Modelos de Regresión Aplicados

Lab de práctica.

Esta práctica la van a realizar en clase. Se hace en grupos. Para obtener una calificación de 100, se tiene que entregar antes de las 11:50 del jueves 10 de noviembre. Si no terminan antes, pueden enviar la respuesta a [Gilbert.brenes.camacho@gmail.com](mailto:Gilbert.brenes.camacho@gmail.com), pero la nota base es de 90.

Toman el archivo antropometria.Rdata, con la submuestra del mes que le tocó a cada persona o grupo (Tienen que hacer un subset, tomando la variable id0a, que es numérica y en formato añomesdia).

Para ello, haga un subconjunto de datos. Por ejemplo,

datos <-antropometria[(anyomes==201012 | anyomes==201101),]

Si son un grupo, juntan los dos o tres meses que les corresponde.

Recuerden que el objetivo principal del análisis es estudiar qué variables antropométricas logran predecir el VEF, controlando además por edad, sexo y condición de fumador. Con el archivo de datos, tiene que estimar modelos con todas las variables y modelos reducidos (a partir de la selección de variables usando stepwise), con las medidas remediales solicitadas más abajo. Para el modelo reducido, primero lo escoge sin ninguna medida remedial, y después le aplica las medidas remediales. Para todos los modelos tiene que evaluar normalidad, homoscedasticidad y calcular el VIF más alto. Puede apoyar sus decisiones en gráficos, que le puede enseñar al profesor para justificar su decisión. Los resultados de los modelos se tienen que usar para llenar las tablas 1 y 2 que están más abajo. Los modelos con medidas remediales son:

1. Modelo con la variable dependiente transformada con una transformación de BoxCox.
2. Modelo con la transformación de la variable dependiente sugerida por el gráfico spread.level.plot().
3. Modelo heteroscedástico
4. Modelo con estimaciones robustas a presencia de valores extremos.
5. Modelo corrigiendo los errores estándar con mínimos cuadrados ponderados (use los predichos para las ponderaciones).

Al final, tiene que contestar cuál modelo escoge y por qué?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estudiante | Mes |  | Estudiante | Mes |
| AMANDA | 200411 |  | JOSTIN | 200605 |
| ANA PAULA | 200412 |  | JOSUE DAVID | 200606 |
| ANDREA | 200501 |  | JUAN JOSE | 200607 |
| BRANDON JOSUE | 200502 |  | KEYLA | 200608 |
| CESAR JOSUE | 200503 |  | LAURA VANESSA | 200609 |
| DANIEL ALFARO | 200504 |  | LUIS FERNANDO | 200411 |
| DANIEL SIBAJA | 200505 |  | MANRIQUE | 200412 |
| FELIPE | 200506 |  | MIRIAM | 200501 |
| IVAN DANIEL | 200507 |  | NATASHA | 200502 |
| JAMES | 200508 |  | OLGER LEONARDO | 200503 |
| JOSE PABLO | 200509 |  | PAMELA | 200504 |
| JOSEPH | 200510 |  | PATSY | 200505 |
|  | 200511 |  | YARLIN | 200506 |
|  | 200512 |  |  | 200507 |
|  | 200601 |  |  | 200508 |
|  | 200602 |  |  | 200509 |
|  | 200603 |  |  | 200510 |
|  | 200604 |  |  | 200511 |
|  |  |  |  |  |

Tabla 1. Modelo para predecir el VEF con todo el conjunto de variables predictoras en la base de datos, usando distintas medidas remediales. (En la columna prob solo pongan un asterisco si el coeficiente es significativamente distinto de cero).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Modelo sin corregir | | Mod. con transf. Box-Cox | | Mod. con transf spread.level.plot | | Mod. heteroscedástico | | Mod. estimac. robustas | | Mod. Mínimos Cuadrados Pond | |
| Variable | Coef | prob | Coef | prob | Coef | prob | Coef | prob | Coef | Prob (tc) | Coef | prob |
| age | -3.1839 | 6.29e-05 \*\*\* | -0.1513075 | 2.35e-05 \*\*\* | -55.223 | 0.000259 \*\*\* | -3.23388104 | 2.124697e-06 | -3.2525 | -3.9465\* | -3.14250 | 2.03e-05 |
| cinturar | -0.0111 | 0.9899 | 0.0007611 | 0.9846 | -0.505 | 0.975976 | -0.09026045 | 9.054858e-01 | 0.0213 | 0.0228 | -0.07178 | 0.9285 |
| caderar | -1.6092 | 0.2390 | -0.0756444 | 0.2173 | -28.370 | 0.276227 | -1.80354009 | 6.802451e-02 | -1.8551 | -1.2847 | -1.71779 | 0.1327 |
| grip | 5.5598 | 1.59e-06 \*\*\* | 0.2417357 | 3.15e-06 \*\*\* | 106.270 | 1.51e-06 \*\*\* | 5.28021534 | 1.603105e-06 | 5.2502 | 4.4174\* | 5.43177 | 1.87e-06 \*\*\* |
| pantorrilla | 2.5727 | 0.3627 | 0.1024020 | 0.4191 | 53.447 | 0.321449 | 2.41193739 | 2.654926e-01 | 2.8288 | 0.9467 | 2.54468 | 0.3005 |
| brazo | 2.5086 | 0.3709 | 0.1055469 | 0.4011 | 51.059 | 0.339558 | 2.12369824 | 3.562805e-01 | 2.2876 | 0.7718 | 2.37726 | 0.3189 |
| tricipital | -3.0093 | 0.0332 \* | -0.1080377 | 0.0875 | -67.134 | 0.012909 \* | -2.33915114 | 4.892719e-02 | -3.1441 | -2.1153\* | -2.65855 | 0.0418 \* |
| subscap | 0.7789 | 0.5271 | 0.0264051 | 0.6325 | 17.708 | 0.450876 | 0.49617902 | 5.912960e-01 | 0.7874 | 0.6044 | 0.60764 | 0.5753 |
| pesokg | 0.3984 | 0.8327 | 0.0164196 | 0.8461 | 7.821 | 0.827833 | 1.00410623 | 5.398055e-01 | 0.6911 | 0.3462 | 0.71215 | 0.6903 |
| tallacm | 1.6775 | 0.2403 | 0.0915262 | 0.1536 | 23.360 | 0.390791 | 0.60982803 | 6.239095e-01 | 2.0142 | 1.3345 | 1.28106 | 0.3357 |
| smoker | 16.5766 | 0.4070 | 0.9665395 | 0.2814 | 223.341 | 0.557793 | 17.44041272 | 2.758510e-01 | 17.7608 | 0.8401 | 17.79794 | 0.3122 |
| hombre | -20.3980 | 0.3920 | -1.0715500 | 0.3162 | -306.560 | 0.499713 | -8.17861010 | 6.973664e-01 | -27.4153 | -1.0881 | -16.06807 | 0.4727 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Prueba de normalidad |  | 0.9202 |  | 0.522 |  | 0.9543 |  | 0.856 |  | 0.9103 |  | 0.9027 |
| Prueba de Variancia constante (ncvTest) |  | 0.0027567 |  | 0.9397 |  | 0.064348 |  | 0.01124 (Breusch Pagan) |  | 0.0013299 |  | 0.1489 |
| R2 |  | 0.4959 |  | 0.4975 |  | 0.4985 |  |  |  |  |  | 0.5042 |
| AIC |  | 2372.463 |  | 1130.692 |  | 1915.6 |  | 2371.81 |  | 2372.822 |  | 2354.889 |
| P2 |  | 0.4650715 |  | 0.4671529 |  | 0.4679823 |  | 0.4314586 |  | 0.4704686 |  |  |

Nota: \*: p<0.05

Tabla 2. Modelo reducido para predecir el VEF con todo el conjunto de variables independientes sugeridas en la Tarea 2, usando distintas medidas remediales. (En la columna prob solo pongan un asterisco si el coeficiente es significativamente distinto de cero).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Modelo sin corregir | | Mod. con transf. Box-Cox | | Mod. con transf spread.level.plot | | Mod. heteroscedástico | | Mod. estimac. robustas | | Mod. Mínimos Cuadrados Pond | |
| Variable | Coef | prob | Coef | prob | Coef | prob | Coef | prob | Coef | Prob t© | Coef | prob |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| age | -3.5934 | 8.30e-07 \*\*\* | -0.17049 | 2.04e-07 \*\*\* | -1.5348 | 5.05e-07 \*\*\* | -3.3965770 | 1.002327e-06 | -3.7451 | -5.1760\* | -3.5086 | 9.05e-07 \*\*\* |
| grip | 5.8500 | 3.56e-08 \*\*\* | 0.24821 | 1.65e-07 \*\*\* | 2.4170 | 5.12e-08 \*\*\* | 5.3968957 | 3.396713e-07 | 5.5230 | 5.2854\* | 5.6973 | 7.19e-08 \*\*\* |
| tricipital | -1.5054 | 0.0532 | -0.05570 | 0.110 | -0.6004 | 0.0653 | -0.9330558 | 1.843182e-01 | -1.4257 | -1.7964 | -1.2657 | 0.0893 |
| tallacm | 1.6037 | 0.0825 | 0.04120 | 0.051 | 0.6993 | 0.0704 | 1.8807957 | 2.996936e-02 | 1.7792 | 1.8883 | 1.6984 | 0.0582 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Prueba de normalidad |  | 0.5844 |  | 0.2925 |  | 0.6575 |  | 0.524 |  | 0.5602 |  | 0.565 |
| Prueba de Variancia constante (ncvTest) |  | 0.0038796 |  | 0.862 |  | 0.04355 |  | 0.003468 |  | 0.0072784 |  | 0.079868 |
| R2 |  | 0.5036 |  | 0.5053 |  | 0.5059 |  |  |  |  |  | 0.5028 |
| AIC |  | 2361.792 |  | 1119.927 |  | 2013.287 |  | 2355.521 |  | 2361.988 |  | 2352.874 |
| P2 |  | 0.4874245 |  | 0.4893334 |  | 0.4898515 |  | 0.4462451 |  | 0.4817902 |  |  |

Nota: \*: p<0.05

Conteste: Cuál modelo escoge y por qué?

El modelo que recomendaríamos sería el modelo reducido con transformación BoxCox. El modelo tiene el segundo R2 ajustado más grande de todos los modelos planteados, quedando atrás del modelo reducido modificado con el valor recomendado del spreadlevel. La diferencia decisiva entre ambos modelos se da en que, con el modelo ajustado con BoxCox, se cumple tanto el supuesto de normalidad como el de heteroscedasticidad, mientras que con el modelo ajustado con el valor del spreadlevel no se corrige la heteroscedasticidad. Del mismo modo, el menor AIC encontrado se da en el modelo reducido corregido con la transformación BoxCox. Con respecto a los modelos con todas las variables, preferimos el modelo reducido no solo porque es más parsimonioso, sino también porque todos los indicadores son mejores en los modelos reducidos. Del mismo modo, el modelo reducido con corrección de BoxCox tiene una cantidad mayor o igual de coeficientes significativos que cualquier otro modelo, específicamente con dos (age y grip). Asimismo, al analizar violación a otros supuestos, el modelo reducido con transformación de boxcox permite observar una nube horizontal al comparar residuos contra valores predichos y vifs menores a 3, por lo que se puede asumir también no multicolinealidad, por lo que transformaciones adicionales podrían no ser necesarias. Al evaluar el supuesto de linealidad, se observa que se cumple para todas las variables predictoras del modelo reducido y la variable dependiente, excepto para tricipital; sin embargo, se mantiene la variable, a pesar de esto y de no ser significativo su coeficiente, por recomendación del proceso del AIC.