Pauta Examen

IN3242 - Estadística, Secciones 1 y 2 Departamento de Ingeniería Industrial, Universidad de Chile
Primavera 2021
Prof. Raimundo Undurraga

Pregunta 1 (60 puntos). Una ONG interesada en mejorar la calidad de vivienda de hogares en situación de vulnerabilidad social ha implementado un programa de mejora a la vivienda para hogares que habitan en campamentos, el cual consiste en la entrega de una vivenda mejorada. La ONG identifica un total de 880 familias elegibles para el programa. En general, las familias elegibles comparten características similares: baja calidad en materialidad de pisos, paredes, y techos; falta de acceso a servicios básicos como agua potable, electricidad y alcantarillado. No obstante, la ONG no tiene recursos para atender las necesidades de vivienda de todos los hogares elegibles identificados, y solo cuenta con 450 viviendas mejoradas para entregar. La ONG acuerda con las familias un plan de entrega de las viviendas en dos años consecutivos. En el año 1 se entregarán las casas a un primer grupo de 450 familias, y en el año 2 se entregarán 430 casas a las familias que no habían recibido la casa en el año 1. De esa manera, todas las familias elegibles podrán tener acceso a las viviendas mejoradas.

Un investigador se percata de que acá hay una oportunidad para evaluar el impacto de las viviendas mejoradas en la calidad de vivienda de los hogares beneficiarios, y sugiere a la ONG que se aleatorice el orden de entrega de las casas a las familias elegibles. La ONG accede a la sugerencia del investigador, y procede a realizar una asignación aleatoria de cuando recibe el programa cada familia: si en el año 1 o en el año 2. Luego, al final del primer año (i.e., antes de que se entreguen las viviendas mejoradas al grupo del año 2), el investigador podrá comparar el nivel de calidad de la vivienda de los hogares que ya recibieron las viviendas mejoradas (asignados al año 1) con el nivel de calidad de la vivienda de los hogares que aún no han recibido el programa.

De hecho, pasado un año desde que se entregaron las casas al grupo del año 1, el investogador accede a una base de datos (descargable aquí) que incluye las siguientes variables:

- *ronda*: igual a 0 si es que la información es ANTES de que se entreguen las viviendas; y 1 si es que la información es DESPUES de que se entreguen las viviendas.
- *grupo_tratamiento*: igual a 1 si el hogar fue asignado al grupo del año 1 (grupo de tratamiento), y 0 si es que fue asignado al grupo del año 2 (grupo de control).
- *prop_piso_buena_calidad*: variable continua que indica la proporción de los pisos de la casa que están en buen estado de materialidad
- $satisfaccion\_calidad\_vivienda$: igual a 1 si es el jefe de hogar esta satisfecho con la calidad de su vivienda; y 0 si es que no está satisfecho con la calidad de su vivienda.
- *ingreso_per_capita*: variable continua que indica el nivel de ingreso per capita del hogar (en dolares americanos).

El investigador le contrata como ayudante de investigación y le solicita resolver las siguientes preguntas:

(a) (10 puntos) En términos de la evaluación de impacto del programa: ¿Qué garantiza la aleatorización de los hogares al grupo del año 1 y año 2? Explique teóricamente y ejemplifique.

Lo que garantizará que los hogares fueron escogidos de manera aleatoria es que su asignación a grupo tratamiento o control sea independiente de los estados potenciales de la variable de resultado, esto es, que $E[Y_{0i}|D_i = 0] = E[Y_{0i}|D_i = 1]$ por lo que no habrá un sesgo de selección. En otras palabras, en el contexto de una regresión lineal que mide el efecto de haber sido asignado al tratamiento (año 1) sobre la calidad de vivienda de los hogares (u otra variable de interés), la aleatorización garantiza que el término de error sea ortogonal a la variablé de asignación del tratamiento, es decir, que $E[e_i|D_i = 0] = E[e_i|D_i = 1] = E[e_i]$, lo cual a su vez garantiza que el efecto del tratamiento puede ser identificado causalmente.

(b) (10 puntos) Demuestre empíricamente que la aleatorización si funcionó. Para ello, explicite la hipótesis a testear, y use las variables *prop_piso_buena_calidad*, *satisfaccion_calidad_vivienda*, e *ingreso_per_capita* para implementar empíricamente el test. Comente brevemente los resultados (Nota: Fíjese bien en la ronda que debe usar para realizar el test.)

Acá debemos comparar los hogares en la ronda cero (esto es antes del tratamiento de ambos) para comparar si las diferencias en los grupos son o no son significativas, esto podemos hacerlo a través de una regresión de las 3 variables que se nos mencionan en función del tratamiento (donde en caso de no ser significativa la aleatorización funcionó y en caso contrario no) o bien un test de medias entre las variables para el grupo control y tratamiento, a continuación se muestran las dos formas para cada variable:

```
> a<-lm(ingreso_per_capita ~ grupo_tratamiento,data=df_0)
> summary(a)

Call:
lm(formula = ingreso_per_capita ~ grupo_tratamiento, data = df_0)

Residuals:
   Min     1Q Median     3Q    Max
 -76.4  -53.9  -29.0    6.2 3907.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         75.969      9.935   7.647 6.92e-14 ***
grupo_tratamiento    1.117     14.070   0.079    0.937
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 185.3 on 692 degrees of freedom
Multiple R-squared:  9.114e-06, Adjusted R-squared:  -0.001436
F-statistic: 0.006307 on 1 and 692 DF,  p-value: 0.9367
```

```
> b<-lm(prop_piso_buena_calidad ~ grupo_tratamiento,data=df_0)
> summary(b)

Call:
lm(formula = prop_piso_buena_calidad ~ grupo_tratamiento, data = df_0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6487 -0.3065  0.1013  0.3513  0.3602

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.639819   0.020359  31.427   <2e-16 ***
grupo_tratamiento 0.008856   0.028834   0.307    0.759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3798 on 692 degrees of freedom
Multiple R-squared:  0.0001363, Adjusted R-squared:  -0.001309
F-statistic: 0.09433 on 1 and 692 DF,  p-value: 0.7588
```

```
> c<-lm(satisfaccion_calidad_vivienda ~ grupo_tratamiento,data=df_0)
> summary(c)

Call:
lm(formula = satisfaccion_calidad_vivienda ~ grupo_tratamiento,
    data = df_0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2341 -0.2341 -0.1954 -0.1954  0.8046

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.19540    0.02202   8.874   <2e-16 ***
grupo_tratamiento  0.03870    0.03118   1.241    0.215
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4107 on 692 degrees of freedom
Multiple R-squared:  0.002221,  Adjusted R-squared:  0.000779
F-statistic:  1.54 on 1 and 692 DF,  p-value: 0.215
```

```
> t.test(df_0$ingreso_per_capita[df_0$grupo_tratamiento==0],df_0$ingreso_per_capita[df_0$grupo_tratamiento==1])

        Welch Two Sample t-test

data:  df_0$ingreso_per_capita[df_0$grupo_tratamiento == 0] and df_0$ingreso_per_capita[df_0$grupo_tratamiento == 1]
t = -0.079265, df = 480.84, p-value = 0.9369
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.81720  26.58237
sample estimates:
mean of x mean of y
 75.96907  77.08649
```

```
> t.test(df_0$prop_piso_buena_calidad[df_0$grupo_tratamiento==0],df_0$prop_piso_buena_calidad[df_0$grupo_tratamiento==1])

        Welch Two Sample t-test

data:  df_0$prop_piso_buena_calidad[df_0$grupo_tratamiento == 0] and df_0$prop_piso_buena_calidad[df_0$grupo_tratamiento == 1]
t = -0.30716, df = 691.9, p-value = 0.7588
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06546508  0.04775315
sample estimates:
mean of x mean of y
0.6398194 0.6486753
```

```
> t.test(df_0$satisfaccion_calidad_vivienda[df_0$grupo_tratamiento==0],df_0$satisfaccion_calidad_vivienda[df_0$grupo_tratamiento==1])

        Welch Two Sample t-test

data:  df_0$satisfaccion_calidad_vivienda[df_0$grupo_tratamiento == 0] and df_0$satisfaccion_calidad_vivienda[df_0$grupo_tratamiento == 1]
t = -1.2409, df = 688.49, p-value = 0.2151
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09993994  0.02253644
sample estimates:
mean of x mean of y
0.1954023 0.2341040
```

Claramente en las 3 regresiones, la variable de tratamiento no es significativa, y en los 3 test de medias no se puede rechazar la hipótesis nula por lo que la diferencia de medias sería de cero con un nivel de significancia de un 95 %, así, la aleatorización funcionó y queda de manera empírica demostrada.

(c) (10 puntos) ¿Cuál es el impacto promedio del tratamiento en la calidad de los pisos? Describa el modelo de regresión lineal que debe correr para estimar impacto. Luego estime el impacto y describa los resultados.

Dado que tenemos una asignación aleatoria entre los hogares, podremos estimar el efecto promedio como el parámetro asociado al tratamiento regresionando la variable calidad de los pisos contra la dummy de tratamiento, siempre en ronda 1 (después de implementado el tratamiento, dado lo mencionado en el foro se tomó como correcto también hacerlo con ronda 2 siempre que se justificara bien como sería el efecto, que sería estar expuesto 2 años al programa versus estarlo solo uno). En efecto, la regresión que corremos es la siguiente:

$$prop\_piso\_buena\_calidad = \beta_0 + \beta_1 grupo\_trat + \varepsilon$$

```
> reg<- lm(prop_piso_buena_calidad ~ grupo_tratamiento, data=df_1)
> summary(reg)

Call:
lm(formula = prop_piso_buena_calidad ~ grupo_tratamiento, data = df_1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8164 -0.1497  0.1836  0.1836  0.2877

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.71228    0.01681  42.363  < 2e-16 ***
grupo_tratamiento  0.10413    0.02339   4.452 9.93e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3054 on 681 degrees of freedom
Multiple R-squared:  0.02829,   Adjusted R-squared:  0.02686
F-statistic: 19.82 on 1 and 681 DF,  p-value: 9.926e-06
```

Así, tendríamos que el efecto promedio del programa sería que aumenta la proporción del piso de buena calidad en un 0,1041 o un 10,41 %. Así, un hogar que fue tratado en el año 1 tiene un 10,41 % más de piso de buena calidad que un hogar que aún no se le aplica el programa de mejora. (dependiendo de como se filtraran los datos o excluyeran podía variar esto pero no se descontó por ello).

(d) (10 puntos) Un colega le sugiere que los resultados de su regresión están sesgados pues el ingreso per cápita es una variable omitida del modelo que correlaciona simultáneamente con la calidad de los pisos y con la probabilidad de ser asignado al grupo de tratamiento. ¿Está en lo correcto el colega? Demuestre teórica y empíricamente.

El modelo de regresión planteado en la parte c) tiene como variable dependiente la calidad de los pisos y como variable independiente el atributo que indica si la familia está en el grupo de tratamiento. Es decir, un modelo con la siguiente forma:

$$prop\_piso\_buena\_calidad = \beta_0 + \beta_1 grupo\_trat + \varepsilon$$

Los ingresos no serían una variable omitida, ya que tal como se demostró en la parte b), los ingresos están bien balanceados entre el grupo de tratamiento y el grupo de control, es decir, no están correlacionados con la dummy de tratamiento. Omitir dicha variable de la regresión no sesga el parámetro de interés.

(e) (10 puntos) Cuál es el impacto promedio del tratamiento en la satisfacción con la calidad de la vivienda? Describa el modelo de regresión lineal que debe correr para estimar el impacto. Luego estime el impacto y describa los resultados.

```
Call:
lm(formula = satisfaccion_calidad_vivienda ~ grupo_tratamiento,
    data = dfP1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4807 -0.4807 -0.3348  0.5193  0.6652

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.33481    0.01869  17.918  < 2e-16 ***
grupo_tratamiento1  0.14588    0.02623   5.562 3.19e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4865 on 1375 degrees of freedom
Multiple R-squared:  0.02201,   Adjusted R-squared:  0.0213
F-statistic: 30.94 on 1 and 1375 DF,  p-value: 3.194e-08
```

El impacto de ser asignado al beneficio es positivo y significativo al 99.99 %. Exactamente, recibir la mejora en la vivienda aumenta en un 14.5 % la probabilidad de estár satisfecho con la calidad del inmueble, relativo a una familia que no recibe el beneficio.

(f) (10 puntos) Un colega le advierte que la satisfacción con la calidad de la vivienda es una variable dicotómica, y le sugiere usar un modelo logit para estimar impacto. Describa los supuestos del modelo logit que estimaría, y luego estimelo empíricamente y describa los resultados obtenidos.

Los supuestos de Logit son que la variable dependiente debe ser una variable categórica. Y en el caso de que sea dicotómica o binaria, la probabilidad estimada de la ocurrencia de un evento particular, se escribe como:

$P(Y = 1|X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$

```
Call:
glm(formula = satisfaccion_calidad_vivienda ~ grupo_tratamiento,
    family = binomial(link = "logit"), data = dfP1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.145  -1.145  -0.903   1.210   1.479

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.68652    0.08138  -8.436  < 2e-16 ***
grupo_tratamiento1  0.60923    0.11115   5.481 4.22e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1862.9  on 1376  degrees of freedom
Residual deviance: 1832.5  on 1375  degrees of freedom
AIC: 1836.5

Number of Fisher Scoring iterations: 4
```

El impacto de ser asignado al beneficio es positivo y significativo al 99.99 %. Es decir, recibir la mejora en la vivienda aumenta la probabilidad de percibir positivamente la calidad de la misma.

Para ver el aumento exacto de esto en el ratio de probabilidad, se calcula la exponencial del coeficiente estimado que se relaciona al grupo de tratamiento. Y se tiene el siguiente resultado:

$e^{0,60923} = 1,839$

Lo que significa que recibir el beneficio implica un aumento del 83.9 % en el ratio de probabilidad asociado a percibir de forma positiva la calidad de la vivienda.

Pregunta 2 (30 puntos).

Vaya al siguiente sitio web y descargue la base de datos Casen 2017.dta. La base contiene alrededor de 216.000 observaciones correspondientes a individuos que habitan el territorio nacional. La variable *pco1* indica la relación del individuo con el jefe de hogar. La variable *sexo* indica si el individuo es hombre o mujer. La variable *edad* indica la edad del individuo. La variable *esc* indica los años de escolaridad del individuo. Finalmente, la variable *yoprcor* indica el salario del individuos (si el individuo está desempleado, los salarios no son observables y corresponden a "missing values.º puntos dentro de la base de datos). Suponga que usted está interesado en modelar el comportamiento de los salarios de los jefes de hogar, y para ello crea el siguiente modelo de regresión:

$$Salario_i = \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Escolaridad_i + e_i \tag{1}$$

(a) (10 puntos) Estime el modelo de regresión usando Stata o R. Según su modelo, ser hombre o mujer, influye en los salarios? Interprete orden de magnitud del parámetro y significancia estadística.

En este caso se define en la dummy respecto al sexo de una persona como cero si es mujer y uno si es hombre por lo que se aprecia en este modelo que el hecho de ser hombre genera una diferencia de $220.101 respecto al salario que perciben las mujeres aunque el R^2 ajustado es de tan solo un 0.173

```
Residuals:
     Min       1Q   Median       3Q      Max
-1423883  -288968  -101618   122560 38897890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -814071      20192   -40.3   <2e-16 ***
esc            82943        887    93.5   <2e-16 ***
sexo          220101       7328    30.0   <2e-16 ***
edad            5721        282    20.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 733000 on 45477 degrees of freedom
Multiple R-squared:  0.173,     Adjusted R-squared:  0.173
F-statistic: 3.17e+03 on 3 and 45477 DF,  p-value: <2e-16
```

En el output se observa que todas las variables son significativas tanto por p-valor como por estadístico crítico e intervalo de confianza. En el modelo se tiene un intercepto de -$814.071 mientras que un año más de escolaridad entrega $82.943 y un aumento de la edad entrega $5.721.

(b) (10 puntos) Cuánto de la variabilidad de los salarios se explican por la variabilidad de las variables del modelo? Construya el indicador utilizando SST, SSR y SSE.

Basta con mencionar el R^2 de la regresión (que el modelo explica un 17.3 % de la variación de los salarios) y como éste está en función de SST, SSR y SSE:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

(c) (10 puntos) Un colega le sugiere que la edad y la escolaridad son variables que no debiesen estar dentro del modelo. Usted le replica indicando que para ello tendríamos que testear la validez de un modelo restringido que excluya ambas variables. Estime el modelo restringido y construya un test F que le permite testear la veracidad de la sugerencia de su colega. Qué concluye?

Dado el output del modelo restringido, definimos el test de hipótesis como:

$$H_0 : \beta_{esc} = \beta_{edad} = 0$$

$$H_1 : \beta_{esc} \text{ o } \beta_{edad} \neq 0$$

Así tendremos un test F el cual comparará la varianza de los errores entre el modelo restringido y el modelo no restringido de la siguiente forma:

$$F = \frac{(SSE_r - SSE_{nr}/g)}{SSE_{nr}/(n-K-1)}$$

donde g es el número de restricciones (en este caso sería 2), SSE_r sería la suma de los errores al cuadrado para el modelo restringido y SSE_{nr} para el no restringido, n sería el número de observaciones de la muestra y K el número de parámetros a estimar en la regresión no restringida por lo que tendremos:

$$F = \frac{(SSE_r - SSE_{nr}/2)}{SSE_{nr}/(45481 - 3 - 1)}$$

$$F = \frac{(2{,}9208 * 10^{16} - 2{,}4448 * 10^{16})/2}{2{,}4448 * 10^{16}/45477}$$

$$F = \frac{(2{,}9208 * 10^{16} - 2{,}4448 * 10^{16})/2}{2{,}4448 * 10^{16}/45477} = 4427{,}16214$$

Luego el valor de F crítico ($F_{2,45477}$) para un $\alpha = 0{,}05$ es igual a 3, por lo que dado que F calculado es mayor al F crítico, rechazamos la hipótesis nula y al menos una de las variables es significativa en el modelo.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1351078  -296204  -104768   125694 38975667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -677762      19869   -34.1   <2e-16 ***
esc            81910        895    91.5   <2e-16 ***
edad            6192        285    21.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 740000 on 45478 degrees of freedom
Multiple R-squared:  0.157,     Adjusted R-squared:  0.157
F-statistic: 4.22e+03 on 2 and 45478 DF,  p-value: <2e-16
```

Pregunta 3 (30 puntos).

(a) (15 puntos) Explique en qué consiste MLE. Discuta al menos 2 aspectos lo diferencian de OLS.

MLE selecciona un set de valores para el conjunto de parámetros θ que caracterizan a un “modelo” dado tal que dicho set de valores maximiza la probabilidad de observar la muestra de estudio. El “modelo” asume que la muestra de estudio es generada a partir de una distribución poblacional determinada. Asumida dicha distribución y el conjunto de parámetros que la caracterizan, la probabilidad de observar la muestra de estudio viene dada por la función de verosimilitud de la muestra, es decir, la función de probabilidad conjunta de cada una de las observaciones contenidas en la muestra. Formalmente, $L(\theta; X_1, ...X_N) = \prod_{i=1}^{N} f_X(X_i; \theta)$, donde $f_X(X_i; \theta)$ representa la función de densidad de X dada la distribución poblacional asumida. En efecto, MLE lo que hace es encontrar los valores del vector θ que maximizan $L(\theta; X_1, ..., X_N)$. [5 puntos]

Dos diferencias fundamentales entre MLE y OLS son, entre otras, las siguientes:

(i) MLE asume una distribución poblacional desde la cual proviene la muestra de estudio. OLS es agnóstico al respecto y no impone restricción alguna a dicha forma funcional. [5 puntos]
(ii) MLE permite estimar modelos no lineales en los parámetros de interés. OLS, en cambio, es un método que tiene entre sus supuestos que el modelo estimado es lineal en los parámetros de interés. [5 puntos]

(b) (15 puntos) Una de las ventajas del estimador de Máxima Verosimilitud (MLE) es que al asumir que los errores del modelo se distribuyen normal, entonces no es necesario asumir que la distribución poblacional desde la cual proviene la muestra de estudio tiene una forma funcional definida. Comente.

Falso. Cuando asumimos que los errores del modelo se distribuyen normal, lo que estamos asumiendo es que la distribución poblacional desde la cual proviene la muestra de estudio es normal.

Pregunta 1 Indique Verdadero (V) o Falso (F). Si es Falso (F), justifique.

(a) [10 puntos] Uno de los supuestos fundamentales para estimar un modelo de regresión lineal vía OLS es que debe ser lineal en las variables. Por ejemplo, el modelo $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 \frac{X_2}{X_1} + u$ no puede ser estimado por OLS, porque no es lineal en las variables.

(b) [10 puntos] Suponga el siguiente modelo lineal: $x_i = \beta Y_i + u_i$, donde (x, Y) son variables observables, mientras que u_i es el error (no observado). Si $cov[u, Y] = 0$, entonces $E[u|Y] = 0$.

(c) [10 puntos] El teorema de Gauss Markov indica que bajo los supuestos de linealidad, aleatoridad de la muestra, no multicolinealidad, independencia entre el error y las variables independientes, y homocedasticidad, entonces el estimador OLS es, dentro de todos los estimadores insesgados, aquel con mínima varianza (lea bien antes de contestar).

(d) [10 puntos] Suponga un modelo de regresión lineal $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + u$ que se estima vı'a OLS. Mientras menor sea la correlación entre X_k y el resto de las variables independientes del modelo, entonces mayor será la varianza muestral del estimador de β_k.

Pregunta 2 Considere el siguiente modelo de regresión lineal expresado en forma matricial $y_i = \mathbf{Z}_i^T \beta + \epsilon_i$, donde $\mathbf{Z}_i$ es un vector $K \times 1$ de variables aleatorias y $\beta = (\beta_1..\beta_k)^T$ un vector $K \times 1$ de parámetros. Supongamos que $\epsilon_i \sim N(0, \sigma^2)$, tal que $Y_i|Z_i \sim N(\mathbf{Z}_i^T \beta, \sigma^2)$.

(a) [15 puntos] Derive la función de verosimilitud condicional.

(b) [15 puntos] Derive la función de log-verosimilitud condicional y obtenga una expresión del estimador de máxima verosimilitud para el vector $\beta = (\beta_1..\beta_k)^T$.

Pregunta 3. [20 puntos] Explique en detalle qué consiste MLE. Discuta al menos 2 aspectos que lo diferencian de OLS.

Pregunta 4 Un grupo de investigadores quiere medir el efecto de un programa que entrega viviendas a hogares de campamento en la probabilidad de que estos mejoren su acceso a agua potable. Los ejecutores del programa deciden asignar el programa aleatoriamente. Algunos

hogares reciben el programa (grupo de tratamiento) y otros no (grupo de control). Los investigadores utilizan el siguiente modelo de regresión $AguaPot_i = \alpha_0 + \alpha_1 Tratamiento_i + e_i$, donde $AguaPot_i$ toma el valor 1 si el hogar tiene agua potable dentro del hogar y 0 si no; la variable $Tratamiento_i$ toma el valor 1 si el hogar fue asignado al grupo de tratamiento y 0 si no.

(a) [10 puntos] Primero se estima el modelo vía OLS (modelo de probabilidad lineal, ver tabla abajo). Interprete los coeficientes α_0 y α_1.

(b) [10 puntos] Suponga que la distribución de probabilidad del acceso a agua es una Bernoulli, tal que $AguaPot_i = 1$ con probabilidad $F(\alpha_0 + \alpha_1 Tratamiento_i)$ y $AguaPot_i = 0$ con probabilidad $1 - F(\alpha_0 + \alpha_1 Tratamiento_i)$. Escriba la función de Máxima Verosimilitud condicional.

Pregunta 1 (40 puntos). Dado el contexto nacional, usted como estudiante está interesado en aportar en el debate de las pensiones. Para tener una opinión informada, usted se consigue los datos de la Superintendencia de Pensiones para analizar qué variables de los pensionados tiene un mayor efecto en el monto de su pensión. Para entender lo que está ocurriendo, usted propone el siguiente modelo:

$$ln(MontoPension) = \beta_0 + \beta_1 Sexo + \beta_2 AnosCotizados + \beta_3 EdadInicio + \epsilon$$

Donde cada variable significa: ln(MontoPension) = logaritmo natural del monto de la primera pensión en pesos considerando el aporte del pilar solidario, Sexo = el sexo del cotizante (1 es hombre y 0 es mujer), AñosCotizados = número de años cotizados, EdadInicio = edad en la que empezó a cotizar.
Los resultados que obtiene son:

Source	SS	df	MS	Number of obs	=	464,249
				F(3, 464245)	=	71883.98
Model	64986.7865	3	21662.2622	Prob > F	=	0.0000
Residual	139900.399	464,245	.301350363	R-squared	=	0.3172
				Adj R-squared	=	0.3172
Total	204887.186	464,248	.441331327	Root MSE	=	.54895

MontoPension	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Sex	.2686708	.0017014	157.91	0.000	.2653361	.2720054
AnosCotizados	.0286848	.0000926	309.63	0.000	.0285033	.0288664
EdadInicio	-.0077055	.000076	-101.44	0.000	-.0078544	-.0075566
_cons	16.22094	.0036753	4413.47	0.000	16.21373	16.22814

(a) (5 puntos) Testee la significancia individual de al menos 2 parámetros del modelo. Explicite el test y el método utilizado.

Dado el output todos son significativos ya que su p-valor es menor a su significancia, en su intervalo de confianza no contiene al cero y el valor del estadístico es mayor al valor crítico.

El test sería:

$$H_0 : \beta_i = 0$$

$$H_0 : \beta_i \neq 0$$

con el estadístico:

$$t = \frac{\hat{\beta}_i}{\sigma_i}$$

(b) (5 puntos) Para cada variable, interprete el valor obtenido. ¿Qué puede aportar al debate? ¿Cómo se podrían mejorar las pensiones?

El sexo influye de manera positiva si se es hombre (por lo que medidas que busquen compensación a esto serían buenas)
Más años de cotización dan mayores montos de pensión (lo cual es intuitivo ya que si cotizo más tiempo debería recibir una pensión mayor)
La edad de inicio es negativa lo que implica que mientras antes se empieza a cotizar es mejor (va en la misma linea que los años cotizados)
Sobre el intercepto, su valor es de 16.22 lo cual es aproximadamente $ 11.070.000 en la variable MontoPension como nivel base

(c) (10 puntos) Un compañero de trabajo le dice que existe alta correlación en el modelo. Identifique al menos un par de variables independientes que pudiesen estar correlacionadas. ¿Cómo afecta esto a la estimación del modelo?

Dos variables que podrían estar fuertemente correlacionadas serían AñosCotizados y EdadInicio ya que se tendrán más años cotizados si la edad de inicio es menor, así ambas explican de cierta manera lo mismo en la regresión por lo que habría una alta multicolinealidad haciendo que el determinante de la matriz X'X sea cercano a cero generando variables con mucha varianza.

(d) (10 puntos) Otro compañero le dice que hay variables relevantes que no están incorporadas al modelo. ¿Qué variable podría ser? ¿Qué implicancias tiene esto en las estimaciones que hizo? Explicite matemáticamente el efecto.

Basta con que mencionen variables que hagan sentido en la regresión que se plantea respecto al monto de las pensiones, además de mencionar el sesgo de variable omitida, donde el efecto de cada parámetro asociado a las variables, el cual sería de la forma:

$$\mathbb{E}(\beta_i) = \beta_i + \beta_{omitido} \frac{Cov(Variable_i, Variable_{omit})}{Var(Variable_{omit})}$$

(e) (10 puntos) Finalmente, usted quiere saber si el modelo efectivamente tiene relevancia. Para ello, fíjese en el R^2 y además realice un test de significancia global. ¿Qué concluye?

Tanto R^2 como $R^2_{ajustado}$ son 0.3172 por lo que el modelo explica aproximadamente el 30 % de la variación del logarítmo de los montos de las pensiones, además notando de la tabla que tenemos $F_{3,464245} = 71883{,}98$ el cual es el estadístico para un test de significancia global respecto al modelo, cuyo valor es mucho mayor que el $F_{critico} = 2{,}6$ por lo que rechazamos la hipótesis nula de que el modelo no es significativo (al menos una de las variables si lo es)

Pregunta 3 (40 puntos). El tiempo de realización en minutos de una determinada tarea dentro de un proceso industrial es una variable aleatoria con función de densidad:

$$f(x) = \frac{x}{\theta^2} e^{\frac{-x}{\theta}}, x > 0$$

El parámetro $\theta > 0$ es desconocido. Buscamos estimarlo mediante el método de máxima verosimilitud. Para ello considere una muestra aleatoria simple de tamaño n.

(a) (10 puntos) Cuál es la función de verosimilitud? Explique intuitivamente que significa dicha función

La función de verosimilitud será:

$$\mathcal{L}(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$\mathcal{L}(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} e^{\frac{-x_i}{\theta}}$$

(b) (10 puntos) Explique en no más de 5 lineas en qué consiste un estimador de máxima verosimilitud

El método de máxima verosimilitud es un método que busca estimar el parámetro tal que ese parámetro haga que la muestra que obtuve sea la más probable de obtener en una distribución. Así se maximiza el valor de la función de verosimilitud la cual se basa en la función de la densidad de probabilidad para una distribución dada

(c) (5 puntos) Obtenga una expresión para la función de log-verosimilitud en base a la expresión obtenida en (a).

$$ln(\mathcal{L}) = ln \prod_{i=1}^{n} \frac{x_i}{\theta^2} e^{\frac{-x_i}{\theta}}$$

$$ln(\mathcal{L}) = \sum_{i=1}^{n} ln(\frac{x_i}{\theta^2}) + ln(e^{\frac{-x_i}{\theta}})$$

$$ln(\mathcal{L}) = \sum_{i=1}^{n} ln(x_i) - 2ln\theta - \frac{x_i}{\theta}$$

(d) (5 puntos) Demuestre que el estimador máximo verosímil viene dado por:

$$\theta_{EMV} = \frac{\bar{X}}{2}$$

$$\frac{\partial ln(\mathcal{L})}{\partial \theta} = -\frac{2n}{\theta} + \sum_{i=1}^{n} \frac{x_i}{\theta^2} = 0$$

$$\sum_{i=1}^{n} \frac{x_i}{\theta^2} = \frac{2n}{\theta}$$

$$\frac{1}{2n} \sum_{i=1}^{n} x_i = \hat{\theta}_{EMV}$$

$$\frac{1}{2n} \sum_{i=1}^{n} x_i = \hat{\theta}_{EMV}$$

$$\frac{\bar{X}}{2} = \hat{\theta}_{EMV}$$

(e) (5 puntos) Qué representa $\theta_{EMV} = \frac{\bar{X}}{2}$. Por qué lo llamamos estimador máximo verosimil? Explique intuitivamente.

Representa el valor del parámetro que al incluirlo hace más probable que se observe la función de densidad en esa muestra

(f) (5 puntos) Calcule el estimador de máxima verosimilitud para E[X]

$$\int_0^\infty xf(x)dx = \int_0^\infty \frac{x^2}{\theta^2} e^{-\frac{x}{\theta}} dx = \frac{1}{\theta^2} \int_0^\infty x^2 e^{-\frac{x}{\theta}} dx$$

Integrando por parte tenemos: $u = x^2$ $du = 2xdx$ $dv = e^{-\frac{x}{\theta}} dx$ $v = -\theta e^{-\frac{x}{\theta}}$

Así tendremos que:

$$= \frac{2}{\theta}(-\theta x e^{-\frac{x}{\theta}}|_0^\infty + \theta \int_0^\infty e^{-\frac{x}{\theta}} dx)$$

$$= \frac{2}{\theta}(0 - \theta^2 e^{-\frac{x}{\theta}}|_0^\infty)$$

$$= \frac{2}{\theta}(\theta^2) = 2\theta$$

Así el EMV de $\mathbb{E}(X) = 2\hat{\theta}_{EMV} = \frac{\sum X_i}{n} = \bar{X}$

Pregunta 1 (60 puntos). En Estados Unidos, al finalizar la educación escolar, se rinden las pruebas SAT para el ingreso a la universidad, donde hay una categoría de Lenguaje (SAT reading y SAT writing) y una de Matemáticas. Se le pide que realice un análisis de los puntajes promedios de la categoría de lenguaje y matemáticas para distintas escuelas, por lo que tienen los siguientes modelos:

$$\text{Average SAT Math} = \beta_0 + \beta_1 \,\%\text{ High Needs} + \beta_2 \text{Average Salary} + \beta_3 \,\%\text{ Females} +$$

$$\beta_4 \,\%\text{ African American} + \epsilon$$

$$\text{Average SAT Language} = \alpha_0 + \alpha_1 \,\%\text{ High Needs} + \alpha_2 \text{Average Salary} + \alpha_3 \,\%\text{ Females} +$$

$$\alpha_4 \,\%\text{ African American} + \varepsilon$$

Donde cada variable significa:

- % High Needs = Porcentaje de estudiantes de la escuela con altas necesidades (estudiantes vulnerables)
- Average Salary = Salario promedio de las familias de los estudiantes de esa escuela
- % Females = Porcentaje de mujeres en la escuela
- % African American = Porcentaje de personas de ascendencia afroamericana en la escuela

(a) (10 puntos) Testee la significancia individual de % Females y % African American para cada modelo. Explicite el test y el método utilizado.

El test sería:

$$H_0 : \beta_i = 0$$
$$H_0 : \beta_i \neq 0$$

con el estadístico:

$$t = \frac{\hat{\beta}_i}{\sigma_i}$$

Donde se podría rechazar o no rechazar por IC, p-valor o valor del estadístico, así en el modelo de Average SAT Math será no significativo % Females y significativo % African American mientras que para el segundo modelo de Average SAT Language ambas variables son significativas.

(b) (10 puntos) Interprete el valor obtenido para cada variable, ¿le hace sentido los valores? ¿Por qué los valores son distintos para cada prueba?

En este caso hacer fundamento de cada variable y que interpretación tendría para el caso de una prueba las variables sociodemográficas que se incluyen en el modelo (un mínimo de reflexión respecto a cada una para puntaje completo)

(c) (15 puntos) Un analista le menciona que sería buena idea incluir a la diversidad cultural que existe en las escuelas. ¿Cómo afecta esto a la estimación del modelo? ¿De qué manera sería bueno realizarlo?

Al incluirla se puede generar un problema de multicolinealidad con % African American que apunta a eso y ya están incluída, por lo que sería bueno incluirla y ver como afecta al $R^2_{ajustado}$ para en caso que mejore el R^2 pero empeore el $R^2_{ajustado}$, eliminar la variable % African American y ver como quedaría el nuevo modelo con diversidad cultural.

(d) (15 puntos) Mirando nuevamente el modelo, observa que hay variables relevantes que no están incorporadas. ¿Qué variables podrían ser? ¿Es posible no omitir variables?

Una variable podría ser cuanto gasta la escuela por estudiante, la cantidad de alumnos por curso, etc (variables que hagan sentido en este caso). No es posible no omitir variables, ya que un modelo completo puede incluir variables que no sean posible medirlas o que bien por temas de recurso o metodología nos sea imposible, por lo que se aspira a buscar modelos que expliquen de la mejor manera posible la variable dependiente dadas estas restricciones.

(e) (10 puntos) Para cada modelo compare el valor de R^2 y $R^2_{ajustado}$, ¿Qué puede comentar respecto a ello? ¿Son los modelos relevantes?

Modelo Average SAT Math $R^2 = 0{,}7547$ $R^2_{ajustado} = 0{,}7514$ Modelo Average SAT Language $R^2 = 0{,}784$ $R^2_{ajustado} = 0{,}7811$

Claramente ambos modelos tienen una penalización por la inclusión de variables no relevantes lo cual se muestra en que el $R^2_{ajustado}$ es menor al R^2, además se aprecia que el modelo de Average SAT Language explica con las mismas variables mucho mejor los puntajes que el modelo para Maths.

Los resultados que obtienen son:

Figura 1: Regresión Average SAT Math

Figura 2: Regresión Average SAT Language

Pregunta 2 (40 puntos). Se desea estudiar el efecto del factor X_1 en la variable y. Se sabe además que existe otro factor X_2, que también afecta y y que $Cov(X_1; X_2) = 0$. La población está descrita por el siguiente modelo lineal:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Donde se sabe que $\beta_2 \neq 0$ (esto es, X_2 si tiene un efecto en y). Se asume además que $E(\varepsilon_i | X_{1i}; X_{2i}) = 0$ y que los errores ï son independientes e idénticamente distribuidos. Se obtuvo una muestra $(y_i, X_{1i}; X_{2i})_{i=1,...,n}$, donde se observa que la varianza muestral de ambas covariables ($X_{1i} y X_{2i}$) es positiva y su covarianza muestral es cero. Dos analistas realizaron análisis independientes de los datos de la muestra.

- El Analista 1 ignoró la variable X_2, y realizó una regresión lineal simple vía Mínimos Cuadrados Ordinarios (MCO) con variable dependiente y_i y X_{1i} como única covariable (más una constante).
- El Analista 2 realizó una regresión lineal multivariada vía MCO con ambas covariables X_{1i} y X_{2i} (más una constante).

Comente la veracidad de las siguientes afirmaciones y justifique su respuesta.

(a) (10 puntos) El Analista 2 obtuvo un estimador insesgado de β_1; mientras que el estimador del Analista 1 es sesgado.

El supuesto clave para la consistencia y insesgadez del estimador MCO es que la esperanza del error, condicional en las covariables, sea cero. Para la regresión del analista 2, esto se cumple, luego el estimador de β_1 y β_2 es insesgado.

Pero el analista 1 no incluyó la covariable $X2$: Luego, esta covariable sera absorbida en el error: $yi = \beta_0^* + \beta_1^* X_{1i} + \epsilon_i^*$

donde $\epsilon_i = \epsilon_i + \beta_2 X_{2i}$. Esta regresión lineal da como coeficiente $\beta_1^* = Cov(y, X_1)/Var(X_1)$. Reemplazando y por el modelo original:

$$Cov(y, X_1) = Cov(\beta_1 X_1, X_1) + Cov(\beta_2 X_2, X_1) + Cov(\epsilon_i, X_1)$$

Bajo los supuestos del problema, los dos últimos términos valen cero, luego $\beta_1^* = \beta_1$. Conclusión: **la aseveración es falsa**, los dos estimadores son insesgados.

(b) (10 puntos) Ambos analistas obtuvieron exactamente el mismo R^2. **Falso.** El analista 2 siempre obtendra un R^2 mayor porque la regresión incluye una covariable adicional con respecto a la regresión del analista 1.

(c) (10 puntos) El error estándar asociado al estimador de β_1 obtenido por el Analista 2 es menor que para el Analista 1.

Correcto. La formula del error estandar esta dada por:

$$SE(\hat{\beta}_k) = \frac{\hat{\sigma}^2}{SST_k(1 - R_k^2)}$$

Como el analista 2 incluye una covariable adicional, la varianza de los residuos es menor ($\hat{\sigma}^2$ más chico). SST_k es igual para los dos. Como la covarianza muestral entre X_1 y X_2 es cero, R_k^2 es cero. Luego el error estándar es inequívocamente menor para el analista 2.

(d) (10 puntos) El Analista 1 obtuvo una constante (estimador de β_0) menor que el del Analista 2.

El estimador de la constante para el analista 2 es $\hat{\beta}_0 = \hat{y} - \hat{\beta}_1\hat{X}_1 - \hat{\beta}_2\hat{X}_2$ y para el analista 1 $\hat{\beta}_0 = \hat{y} - \hat{\beta}_1\hat{X}_1$. Vimos que ambos obtienen el mismo $\hat{\beta}_1$, luego no esta claro si el estimador de la constante es mayor o menor para el analista 1 (depende de $\hat{X}_2$ y la magnitud de $\hat{\beta}_2$).

Pregunta Bonus (40 puntos) Sea $Y_1, Y_2, ..., Y_n$ una secuencia de variables aleatorias normales e independientes. Es decir que la densidad marginal de cada una de ellas está dada por:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}}$$

Además se sabe que $\mathbb{V}(Y_i) = \sigma^2$ y $\mathbb{E}(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$ donde β_0, β_1 y X_i son valores fijos (no aleatorios).

(40 puntos) Obtenga el estimador de máxima verosimilitud para β_0 y β_1 en función de Y_i y X_i.

La verosimilitud viene dada por:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}}$$

$$ln(L(\beta_0, \beta_1)) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial ln L(\beta_0, \beta_1)}{\partial \beta_0} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

$$\frac{\partial ln L(\beta_0, \beta_1)}{\partial \beta_1} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$