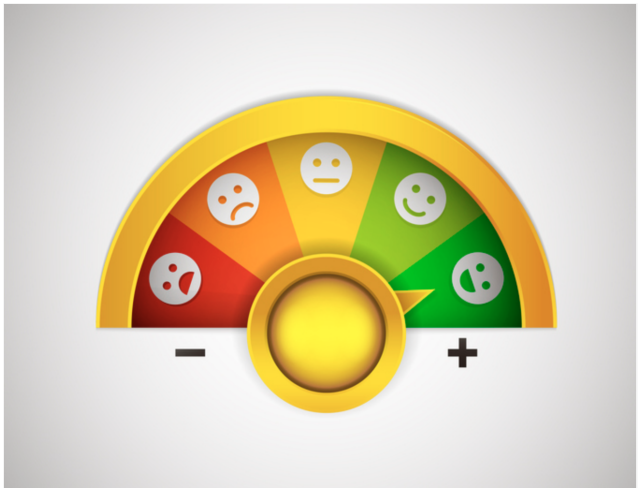
Segundo bimestre de 2022 TP 2: Modelo lineal generalizado

Herramientas de Modelado Estadístico

Reglas del TP:

- Este trabajo debe hacerse de forma individual o de a dos.
- Deben enviarse el script de R y un informe que contenga las respuestas a todas las preguntas y los gráficos pedidos, en lo posible en RMarkdown. No hace falta explicar en el informe que es lo que hace cada una de las funciones del script.
- El script debe estar prolijo. Esto en particular implica que las variables tienen que tener un nombre descriptivo (es decir, no llamar `a`, `b`, `c` a las variables).

Regresión ordinal



Vamos a usar un conjunto de datos correspondiente a una encuesta con escala de tipo Likert (es decir, se pide al encuestado marcar un entero entre 1 y 5, donde 1 = Totalmente en desacuerdo y 5 = Totalmente de acuerdo). La encuesta consiste de 44 preguntas muy variadas, como "Disfruto de bailar" o "Creo que un desastre climático podría llegar a ser divertido". Para los individuos encuestados, se tienen también otras variables extra-encuesta que pueden ser de interés, como por ejemplo edad, género, religión, etc. Los datos están `encuesta.csv` y el archivo `codebook.txt` contiene una descripción de las preguntas y las variables extra-encuesta.

1. Dividir al conjunto de datos en data de entrenamiento y testeo.
2. Leer las preguntas y elegir alguna que parezca interesante. Llamaremos Q a esta pregunta.
3. Supongamos que queremos modelar la respuesta Q en función de la edad y el género. ¿Cuál sería el problema teórico de usar una regresión lineal para esto? ¿Cuál sería el problema de usar una regresión multinomial en este problema?

4. Leer (en Wikipedia, por ejemplo) acerca de **Regresión Ordinal**. Explicar, en tus propias palabras, en qué consiste este modelo.

5. Usando el paquete `MASS` y la función `polr`, aplicar el modelo de regresión ordinal para predecir Q en función de la edad.

6. Estimar la probabilidad de que a una persona de 25 años esté **al menos** de acuerdo con la frase "me gustan las armas" (pregunta 9).

7. Para la pregunta Q, definamos la siguiente función de pérdida:

$$L(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

   donde y_i es la respuesta del individuo i a la pregunta Q y $\widehat{y_i}$ es la correspondiente predicción. Notar que tanto y_i como $\widehat{y_i}$ son números enteros entre 1 y 5. Implementar esta funcón de pérdida en `R`.

8. Implementar un modelo lineal que prediga la respuesta a la pregunta Q en función de la edad. Este modelo tendrá predicciones $\widehat{y}$ que pertenecen a toda la recta real. Para hacerlo comparable con el modelo de regresión ordinal, tomar como predicción final al número entero entre 1 y 5 más cercano a y_i.

9. Comparar el valor de la pérdida L para el modelo de regresión ordinal y el modelo de regresión lineal (modificado) del item anterior, aplicando ambos. Decidir cuál de los dos es preferible (recordar entrenar los modelo en el conjunto de entrenamiento y evaluarlo en la data de testeo)

10. Probar al menos 5 modelos modelos que le parezca que tengan sentido, agregando nuevas variables, interacciones, probando otros algoritmos, etc, intentando minimizar la pérdida L.

11. Consideremos ahora el problema de regresión logística que consiste en predecir si una persona tiene un titulo universitario o no (ver variable `education`, categorías 3 y 4). Para esto se pueden tomar como covariables las preguntas que deseen y el resto de las variables extra-encuesta. Elegir un modelo M para este problema de clasificación. Tomaremos como pérdida la exactitud, es decir, la proporción de predicciones correctas.

12. Analizar cómo varía la exactitud en función del tamaño de muestra. Usar una regresión beta para predecir cuál es el mínimo tamaño de muestra necesario para obtener una exactitud de 0.9.

Importante / para discutir: en general está mal armar modelos que usen al género, raza y religión como covariables. La data de entrenamiento representa al status quo, que puede reflejar sesgos discriminatorios en ciertos grupos. El modelo no sabe de ética y, por lo tanto, va a reproducir esos sesgos. En este TP se pueden usar, pero en general hay que tener cuidado.