Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística

Trabajo Final

Modelo lineal general II

Caso 1 (70 %)

Por medio de análisis de regresión se quiere determinar los factores que más influyen sobre el salario de jugadores profesionales de la NBA. Adicionalmente, se quiere proponer un modelo predictivo. Para alcanzar estos objetivos, se tomó una muestra de 216 jugadores profesionales durante las temporadas 2010-2011 y 2012-2013. Dado que el salario que se acuerda en los contratos está influenciado por el rendimiento en las temporadas anteriores, la base de datos (`NBA.csv`) contiene diferentes indicadores de juego en la última temporada del contrato previo y el salario acordado en el nuevo contrato.

Las variables observadas fueron:

- `Player`: nombre del jugador.
- `Position`: posición de juego en la mayoría de partidos.
- `Salary`: salario anual (en miles de dólares) acordado en el contraro.
- `Length.of.Contract`: duración (en años) del contraro.
- `Age`: edad.
- `GS`: número de juegos iniciados.
- `MP`: promedio de minutos jugados por partido.
- `FG`: razón del promedio de tiros acertados/realizados por juego.
- `FG`: razón del promedio de tiros acertados/realizados por juego.
- `X3P`: razón del promedio de tiros más allá la línea de tres puntos acertados/intentados por juego.
- `X2P`: razón del promedio de tiros dentro de la línea de tres puntos acertados/intentados por juego.
- `FT`: razón del promedio de tiros libres acertados/realizados por juego.
- `TRB`: promedio de rebotes capturados por juego.
- `AST`: promedio de asistencias por juego.
- `STL`: promedio de robos por juego.
- `BLK`: promedio de bloqueos por juego.
- `TOV`: promedio de perdidas de balón por juego.
- `PF`: promedio de faltas personales por juego.
- `PTS`: promedio de puntos anotados por juego.
- `PER`: índice de eficiencia del jugador. Medida global de rendimiento del jugador. 0-10 bajo, 10-20 medio, 20-25 estrella, 25-30 extraordinario, 30+ super-estrella.

- `TS`: porcentaje de disparo real. Indicador que refleja que tan bien dispara un jugador. Se calcula de la misma forma que FG pero considerando la dificultad de los tiros. Por ejemplo, los tiros de 3 puntos acertados tienen mayor peso.
- `USG`: porcentaje de uso. Estimación del porcentaje de jugadas de equipo hechas por el jugador mientras estaba en la cancha.
- `ORtg`: Calificación ofensiva. Estimación de los puntos producidos por cada 100 posesiones del jugador.
- `DRtg`: Calificación defensiva. Estimación de los puntos permitidos por cada 100 posesiones del jugador.
- `OWS`: Acciones ganadoras ofensivas. Estimación del número de victorias aportadas por el jugador debido únicamente a la producción ofensiva.
- `DWS`: Acciones ganadoras defensivas. Estimación del número de victorias aportadas por el jugador debido únicamente a la producción defensiva.
- `WS`: Acciones ganadoras globales. Estimación general del número de victorias aportadas por un jugador.

Para el análisis de los datos, divida la muestra en dos partes de forma aleatoria: una para estimar el modelo - entrenamiento - (190 observaciones) y otra para hacer una evaluación del modelo ajustado - validación - (26 observaciones)

Con la sub-muestra de entrenamiento ajuste un modelo de regresión lineal considerando la variable `Salary` como respuesta y todas las demás como covariables (excepto `Player`, obviamente). Si es necesario, utilice transformaciones y/o elimine (pocas) observaciones atípicas. En caso de eliminar a algún jugador, justifique claramente porque considera que es atípico. Evalúe multicolinealidad. Comente los resultados más relevantes.

Realice un proceso de selección de variables. Proponga al menos tres modelos diferentes. Justifique claramente el criterio de selección utlizado. Compare estos modelos con el modelo completo. Comente los resultados más relevantes.

A partir de los modelos propuestos (y el modelo completo) haga predicciones para los jugadores de la sub-muestra de validación. Compare los modelos usando el error cuadrado medio de predicción.
A partir de este análisis, ¿cuáles son los factores que más influyen sobre el salario de los jugadores? y ¿los modelos propuestos proporcionan buenas predicciones?

Caso 2 (30 %)

Los datos `docvisit.csv` corresponden a una submuestra de una encuesta de gastos médicos de los Estados Unidos para el año 2003. Esta contiene información de población mayor de 65 perteneciente a Medicare (programa nacional de seguridad social que proporciona cobertura de seguro médico básico). Particularmente, se quiere determinar los factores que influyen significativamente sobre las visitas al médico.
Para esto, se debe ajustar un modelo para el número de visitas anuales al médico (`DOCVIS`) en función de las siguientes covariables:

- `educyr`: años de educación.
- `totchr`: número de condiciones crónicas.
- `private`: seguro privado complementario al seguro médico del estado (1: si, 0: no).
- `medicaid`: seguro complementario para personas de bajos ingresos (1: si, 0: no).
- `age`: edad del paciente.

- `income`: ingreso familiar total anual (miles de dolares).
- `female`: genero (1: mujer, 0: hombre).

Considere como primera opción un modelo Poisson. Evalúe si un modelo que tenga en cuenta sobredispersión o inflación de ceros proporciona mejores resultados.

A partir del ajuste del mejor modelo identifique que factores influyen positiva/negativamente sobre el número de visitas al médico.

Aspectos a tener en cuenta

El reporte final no puede exceder 10 páginas. Puede presentarse en grupos máximo de 2 integrantes.

Todas las tablas y figuras deben estar enumeradas. Solo incluya tablas y figuras que sean relevantes (es decir, deben tener una referencia en el texto). Recuerde que el número de páginas es limitado, así que use el espacio de forma inteligente.

Para cada caso, debe incluir una **sección con las conclusiones y recomendaciones**.

En el reporte no incluya códigos de R.

Fecha de entrega: 26 de julio (físico y a través del campus virtual).
Fecha de sustentación: 28 de julio a partir de las 7am.