Santiago de Cali, Octubre de 2025
Universidad Autónoma de Occidente
Inferencia Estadística
Desafío #3
Alejandro Meneses Portilla.

Situación 1

Se presenta un modelo compuesto por un sistema de ecuaciones de la siguiente manera:

$$Y_1 = \theta + \epsilon_1 \quad Y_2 = 2\theta - \varphi + \epsilon_2 \quad Y_2 = \theta + 2\varphi + \epsilon_3$$

Dado que el objetivo del método de los Mínimos Cuadrados es encontrar las estimaciones de los parámetros θ y φ para un modelo lineal que **minimice** la suma de los cuadrados de los errores. Para esto, se debe despejar el error de cada ecuación del sistema:

$$\epsilon_1 = Y_1 - \theta \quad \epsilon_2 = Y_2 - (2\theta + \varphi)\, \epsilon_3 = Y_3 - (\theta - 2\varphi)$$

Para obtener las estimaciones de θ y φ, se debe minimizar la función de suma de los cuadrados de los errores, $S(\beta_0, \beta_1)$, definida como:

$$S(\theta, \varphi) = \sum_{i=1}^{3} (\epsilon_i)^2 = (Y_1 - \theta)^2 + (Y_2 - (2\theta + \varphi))^2 + (Y_3 - (\theta - 2\varphi))^2$$

Para encontrar los valores que minimizan $S(\theta, \varphi)$ se toman las derivadas (aplicando regla de la cadena) con respecto a cada parámetro y se iguala el resultado a cero:

$$\frac{dS}{d\theta} = 2(Y_1 - \theta)(-1) + 2(Y_2 - (2\theta + \varphi))(-2) + 2(Y_3 - (\theta - 2\varphi))(-1)$$

$$0 = -2Y_1 + 2\theta - 4Y_2 + 8\theta - 4\varphi - 2Y_3 + 2\theta + 4\varphi$$

$$0 = -2Y_1 - 4Y_2 + 12\theta - 2Y_3 \rightarrow \theta = \frac{2(Y_1 + 2Y_2 + Y_3)}{12}$$

$$\theta = \frac{Y_1 + 2Y_2 + Y_3}{6}$$

Haciendo lo mismo para φ tenemos:

$$\frac{dS}{d\varphi} = 2(Y_2 - 2\theta + \varphi) - 4(Y_3 - \theta - 2\varphi)$$

$$0 = 2Y_2 - 4\theta + 2\varphi - 4Y_3 + 4\theta + 8\varphi$$

$$0 = 2Y_2 - 4Y_3 + 10\varphi \rightarrow \varphi = \frac{2(-Y_2 + 2Y_3)}{10}$$

$$\varphi = \frac{-Y_2 + 2Y_3}{5}$$

Por lo tanto, los estimadores son: $\theta = \frac{Y_1 + 2Y_2 + Y_3}{6}$ y $\varphi = \frac{-Y_2 + 2Y_3}{5}$.

Santiago de Cali, Octubre de 2025
Universidad Autónoma de Occidente
Inferencia Estadística
Desafío #3
Alejandro Meneses Portilla.

Situación 2 (ver funcion situacion_2()**)**

En esta situación se busca estudiar cómo la dieta de los caimanes depende principalmente de dos factores: **Lago de Origen** y **el tamaño** (< 2.3m vs >2.3m). La dieta se clasifica en 5 categorías: **Peces**, **Invertebrados**, **Reptiles**, **Aves** y **Otros.**

Dado que las variables son nominales y son múltiples, se utiliza un **modelo de regresión logística multinomial** que extiende la regresión logística binaria a **más de dos categorías de respuesta**, para ello se utilizan los siguientes supuestos:

- **Peces** como variable de referencia.
- **Lago y tamaño** como predictores en el modelo.
- Se calculan los coeficientes, ratios odd, intervalos de confianza y valores p para evaluar significancia estadística.

A. Dado que el modelo ajustado es un **modelo de regresión logística multinomial** y la categoría de referencia se fija en **Peces,** se tiene el siguiente modelo:

$$log\left(\frac{P(Y=i)}{P(Y=Peces)}\right) = \beta_{0i} + \beta_{1i} + \beta_{2i} \; donde \; i \in \{Invertebrados, \; Aves, \; Reptiles, \; Otros\}.$$

B. Las razones de chance (u “Odds Ratios”) permiten interpretar el efecto de cada variable sobre la probabilidad de elegir un tipo de alimento en lugar de la variable de referencia (en este caso, **Peces**). A continuación se presenta la tabla con estas razones:

C.

Variable	Odds Ratio	Desv. Estándar	Valor P.
Aves (vs. Peces)			
(Intercept)	0.0656	0.7104	0.0001
LagoHancock	2.0041	0.7813	0.3735
LagoOcklawaha	0.5203	1.2021	0.5868
LagoTrafford	2.9679	0.8417	0.1962
Tamano>2.3	1.8790	0.6425	0.3262
Invertebrados (vs. Peces)			
(Intercept)	0.9132	0.3080	0.7681
LagoHancock	0.1905	0.6129	0.0068

LagoOcklawaha	2.5527	0.4719	0.0470
LagoTrafford	3.0709	0.4905	0.0222
Tamano>2.3	0.2327	0.3959	0.0002
Otros (vs. Peces)			
(Intercept)	0.2075	0.4748	0.0009
LagoHancock	2.2848	0.5575	0.1383
LagoOcklawaha	1.0055	0.7766	0.9943
LagoTrafford	4.5559	0.6214	0.0147
Tamano>2.3	0.7179	0.448	0.4598
Reptiles (vs. Peces)			
(Intercept)	0.0256	1.0590	0.0005
LagoHancock	3.4666	1.1854	0.2943
LagoOcklawaha	11.6910	1.1181	0.0279
LagoTrafford	18.8271	1.1164	0.0086
Tamano>2.3	1.4210	0.5800	0.5446

D. En la tabla anterior, el **Tamano > 2.3** refleja el efecto de ser un caimán grande versus uno pequeño (≤ 2.3 m), tomando como referencia **Peces**. Teniendo en cuenta esto:

- **Invertebrados**

$$Coeficiente = -1.4581,\ Desv.\ Estándar = 0.3959$$
$$Odds\ Ratio = exp(-1.4581) \approx 0.2327$$

Esto indica que los caimanes grandes tienen **77% menos probabilidad de elegir invertebrados versus peces** que los caimanes pequeños.

- **Reptiles**

$$Coeficiente = 0.3514,\ Desv.\ Estándar = 0.5800$$
$$Odds\ Ratio = exp(0.3514) \approx 1.4210$$

Los caimanes grandes tienen **42% más probabilidad de elegir reptiles frente a peces**, aunque el error estándar es relativamente grande.

- **Aves**

$$Coeficiente = 0.6307,\ Desv.\ Estándar = 0.6425$$
$$Odds\ Ratio = exp(0.6307) \approx 1.8789$$

Los caimanes grandes parecen tener casi el doble de probabilidad de elegir aves frente a peces. Sin embargo, la incertidumbre es alta (Desv. Estándar = 0.64), por lo que no parece determinante en primera instancia.

- **Otros**

$$Coeficiente = -0.3314,\ Desv.\ Estándar = 0.4483$$
$$Odds\ Ratio = exp(-0.3314) \approx 0.7179$$

Los caimanes grandes tienen 28% **menos probabilidad** de elegir "otros" alimentos frente a peces, aunque el efecto **no es significativo** (coeficiente pequeño).

Conclusión:

El tamaño de los caimanes **es un factor determinante en su dieta**: los pequeños parecen consumir más presas pequeñas (invertebrados), mientras que los grandes tienden a cazar presas de mayor tamaño (reptiles, aves, incluso mamíferos). Este resultado coincide con la hipótesis de que la necesidad energética y la capacidad de depredación aumentan con el tamaño corporal.

E. Aquí se comparan los lagos con respecto a la referencia **George** (porque en el modelo de **multinom** en R, la primera categoría alfabética se suele tomar como base, y aquí "George" quedó como referencia del factor).

Las columnas más relevantes son **LagoHancock, LagoOcklawaha y LagoTrafford**. Observando cada categoría por lago, tenemos:

Invertebrados vs Peces

- **Hancock:** $coef = -1.6583 \rightarrow Odds\ Ratio \approx 0.19$. Menor probabilidad de consumir invertebrados que en George.

- **Ocklawaha:** $coef = 0.9372 \rightarrow Odds\ Ratio \approx 2.55$. Más del doble de probabilidad de consumir invertebrados que en George.

- **Trafford:** $coef = 1.1220 \rightarrow OR \approx 3.07$. Tres veces más probabilidad que en George.

Los caimanes de **Ocklawaha y Trafford** tienen clara preferencia por invertebrados frente a peces, mientras que en **Hancock** es menos probable.

Santiago de Cali, Octubre de 2025
Universidad Autónoma de Occidente
Inferencia Estadística
Desafío #3
Alejandro Meneses Portilla.

Reptiles vs Peces

- **Hancock:** $coef = 1.2432 \rightarrow Odds\ Ratio \approx 3.47$. Hay mayor probabilidad que en George.
- **Ocklawaha:** $coef = 2.4588 \rightarrow Odds\ Ratio \approx 11.7$. Muchísima mayor probabilidad que en George.
- **Trafford:** $coef = 2.9353 \rightarrow OR \approx 18.8$. Probabilidad todavía más alta que en George, comparado con Ocklawaha.

Los lagos **Ocklawaha y Trafford** muestran un consumo muchísimo más marcado de **reptiles** frente a peces, aunque **Hancock** también es alto.

Aves vs Peces

- Ninguno de los coeficientes (Hancock 0.695, Ocklawaha -0.653, Trafford 1.088) parece significativo: SE's son grandes (≈0.8–1.2), por lo que no hay evidencia clara de que el lago influya en la elección de aves como alimento primario.

Otros vs Peces

- **Hancock:** $coef = 0.8263 \rightarrow OR \approx 2.28$. Más probabilidad de "otros" que en George.
- **Ocklawaha:** $coef = 0.0055 \rightarrow OR \approx 1.01$. Prácticamente igual que George.
- **Trafford:** $coef = 1.5164 \rightarrow OR \approx 4.56$. Más probabilidad que en George.

Los lagos **Hancock y Trafford** tienen mayor proporción de "otros" frente a peces, mientras que **Ocklawaha** es similar a George.

Conclusiones

1. La **dieta de los caimanes** depende tanto del **lago** como del **tamaño**, aunque los efectos varían según la categoría alimenticia.
2. El hallazgo más claro es que los caimanes grandes casi no consumen **invertebrados** como dieta principal, su probabilidad de consumir invertebrados frente a peces es **mucho menor** que la de los caimanes pequeños.

3. En cambio, para **Aves, Reptiles y Otros**, el tamaño no mostró un efecto concluyente.

4. El **lago** influye considerablemente: las probabilidades relativas de las dietas cambian según el hábitat, en especial para los reptiles.

Situación 3

Para esta situación, se hace un breve estudio de los datos de Enero de 2023 de los taxis amarillos proporcionados por la ***Taxis & Lomousines Commission (TLC) de NY*** (datos disponibles en https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page) con el objetivo de estructurar 3 modelos que expliquen diferentes aspectos de los viajes en taxis:

1. **Modelo 1**: Probabilidad de recibir propina.
2. **Modelo 2**: Número de pasajeros.
3. **Modelo 3**: Monto de la propina condicionado a que sea mayor a cero.

Modelo 1: Modelado de Ocurrencia de Propinas (ver funcion situacion_3("modelo1")**)**

El objetivo es predecir la probabilidad de que un cliente deje propina ($recibio\_propina$). El modelo base es una Regresión Logística.

Forma del Modelo:

$$recibio\_propina \sim trip\_distance + passenger\_count + payment\_type + pickup\_hour + RatecodeID + store\_and\_fwd\_flag$$

Variable dependiente: $recibio\_propina\ (1\ si\ tip\_amount\ >\ 0,\ 0\ de\ lo\ contrario)$

En la tabla de valores del Anexo **1**, se observan los valores del modelo y varios de estos son muy pequeños (**p < 0.001**). Esto significa que muchas variables son **estadísticamente significativas** para explicar la probabilidad de recibir propina.

- **trip_distance**: Esto demuestra poca significancia, por lo cual la distancia no es un factor determinante en la probabilidad de recibir propinas.

- **payment_type 2, 3, 4**: Muestra que el tipo de pago es clave en la probabilidad de dejar propina pues muestra que los **pagos con tarjeta** tienen más probabilidad de registrar propina que pagos en efectivo.

- **pickup_hour (varias categorías)**: casi todas significativas, por lo cual el horario también influye, donde los horarios de recogida entre la **1 PM** y las **4 PM** son bastante significativos, entre las **9 PM** y las **11 PM** son poco significativas, por lo que

la probabilidad de dejar propinas entre las **9 PM** y **11 PM es poca**.

- **RatecodeID**: algunos niveles aparecen significativos, pero el **RatecodeID6** (Viajes en Grupo) no tiene significancia, en este contexto se podría afirmar que la probabilidad de dejar propina tiende a cero cuando se realizan viajes en grupo.

- **store_and_fwd_flagY**: **Significativo**. Al parecer este metadato técnico afecta la probabilidad de propina, se tuvo en cuenta inicialmente dado que es un metadato que confirma si se tuvo que guardar información en local por falta de conexión (zona apartada, que podría implicar viaje largo).

Alternativa a investigar: Utilizando el criterio de Akaike (AIC) sobre un modelo ajustado sobre una submuestra de 10% y otro utilizando ***speedglm***, se obtuvo lo siguiente:

Modelo	AIC
Submuestra 10%	75869.11
Submuestra 50%	75869.11
Speedglm	764459.75

A simple vista es preferible utilizar submuestras (de ser posible) incluso del 10% de los datos, sin embargo, speedglm puede ser una gran alternativa en contextos de big data donde los datasets pueden constar de cientos o miles de millones de registros.

Modelo 2: Número de pasajeros (ver funcion situacion_3(“modelo2”)**)**

- **Forma del Modelo:**

$$passenger\_count \sim trip\_distance + pickup\_hour + PULocationID + DOLocationID$$

- **Estimador base:** Poisson.

- **Chequeo de equidispersión:** Este es un indicador de si los datos siguen una distribución **Poisson**, donde $RD = Varianza\ /\ Media$. En este caso, tenemos un ratio de dispersión de **0.5893**, lo cual es una **Subdispersión** (los conteos son menos variables de lo que podría predecir de **manera óptima** una Poisson).

- **Resultados resumidos**

Variable	Odds Ratio	Interpretación
Intercepto	1.5728621	Valor base del modelo

trip_distance	1.0000508	Por cada unidad adicional de distancia, las probabilidades de tener más de 1 pasajero aumentan un 0.005%. **No es significativo**.
factor(pickup_hour)1	0.9667184	Los viajes iniciados a esta hora tienen una probabilidad ligeramente menor de tener más de 1 pasajero.
factor(pickup_hour)5	1.0341649	Aumenta 3.4% la cantidad de múltiples pasajeros.
factor(pickup_hour)19, 20, 21, 22	1.111, 1.114, 1.123, 1.100	Las horas de recogida entre las 19-22 aumentan en un 10%-12% aprox. la probabilidad de múltiples pasajeros (tiene sentido por after-hours, reuniones, etc).
PULocationID3–265	0.67 – 1.58	Diferencias según el punto de recogida, pero lugares como **PULocationID6** aumentan hasta **58%** la cantidad de pasajeros.
DOLocationID2–265	0.55 – 0.9	Diferencias según el destino, se observa que la mayoría de destinos influyen un poco en la cantidad de pasajeros.

Alternativa a investigar: Se obtuvo un ratio de dispersión de **0.589**, por lo que **no se detectó sobredispersión.**

Modelo 3: Monto de la propina (ver funcion situacion_3(“modelo3”)**)**

- **Forma del Modelo:**

$$tip\_amount \sim trip\_distance + passenger\_count + payment\_type + pickup\_hour$$

- **Estimador base:** Gamma.
- Función de enlace **“inverse”:** No puede converger a la solución.
- **Resultados Resumidos:**

Tipo	Coeficientes	Interpretación
trip_distance	1.035	Por cada unidad adicional de distancia del viaje, el monto promedio de la propina aumenta en aproximadamente un 3.6%.
passenger_count	1.017	Por cada pasajero adicional, el monto promedio de la propina aumenta en aproximadamente un 1.7%.

Ambos factores, la **distancia** y el **número de pasajeros**, tienen un efecto positivo y multiplicativo sobre el monto de la propina. La distancia es el factor más influyente, con un impacto de casi el doble que el conteo de pasajeros.

Factor tipo de pago (payment_type)

Tipo	Coeficientes	Interpretación (vs. Tarjeta de Crédito, Tipo 1)
payment_type2	1.269	Si es Pago Tipo 2 (ej. Efectivo), la propina promedio es un 27% mayor que con Tarjeta de Crédito.
payment_type3	0.394	Si es Pago Tipo 3 (ej. No tip) la propina promedio es un 60.6% menor (1−0.394) que con Tarjeta de Crédito.
payment_type4	1.200	Si es Pago Tipo 4 (Dispute), la propina promedio es un 20% mayor que con Tarjeta de Crédito.

El coeficiente alto para los tipos 2 y 4 (1.27 y 1.20) sobre el tipo 1 (Tarjeta de Crédito) es un poco contra-intuitiva. Tradicionalmente, la tarjeta de crédito es la principal fuente de propinas dado se facilidad de uso y poca fricción de pago.

Factor tiempo (pickup_hour)

Rango de Horas	Coeficientes	Efecto (vs. Hora 0)
1 AM - 12 PM	0.728 a 0.860	Propina promedio menor.
1 PM - 11 PM	0.783 a 0.942	Propina promedio menor, pero acercándose a la base.

Las horas nocturnas suelen asociarse con viajes relacionados con entretenimiento, restaurantes y aeropuertos, donde la propina puede ser más generosa, lo cual es capturado por estos coeficientes más elevados (cercanos a 1).

Santiago de Cali, Octubre de 2025
Universidad Autónoma de Occidente
Inferencia Estadística
Desafío #3
Alejandro Meneses Portilla.

Anexo 1: Resultados Modelo 1

Tabla 1: Coeficientes

Variable	Odds Ratio
(Intercept)	1.275771e+01
trip_distance	9.997256e-01
passenger_count	9.756440e-01
payment_type2	9.646457e-06
payment_type3	5.750749e-04
payment_type4	3.515839e-04
factor(pickup_hour)1	1.529780e+00
factor(pickup_hour)2	2.175817e+00
factor(pickup_hour)3	2.567395e+00
factor(pickup_hour)4	2.742523e+00
factor(pickup_hour)5	2.608123e+00
factor(pickup_hour)6	2.624627e+00
factor(pickup_hour)7	2.526716e+00
factor(pickup_hour)8	2.443633e+00
factor(pickup_hour)9	2.330624e+00
factor(pickup_hour)10	2.329137e+00
factor(pickup_hour)11	2.261326e+00
factor(pickup_hour)12	2.308061e+00
factor(pickup_hour)13	2.435415e+00
factor(pickup_hour)14	2.333776e+00

factor(pickup_hour)15	2.357764e+00
factor(pickup_hour)16	2.489548e+00
factor(pickup_hour)17	2.185597e+00
factor(pickup_hour)18	1.710715e+00
factor(pickup_hour)19	1.464402e+00
factor(pickup_hour)20	1.300286e+00
factor(pickup_hour)21	1.088787e+00
factor(pickup_hour)22	9.900055e-01
factor(pickup_hour)23	9.076477e-01
RatecodeID2	7.179009e-01
RatecodeID3	4.353722e-01
RatecodeID4	2.666845e-01
RatecodeID5	1.489181e-01
RatecodeID6	6.944943e-01
RatecodeID99	8.116170e-06
store_and_fwd_flagY	7.351953e-01

Tabla 2: Coeficientes con Error Estándar y Significancia

Variable	Estimado	Std. Error	z value	Pr(>|z|)	Signif.
(Intercept)	2.546e+00	3.399e-02	74.8984	0.000e+00	***
trip_distance	-2.744e-04	6.185e-05	-4.4373	9.109e-06	***
passenger_count	-2.466e-02	3.694e-03	-6.6742	2.486e-11	***
payment_type2	-1.155e+01	8.396e-02	-137.5496	0.000e+00	***
payment_type3	-7.461e+00	6.169e-02	-120.9487	0.000e+00	***
payment_type4	-7.953e+00	5.793e-02	-137.2784	0.000e+00	***
factor(pickup_hour)1	4.251e-01	4.178e-02	10.1752	2.558e-24	***

factor(pickup_hour)2	7.774e-01	3.920e-02	19.8305	1.624e-87	***
factor(pickup_hour)3	9.429e-01	3.848e-02	24.5015	1.423e-132	***
factor(pickup_hour)4	1.009e+00	3.830e-02	26.3395	6.763e-153	***
factor(pickup_hour)5	9.586e-01	3.780e-02	25.3636	6.355e-142	***
factor(pickup_hour)6	9.649e-01	3.757e-02	25.6842	1.754e-145	***
factor(pickup_hour)7	9.269e-01	3.707e-02	25.0012	5.927e-138	***
factor(pickup_hour)8	8.935e-01	3.679e-02	24.2884	2.597e-130	***
factor(pickup_hour)9	8.461e-01	3.640e-02	23.2464	1.548e-119	***
factor(pickup_hour)10	8.455e-01	3.631e-02	23.2825	6.674e-120	***
factor(pickup_hour)11	8.160e-01	3.622e-02	22.5305	2.087e-112	***
factor(pickup_hour)12	8.364e-01	3.607e-02	23.1901	5.734e-119	***
factor(pickup_hour)13	8.901e-01	3.612e-02	24.6450	4.160e-134	***
factor(pickup_hour)14	8.475e-01	3.629e-02	23.3543	1.245e-120	***
factor(pickup_hour)15	8.577e-01	3.671e-02	23.3669	9.284e-121	***
factor(pickup_hour)16	9.121e-01	3.692e-02	24.7028	9.989e-135	***
factor(pickup_hour)17	7.819e-01	3.685e-02	21.2188	6.405e-100	***
factor(pickup_hour)18	5.369e-01	3.694e-02	14.5358	7.183e-48	***
factor(pickup_hour)19	3.814e-01	3.757e-02	10.1539	3.185e-24	***
factor(pickup_hour)20	2.626e-01	3.859e-02	6.8050	1.011e-11	***
factor(pickup_hour)21	8.506e-02	3.968e-02	2.1435	3.207e-02	*
factor(pickup_hour)22	-1.004e-02	4.207e-02	-0.2387	8.113e-01	
factor(pickup_hour)23	-9.690e-02	4.615e-02	-2.0999	3.574e-02	*
RatecodeID2	-3.314e-01	1.572e-02	-21.0796	1.225e-98	***
RatecodeID3	-8.316e-01	4.645e-02	-17.9026	1.126e-71	***
RatecodeID4	-1.322e+00	5.599e-02	-23.6049	3.433e-123	***
RatecodeID5	-1.904e+00	2.222e-02	-85.6934	0.000e+00	***

RatecodeID6	-3.646e-01	3.300e+00	-0.1105	9.120e-01	
RatecodeID99	-1.172e+01	5.767e-01	-20.3260	7.578e-92	***
store_and_fwd_flagY	-3.076e-01	3.659e-02	-8.4074	4.193e-17	***