Examen Parcial

¡Bienvenid@ a tu examen parcial!

El link de Canvas estará disponible hasta las 9:10pm o una vez que se cumplan 120 minutos del examen (lo que suceda primero). A esa hora se cerrará de forma automática. Al final del examen hay una pregunta donde pueden subir archivos de respaldo a sus respuestas. En caso de hacerlo, asegúrense que el archivo que suban esté claramente asociado a su examen.

Academia de natación *Phelps*

En preparación a los Juegos Olímpicos de Paris 2024, *Michael Phelps* (el atleta más galardonado de la historia con un total de 28 medallas olímpicas) decide abrir su clínica de natación para preparar atletas de alto rendimiento para poder competir en natación durante los próximos juegos olímpicos. Utilizando una base de datos donde la unidad de observación (el renglón de la base de datos) es el resultado de una prueba de natación de 100 metros, la Federación Internacional de Natación (FINA) te solicita que lleves a cabo un análisis estadístico.

1 Planteamiento de especificaciones.

En primera instancia, describiremos algunos de los resultados de un grupo de estimaciones que llevaste a cabo para la FINA. En estas estimaciones utilizaste la variable $Tiempo_i$ como variable dependiente (medida en segundos). $Tiempo_i$ indica el total de segundos que el atleta tardó en completar la prueba i. En algunos casos, dicha variable dependiente se utilizó por si sola y en otros casos se empleó una transformación de la misma (ej. $\log(Tiempo_i)$). A continuación se describe el resultado de dichas estimaciones. Empezamos sin prestar atención a la significancia, sino solo a describir los resultados de las estimaciones. En cada caso de los siguientes incisos deberás indicar la ecuación que se estimó y mostrar a qué coeficiente(s) corresponde(n) el(los) valor(es) que se describe(n) en el inciso correspondiente. Asegúrate de indicar claramente las variables explicativas (X's) que estás utilizando en cada caso. Tu las debes definir. Cualquier variable con un coeficiente que no puedas calcular, déjalo indicado como $\beta_j X_{ji}$ en la especificación.

1. *(6 puntos)* En la primera especificación encontramos que los hombres por nacimiento tienen un tiempo promedio de 52 segundos, mientras que las mujeres por nacimiento un tiempo promedio de 57 segundos.

   R: Creamos la variable $Mujer_i$ que es una dummy igual a uno si i es mujer por nacimiento, e igual a cero si es hombre. Nuestra especificación sería:

   $$Tiempo_i = 52 + 5\ Mujer_i$$

2. *(6 puntos)* En la segunda especificación encontramos que el individuo promedio (de 23 años de edad) tiene un tiempo promedio de 54 segundos y que un aumento de un año de edad esta relacionado a un incremento de 2.3% en el tiempo.

   R: Creamos la variable de $Edad_i$ que corresponde a los años del atleta i durante la prueba. Con esto, nuestra especificación sería:

   $$\log(Tiempo_i) = 3.46 + 0.023\ Edad_i$$

   Para deducir β_0 sabemos que:

   $$log(54) = \beta_0 + 0.023(23)$$

3. *(6 puntos)* En la tercera especificación encontramos una comparación entre nado en diferentes estilos (siendo los 4 estilos: crol, dorso, pecho y mariposa). Ahi podemos observar que el estilo más rápido es crol seguido de mariposa que es 1.5 segundos más lento, en promedio. Dorso es 3 segundos más lento que crol en promedio. Y pecho es el estilo más lento, siendo 5 segundos más lento que dorso. Además, en esta especificación se encontró que los nadadores que forman parte de la academia de Phelps obtienen los mismos resultados promedio antes descritos en todas las disciplinas, excepto en mariposa (la especialidad de Phelps), donde son 1.5 segundos más rápidos.

   R: Creamos dummies para los estilos de nado: $\{Mariposa_i, Pecho_i, Dorso_i\}$ que son mutuamente excluyentes y dejamos crol como el estilo de referencia. Tenemos tambien $Phelps_i$ como dummy qie indica si el atleta i es parte de la academia de Phelps. Con ello nuestra especificación sería:

   $$\begin{aligned} Tiempo_i =& \beta_0 + 0Phelps_i + 1.5Mariposa_i + 3Dorso_i + 8Pecho_i \\ &- 1.5Mariposa_iPhelps_i + 0Dorso_iPhelps_i + 0Pecho_iPhelps_i \end{aligned}$$

4. *(6 puntos)* En la cuarta especificación vemos que un aumento de 1% en la altura del atleta está relacionado con una reducción de 0.075 desviaciones estándar en el tiempo.

   R: Estandarizamos la variable dependiente, restando la media y dividiendo entre su desviación estándar. Con ello creamos la variable $\widetilde{Tiempo_i}$. Utilizando esta variable como dependiente y la variable $Altura_i$ que indica en centímetros la altura del individuo, la siguiente sería nuestra especificación:

   $$\widetilde{Tiempo_i} = \beta_0 - 7.5\log(Altura_i)$$

   De hecho, el valor de β_0 tambien se puede deducir, aunque no les daría penalización por no deducirlo. Para ello, vale la pena notar que al promediar $\widetilde{Tiempo}$ obtendrían un cero por el hecho de que está estandarizada. Si promedian del lado derecho obtendrían:

   $$0 = \beta_0 - 7.5\overline{\log(Altura)} \tag{1}$$

Por lo tanto $\beta_0 = +7.5\overline{\log(Altura)}$

5. *(6 puntos)* En la quinta especificación, utilizamos la variable $Tiempo_i$ para crear una nueva variable dependiente: $Ganador_i$, una dummy que indica si ese renglón corresponde al registro de una prueba en la cual el competidor resultó ganador. En dicha especificación podemos ver que un aumento de una competencia previa incrementa la probabilidad de ser ganador en 1.4 puntos porcentuales para las mujeres y en 1.7 puntos porcentuales para los hombres. Asimismo, una mujer sin competencias previas tiene una probabilidad de 9% de resultar ganadora.

   R: La variable $Competencia_i$ indica el número de competencias previas que el atleta i ha tenido antes de la prueba. En este caso, a diferencia que en el inciso (a) dejaremos que Mujer sea el grupo de referencia debido a que, con la información provista, podemos determinar el valor de un coeficiente más. Con ello, nuestra especificación sería:

$$Ganador_i = 0.09 + \beta_1 Hombre_i + 0.014 Competencia_i + 0.003 Competencia_i Hombre_i$$

2 Preguntas de desarrollo.

Contesta las siguientes preguntas con la información que se te entrega. Todas ellas corresponden también al análisis que llevaste a cabo para FINA con la base de datos antes mencionada.

6. *(15 puntos)* Una compañera de tu oficina estimó todas las especificaciones de la sección 1 utilizando $Tiempo_i$ medido en minutos en vez de segundos. Indica cuáles de las especificaciones cambiarían y TODOS los coeficientes resultantes serían diferentes que al utilizar $Tiempo_i$ medido en segundos. OJO: si al menos uno de los coeficientes sería el mismo, no deberás marcar dicha especificación.

   R: Para todas las especificaciones, lo que notamos es que:

$$Tiempo_i = Tiempo\_min_i \cdot 60$$

   Por lo tanto veamos para cada especificación:

$$\begin{aligned} Tiempo\_min_i \cdot 60 &= 52 + 5\ Mujer_i \\ Tiempo\_min_i &= \frac{52}{60} + \frac{5}{60}\ Mujer_i \end{aligned} \tag{2}$$

   En la primera especificación ambos coeficientes cambian porque ambos están en las mismas unidades que Y. Por lo tanto, SI lo marcamos.

$$\begin{aligned} \log\left(Tiempo\_min_i \cdot 60\right) &= 3.46 + 0.023 Edad_i \\ \log\left(Tiempo\_min_i\right) &= \left[3.46 - \log\left(60\right)\right] + 0.023 Edad_i \end{aligned} \tag{3}$$

En la segunda especificacioón el coeficiente de Edad no cambia, pero la constante sí. Por lo tanto, NO lo marcamos.

La tercera especificación sigue la misma lógica que la primera, por lo tanto, SI la marcamos.

Para la cuarta especificación veamos que sucede si estandarizamos utilizando la variable en minutos:

$$\begin{aligned} \widetilde{Tiempo\_min}_i &= \frac{Tiempo\_min_i - \overline{Tiempo\_min}}{\sqrt{Var(Tiempo\_min)}} \\ &= \frac{Tiempo_i \cdot 60 - \overline{Tiempo} \cdot 60}{\sqrt{Var(60 \cdot Tiempo)}} \\ &= \frac{60 \cdot (Tiempo_i - \overline{Tiempo})}{\sqrt{60^2 Var(Tiempo)}} \\ &= \frac{60 \cdot (Tiempo_i - \overline{Tiempo})}{60 \cdot \sqrt{Var(Tiempo)}} \\ &= \frac{Tiempo_i - \overline{Tiempo}}{\sqrt{Var(Tiempo)}} \\ &= \widetilde{Tiempo}_i \end{aligned} \tag{4}$$

Como la variable dependiente no cambia en la cuarta especifciación, los coeficientes serían los mismos, ninguno cambio. Por lo tanto, NO la marcamos.

Finalmente, la quinta especificación, la dummy de ganador no va a cambiar. Ya que no importa si el tiempo se mida en minutos o segundos, el ganadro no cambiaría. Por lo tanto, NO la marcmos tampoco.

7. *(15 puntos)* Te interesa evaluar si los nadadores de la academia de natación *Phelps* obtienen mejores resultados (estadísticamente significativos) para el estilo de mariposa. Para ello, siguiendo el resultado de la tercera especificación, te quedas sólo con las observaciones de las pruebas de mariposa. Usando errores homocedásticos obtienes una diferencia estadísticamente significativa con un valor-p igual a 0.04. Sin embargo, los errores heterocedásticos son 15% mayores que los homocedásticos. ¿Qué valor-p obtendrías si utilizaras errores heterocedásticos? (redondea el valor-p a tres decimales)

R: Para empezar, si tenemos un valor-p de 0.04, eso quiere decir que el estadístico-t asociado sería (en una prueba bilateral, que es lo usual para determinar la significancia estadística):

$$t = \Phi^{-1}\left(1 - \frac{p}{2}\right) = 2.0537$$

Dicho estadístico-t resultado de la diferencia en mariposa entre la academia *Phelps* y los demás entre el error estándar. En la pregunta 1(c) sabíamos que la diferencia en segundos en mariposa entre la academia *Phelps* y los demás era de 1.5 segundos. Por lo tanto, podemos deducir el error estándar:

$$t = 2.0537 = \frac{1.5}{SE_{homo}}$$

Con ello, sabemos que los errores homocedásticos para la diferencia es 0.7304. Como los errores heterocedásticos son 15% mayores sabríamos que son 0.8399. Por ello, utilizando estos errores y la diferencia de 1.5, el nuevo estadístico-t sería:

$$t = \frac{1.5}{0.8399} = 1.7859$$

De ahí deducimos que el valor-p con errores heterocedásticos sería:

$$pval = 2 \cdot (1 - \Phi(1.7859)) = 0.074$$

8. *(10 puntos)* Siguiendo con el test anterior, alguien te señala que un problema con tus estimaciones es que diferentes renglones le corresponden a un mismo competidor. Es decir, más de un renglón le corresponde a una misma persona en diferentes competiciones debido a que tus datos se recopilan entre 2021 y 2023. Tu estabas asumiendo que tu muestra es i.i.d. Explica: ¿por qué el supuesto de i.i.d. puede ser incorrecto en este caso?

   R: Porque en este caso, al dos renglines o más corresponderle al mismo individuo se rompa la independencia de los diferentes componentes de la muestra. El hecho de que en tu muestra se elja a un competidor muy exitoso o de alto rendimiento, eso hace muy probable que sea observado en diversas ocasiones. Esos diferentes resultados observados no son independientes entre si al corresponderle al mismo individuo. Los valores observados en esos renglones estarán fuertemente correlacionados entre sí.

9. *(15 puntos)* Explica por qué el resultado de la pregunta (3) donde evaluamos si la diferencia de los nadadores de la academia *Phelps* es significativamente distinta no puede tomarse como causal. Es decir, ¿por qué no podemos atribuir la diferencia de tiempos a lo que aprenden los nadadores en la academia *Phelps*?

R: Simplemente podríamos tener un problema de autoselección o de sesgo por variables omitidas. Si los nadadores con mayores aspiraciones y talento (particulamrmente para el estilo mariposa) deciden inscribirse en la academia *Phelps* inspirados por los resultados de Michael, eso significaría que los mejores resultados observados para la academia *Phelps* pudieran simplemente deberse a que mejores nadadores decidieron inscribirse. Es decir, aun sin estar esos mismos nadadores en la academia *Phelps* tal vez de cualquier manera les hubiera ido mejor, particulamrente en mariposa. La variable omitida (no observable) pudiera ser *aspiracion profesional*. Es muy difiícil de medir u observar. Pero podría estar positivamente correlacionada con ser nadador de la academia *Phelps* y también relacionada a obtener un menor tiempo. Por lo tanto, el (-1.5) que observamos para el coeficiente de $Mariposa_i Phelps_i$ podría estar sesgado de forma negativa. Es decir, de controlar por *aspiracion profesional* obtendríamos un valor mayor, con lo cual la diferencia entre los nadadores de la acamia *Phelps* con los demás se diluiría.

10. *(15 puntos)* Imagina que una vez recabada la base de datos se descubre que todos los tiempos registrados tienen un error ya que el cronómetro utilizado sumaba 0.2 segundos al tiempo verdadero de los competidores.

    a. Propón un estimador insesgado para la media de $Tiempo_i$. Demuestra que tu estimador es insesgado.

    R: Para responder esta pregunta podemos empezar por ver cómo afecta el cálculo de la media. Sea $\tilde{X}_i = X_i + k$, donde en nuestro caso $k = 0.2$. Con esto vemos para el caso de la media:

$$\begin{aligned}\overline{\tilde{X}} &= \frac{1}{n}\sum_i \tilde{X}_i = \frac{1}{n}\sum_i \Big(X_i + k\Big) \\ &= \frac{1}{n}\sum_i X_i + \frac{1}{n}\sum_i k \\ &= \overline{X} + k\end{aligned}$$

    Por lo tanto, para proponer un estimador insesgado simplemente tendríamos que proponer $\overline{\tilde{X}} - k$

    b. Si no te hubieras percatado del error, ¿afectaría esto al cálculo del estimador de la varianza: S^2_X? Demuestra si afecta o no.

    R: No la afecta. Para ver esto:

$$
\begin{aligned}
S^2_{\tilde{X}} &= \frac{1}{n-1}\sum_i \left(\tilde{X}_i - \overline{\tilde{X}}\right)^2 \\
&= \frac{1}{n-1}\sum_i \left(X_i + k - \overline{X} - k\right)^2 \\
&= \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2 \\
&= S^2_X
\end{aligned}
$$