Primer trabajo práctico: Modelo lineal

El trabajo práctico puede realizarse en grupos de como máximo 2 personas. Para la entrega debe crearse un informe explicando cada ítem, con los gráficos y justificaciones que crea pertinentes. La entrega debe realizarse en un archivo formato pdf, a jmgarcia@fi.uba.ar. Fecha límite de entrega: domingo 19 de junio.

El dataset *car-following_trajectory.csv* se cuenta con 120 observaciones de las siguientes variables:

- v_Vel: velocidad en m/s del vehículo de referencia (ego)
- v_Acc: aceleración del vehículo de referencia en m/s^2
- Space_Headway: distancia en metros entre el vehículo de referencia y el auto de enfrente
- Preceding_Distance: Distancia en movimiento con respecto al auto de enfrente en la medición anterior
- Following: un indicador que vale 1 si el vehículo de referencia está siendo seguido por otro vehículo.
- Local_Y_diff: Variación de distancia (en metros) del vehículo de referencia.

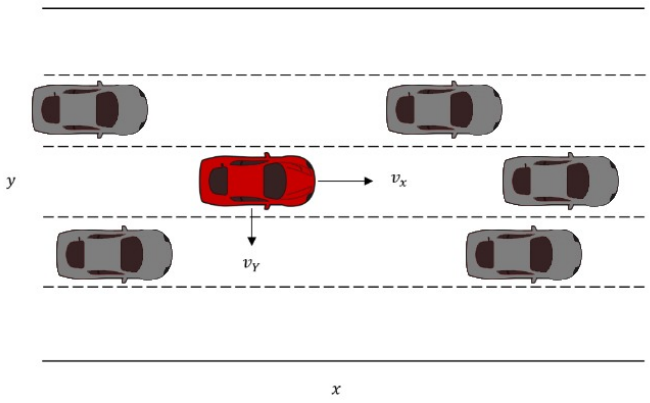

Figura 1: Vehículo de referencia (ego) en rojo

1. Cargar los datos del archivo. La variable Following es una variable categórica donde el 1 indica si el auto está siguiendo y 0 si no. Transformarla en un factor. Finalmente revisar que todas las variables contenidas en el dataframe estén correctamente definidas.

2. Se desea ajustar un modelo de regresión múltiple para predecir la variable $Local\_Y\_diff$ en función del resto de las variables en el data set. Escribir el modelo propuesto, indicando los supuestos del mismo.

3. Realizar un scatterplot de las variables con la función $ggpairs$.

4. A partir de la tabla de correlaciones estimadas entre las variables, si tuviera que elegir una sola variable para proponer un modelo de regresión simple, ¿cuál eligiría y porqué?

5. Realizar un ajuste de regresión lineal múltiple. ¿Es la regresión significativa? Especificar las hipótesis nula y alternativa de este test. ¿Cómo se calcula el p-valor en este caso? ¿Rechazaría a un nivel de significación de 0.05?.

6. A partir de la tabla de coeficientes estimados, ¿Qué variables resultan significativas? ¿A qué nivel? ¿Cuál es el valor de la estimación para σ^2? Especificar las hipótesis nulas y alternativas para *alguno* de los test t reportados en la tabla, el estadístico del test y la regla de decisión. ¿Cómo se calcula el p-valor para este test?

7. Evaluar la bondad del ajuste realizado, a través del coeficiente de determinación. Indicar cuánto vale y qué significa.

8. Validar los supuestos expresados en el ítem 2 a partir del análisis de los residuos, para el modelo seleccionado. ¿Observa algo extraño en los gráficos? ¿Qué propone?

9. ¿Cuál sería la estimación de la esperanza de la variable a predecir para una observacion con los siguientes valores: $v\_vel = 4{,}06, v\_Acc = 0{,}0568, Space\_Headway = 8{,}575, Preceding\_Distance = 7{,}62, Following = 1$

10. Hallar un intervalo de confianza y de predicción de nivel 0.95 para la estimación hallada en el ítem anterior.

11. *Selección de modelos.* Plantear un nuevo modelo en el que intervengan aquellas variables que contribuyen significativamente y estimar los parámetros por mínimos cuadrados. ¿Qué modelo elegiría finalmente? Utilizar medidas de bondad de ajuste y de predicción, tal como la estimación del error cuadrático medio de validación cruzada.