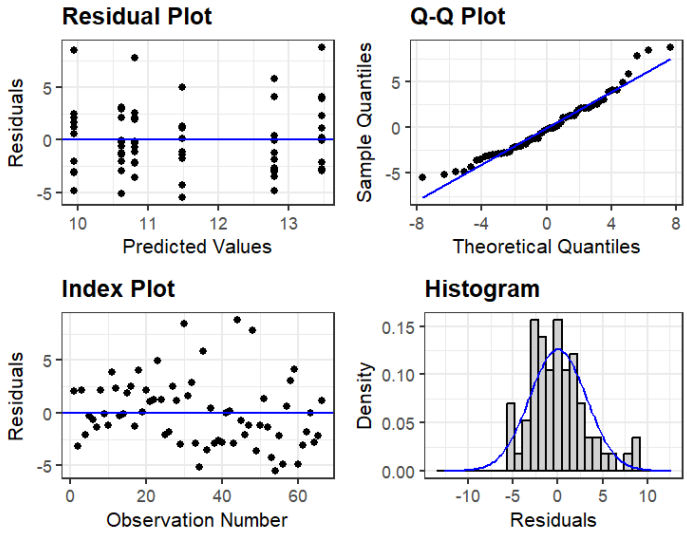
- Cuando la posición de salida es (+) y el resto de los coeficientes permanecen constantes se espera que la persona dure 0.675 segundos menos al recorrer 100 m planos **respecto a la posición de salida (-)**. Note que la interpretación se hace con respecto al término ausente.

- Cuando el calentamiento es B y el resto de los coeficientes permanecen constantes se espera que la persona dure 2.848 segundos más respecto al calentamiento A. Observe la prueba de hipótesis, note que el valor P es muy bajo, por lo que podemos decir que *hay evidencia suficiente para rechazar la hipótesis nula de que la media de B es igual a la media de A*, es decir, son diferentes.
- Interprete usted a modo de práctica el calentamiento C.

3. *Adecuación del modelo*

En esta sección revisamos los supuestos de aplicación. Recuerde que esto se puede realizar de forma gráfica y/o analítica.

Forma gráfica:



De la Figura anterior se observa qué:

- Los residuales son independientes, pues no se observa un patrón según el orden de observación.
- No hay grandes desviaciones de la normalidad de los residuos.
- Se cumple con la homocedasticidad, pues no hay indicio de que las varianzas no sean constantes.

Si lo considera conveniente, realiza pruebas analíticas para comprobar los puntos anteriores.

- Mediante la prueba de autocorrelación de Durbin-Watson se contrasta la hipótesis $H_o: \rho = 0$ (Nota: ρ es el símbolo para correlación).

- Se obtiene como resultado un valor p = 0.874, el cual es lo suficientemente alto para decir que *no existe evidencia suficiente para rechazar la hipótesis nula de que los residuales no están correlacionados.*

- Para la prueba de normalidad se utiliza la prueba de Lilliefors, recuerde que la escogencia de cada una de estas pruebas debe justificarse completamente. De nueva cuenta, esto se deja como práctica para la persona estudiante. Toda prueba de normalidad tiene la siguiente hipótesis nula $H_o: los\ datos\ siguen\ una\ distribución\ normal$. Note la redacción.
  - Con un valor p = 0.3058 *no hay evidencia suficiente para rechazar la hipótesis nula*. Es decir, los residuos siguen una distribución normal.
- Dado que los residuos siguen una distribución normal, se puede aplicar una prueba de homocedasticidad basado en la distribución χ^2, en este caso particular se utilizará la prueba de Breush-Pagan, cuya hipótesis nula es $H_o: los\ residuales\ son\ homocedásticos$
  - Con un valor p = 0.8493 *no existe evidencia suficiente para rechazar la hipótesis nula*.

Ahora bien, dado que todos los supuestos se cumplen, se pueden analizar estadísticos de bondad de ajuste como el R^2.

r.squared	adj.r.squared	sigma
0.135	0.093	3.239

Aquí hay un problema, para los investigadores de la carrera de CIMOHU este valor es demasiado bajo. ¿Qué procede? Detenerse, desde este punto se sabe que el modelo no va a servir. Pero ¿qué hacemos? Cuestionarse la especificación del modelo. Vamos a ello.

Solución No. 2:

A partir de este punto, los pasos repetitivos se van a presentar de forma resumida.

1. *Especificación del modelo:*

Las personas investigadoras se dan cuenta que algo falta en el modelo. Por que sus hipótesis si tienen sentido, pero hay algo que genera mucho "ruido". Uno de ellos señala que la estatura y el peso de todos los sujetos es diferente y que eso tiene un impacto sobre la velocidad a la que corren.

$$y_{segundos} = \beta_0 + \beta_1 Salida(+) + \beta_2 Calentamiento(B) + \beta_3 Calentamiento(C) + Peso + Estatura + \varepsilon_{ij}$$

2. *Estimación de los parámetros*

Dicho esto, se disponen a recoger la información de Peso (kg) y estatura (cm) de las personas que participaron del estudio. Y con base en ello obtienen estas estimaciones:

term	estimate	std.error	statistic	p.value
(Intercept)	30.770	4.591	6.703	0.000
pos_salida+	-0.283	0.472	-0.600	0.551
calentamientoB	2.451	0.574	4.269	0.000
calentamientoC	0.264	0.575	0.459	0.648
peso	0.255	0.023	10.989	0.000
estatura	-0.217	0.032	-6.822	0.000

Observe, rápidamente, como ambas variables si tiene un impacto sobre la variable de respuesta, pues ambas resultan significativas.

3. *Adecuación del modelo*

Se revisan todos los supuestos de nueva cuenta, todos se cumplen, no se presentan para evitar repeticiones innecesarias (en términos didácticos) y centrarse en los cambios ocurridos. Los investigadores observan sus nuevos estadísticos de bondad de ajuste. La mejora es sustancial y satisface los requisitos de la investigación.

r.squared	adj.r.squared	sigma
0.713	0.689	1.897

Ahora bien, se prueba la colinealidad de las variables en el modelo, para ello se hace uso del FIV obteniendo valores menores a 1.65 en todos los casos, por lo que no hay multicolinealidad.

¿Qué inconvenientes tiene el modelo anterior? Pues que la posición de salida no es significativa respecto a la categoría de referencia (-), por lo que ensucia el modelo y podemos considerar quitarla.

¿Puedo hacer eso con calentamiento C? NO, porque uno de los calentamientos (B) si es significativo con respecto a la categoría de referencia (Calentamiento A).

Nota: de ser necesario limpiar varias variables, debería hacer el proceso UNA a la vez y no todas simultáneamente.

- Se re - especifica el modelo

$$y_{segundos} = \beta_0 + \beta_2 Calentamiento(B) + \beta_3 Calentamiento(C) + Peso + Estatura + \varepsilon_{ij}$$

- Se re – estiman los coeficientes

term	estimate	std.error	statistic	p.value
(Intercept)	30.397	4.524	6.718	0.000
calentamientoB	2.453	0.571	4.294	0.000
calentamientoC	0.259	0.572	0.453	0.652
peso	0.257	0.023	11.127	0.000
estatura	-0.216	0.032	-6.836	0.000

- Se re-verifica la adecuación de modelo

Cumple con todos los supuestos y los estadísticos de bondad de ajuste nuevos son:

r.squared	adj.r.squared	sigma
0.711	0.692	1.887

Calcule la diferencia entre R^2's y compárela con la diferencia del modelo "sucio" (el anterior).

Nótese que el modelo ya no se puede limpiar más.

4. Validación del modelo

En esta situación se valida el modelo vaya a funcionar en la forma en la que fue concebido.

Ejercicio para la persona estudiante

1. Interprete cada uno de los coeficientes obtenidos en el modelo final.
2. Realice un ANOVA con los datos del modelo ANTES de limpiarlo. Verifique que se llegan a exactamente las mismas conclusiones, por ejemplo, el calentamiento A y C son iguales y a su vez diferentes al calentamiento B y que la posición de salida no es significativa (+ y – son iguales, pues).
3. Uno de los investigadores señala que se debería incluir el IMC al modelo, pues es una variable con más sentido teórico que la estatura y el peso de forma individual. Usted como persona experta en análisis de regresión le indica que esto es un error, que, si se desea incluir la variable IMC, deben eliminarse el peso y la estatura. Explíquele el por qué.
4. Siga todos los pasos anteriores para la realización del modelo propuesto en la pregunta 3.
5. Con base en el principio de parsimonia y la teoría al respecto del tema tratado por CIMOHU, ¿cuál modelo es preferido, el de la pregunta 4 o el realizado en esta práctica?