# Biometria II
# TP Nº 6
# Modelos Lineales Generalizados Regresion Poisson

Hasta ahora teniamos para los modelos lineales, lo siguiente:

$$Y_i \sim N(\mu, \sigma^2)$$
$$Y_i = g(x_i) + \epsilon_i$$

Con lo cual:

$$E[Y_i] = \mu_i = g(x_i)$$

Sin embargo ahora supondremos:

$$\mu_i = e^{g(x_i)}$$

o

$$log(\mu_i) = g(x_i)$$

Entonces:

$$Y_i \sim P(\mu_i)$$

y

$$E[Y_i/continuo_i] = \mu_i = e^{g(x_i)} = e^{\beta_0 + \beta_1 * X_{i1} + ... + \beta_p * Xip}$$

## Distribucion de Poisson

La distribucion de *Poisson* es muy utilizada para modelar el numero de ocurrencias de un evento en un intervalo de tiempo dado. Por ejemplo, la espera de un colectivo, cantidad de presas encontradas por un predador por mes, o el numero de individuos / $m^2$ tipicamente se pueden modelar con una distribucion de *Poisson*. Uno de los supuestos basicos sobre los que esta distribucion se construye es, que para intervalos de tiempo cortos, la probabiidad de ocurrencia del evento es proporcional a la medida del tiempo esperado. Esta distribucion posee solo un parametro $\lambda$, que es la esperanza. Entonces una *v.a* tendra una distribucion de *Poisson* si:

$$P(X = x|\lambda) = (e^{-\lambda} * \lambda^x)/x!$$

$$E(X) = \lambda$$

$$Var[X] = \lambda$$

Antes de arrancar vamos a inspeccionar un poco como se ve la distribucion de Poisson conforme cambia el valor de $\lambda$. Abra una explorador de internet y vaya al siguiente **https://homepage.divms.uiowa.edu/~mbognar/**.

Explore la *applet* para la distribucion de Poisson y responda:

-Cual es la probabilidad de que X=2 cuando $\lambda$=7?

-Cual es la probabilidad de que X=2 cuando $\lambda$=7?

# Problema 1. Reservas Urbanas y conservación de aves

Las reservas urbanas pueden, entre otros servicios ecosistémicos, brindar refugio para las aves nativas. Esto representa un factor importante en la conservación de la biodiversidad. Con el fin de evaluar este servicio, se llevo a cabo un estudio para cuantificar la diferencia en abundancia de aves nativas entre ambientes antropizados y reservas urbanas con planes de manejo. Para ello, se seleccionaron 35 puntos de muestreo en la Ciudad de Buenos Aires de los cuales 21 correspondieron a ambientes antropizados y 14 a la Reserva Ecologica Costanera Sur. En cada uno de los puntos se realizaron observaciones de aves.

El protocolo original consistio en realizar conteos de aves nativas en cada punto durante 5 minutos, pero dependiendo de las condiciones en las que se daba el registro, la duracion final de las observaciones vario entre 2 y 7 minutos. Por ello para cada conteo se cuenta tambien con el registro del tiempo de observacion (en minutos). Base de datos "*aves.txt*".

- Indique cual es la variable dependiente o respuesta. ¿Cual es su potencial distribucion de probabilidades? ¿Cual es la variable explicatoria? ¿De que tipo es?
- Plantee y escriba el modelo en terminos de regresion.
- Describa grafica y estadisticamente los datos.
- Pruebe los supuestos del modelo. ¿Se verifican?

Chequeamos residuos vs. predichos.

- Valide el modelo.
- En que medida las reservas favorecen la conservacion de aves nativas?. Informe la magnitud del efecto en escala de la variable respuesta.

Notar que el summary nos devuelve: a)Intercept y b) Ambiente Reserva.

- Intercept (tasa para el area metropolitana)

$$\hat{\beta}_0 = log(\widetilde{aves}_{metropolitana})$$

-Ambiente Reserva (efecto reserva sobre metropolitana)

$$\hat{\beta}_1 = log(\widetilde{aves}_{reserva}) - log(\widetilde{aves}_{metropolitana})$$

Es decir, el incremento respecto del ambiente metropolitano.

# Problema 2. Atropellamiento de anfibios en una carretera en las cercanias de un parque natural (Modificado a partir de base de datos de Zuur 2009)

Los datos presentados provienen de un estudio de dos anios sobre vertebrados atropellados en una ruta nacional del sur de Portugal, pavimentada y con trafico moderado. En las cercanias se encuentran ambientes boscosos, tierras abiertas, incluyendo pastos, prados y barbecho.
La ruta fue inspeccionada cada dos semanas por dos anios. Se identifico a cada animal encontrado muerto a nivel de especie, siempre que hubiese sido posible, y se registro su ubicacion geografica en coordenadas. Para fines del analisis de datos, la ruta se dividio en 52 segmentos de 500 mts y se presenta el numero total de anfibios muertos por segmento ("TOT.N"). Los datos se encuentran en la base *roadkills.txt*.
En cada punto se registraron variables ambientales (ver tabla). En particular se desea conocer si el numero de anifibios atropellados (TOT.N) se encuentra relacionado con las variables OPEN.L, MONT.S, POLIC, SHRUB, WAT.RES, L.WAT.C, L.P.ROAD, D.WAT.COUR y D.PARK. No se tienen evidencias de que haya interaccion entre las variables.

- Realice un analisis exploratorio de las variables involucradas. ¿Detecta datos atipicos? ¿Como son las relaciones entre las variables explicatorias? ¿Y entre las explicatorias y la respuesta?
- Se decide aplicar raiz cuadrada sobre las variables POLIC, SHRUB, WAT.RES, L.P.ROAD, D.WAT.COUR. ¿Por que cree que se realizo este procedimiento?

Se aplica la raiz cuadrada debido a los altos valores que toman las variables.

- Plantee el modelo adecuado.

- Calcule la sobredispersion del modelo aditivo completo.
- ¿Es correcto modelar suponiendo distribucion de Poisson? ¿Por que? Si considera que no es correcto utilizar la distribucion *Poisson*, ¿que alternativas de analisis propone?.

## A que llamamos sobredispersion o subdispersion? Como se ve?

Supongamos que estamos modelando la disposicion en un determinado lugar de individuos de una especie. Si bien esta disposicion de los individuos puede ser modelada usando una distribucion de Poisson, algunos individuos podrían **agruparse** para ocupar el terreno (dispersion por contagio) o **alejarse** entre si lo más posible (dispersion por rechazo). Estos casos no pueden ser adecuadamente modelados por la distribucion de *Poisson* porque se viola el supuesto de la independencia entre los eventos. En estos casos la variabilidad se aleja de la media. El coeficiente de dispersion (CD = varianza/media) permite identificar estas situaciones. Si los individuos se agrupan siguiendo un patron de contagio, la variabilidad aumenta mas alla de la explicada por la distribucion de *Poisson* (sobredispersion). En cambio, si los individuos se agrupan siguiendo un patron de rechazo, la variabilidad estara por debajo de la explicada por la distribucion de *Poisson* (subdispersion).

## Time out para un ejemplo

- Se sabe que la cantidad de hojas de una plantula a los 30 dias post-germinacion tiene una media de 8.56 hojas. Corra el script.
- ¿Que media y que varianza esperaria obtener? ¿Es razonable suponer una distribucion *Poisson*?

Una forma efectiva de modelar datos de conteo con sobredispersión es usando la distribución binomial negativa.

## Que ventaja tiene la distribucion Binomial Negativa?

La distribucion *Binomial* cuenta el numero de exitos en un numero prefijado de ensayos de *Bernoulli*. Supongamos, en cambio, que contamos el numero de ensayos de *Bernoulli* requeridos para conseguir un numero prefijado de exitos. Esta formulacion nos anticipa la distribucion *Binomial Negativa*.
Entonces, en una secuencia de ensayos independientes de *Bernoulli* $(p)$, sea la *v.a.* X, que denota el ensayo para el cual el r-esimo exito ocurre, donde r es un entero prefijado. Entonces:

$$P(X = r|r, p) = \binom{x-1}{r-1} * p^r * (1-p)^{x-r} \qquad (1)$$

donde (1) tiene distribucion *Binomial Negativa* con parametros $r$ y $p$.

La obtencion de esta formula se entiende rapidamente de la distribucion binomial. El evento {X=x} puede ocurrir solamente si hay exactamente r-1 exitos en los primeros x-1 ensayos y un exito en el ensayo x. La probabilidad de r-1 exitos en x-1 ensayos es la probabilidad binomial:

$$\binom{x-1}{r-1} * p^{r-1} * (1-p)^{x-r}$$

y con probabilidad p hay un exito en el ensayo x. Multiplicando estas probabilidades se llega a (1).

La distribución *Binomial Negativa* esta relacionada con la distribucion de *Poisson* y puede parametrizarse usando $\mu$ y *alpha* ($\alpha$>0). La media de la distribucion *Binomial Negativa* es $\mu$ y su varianza $\mu + \alpha^*\mu^2$. De esta manera si $\alpha = 0$, la distribucion Binomial Negativa es igual a la Poisson.

Ahora vamos a caracterizar a la distribucion *Binomial Negativa* para entender como influyen los diferentes parametros sobre ella.

En R tenemos que mu=media y varianza=mu+$(mu^2)$/size

Por otra parte, es conveniente visualizar que relacion guarda la *esperanza* con la *varianza*.

## Seguimos con el ejercicio

- Realice el modelo utilizando la distribucion Binomial negativa. Aplique un metodo de seleccion de modelos de manera de incorporar variables de a una. Utilice AIC para decidir la incusion o no de una variable en el modelo.

Se puede chequear que el modelo elegido tiene todos los factores importantes mediante *drop1*.

- Evalue los supuestos del modelo seleccionado.

Chequeamos colinealidad entre las variables.

Chequeamos residuos vs. predichos.

- Concluya en relacion a las condiciones que favorecen el atropellamiento de anfibios.
- Para saber mas: Zuur (2009), plantea este caso y realiza diversos analisis con distinto nivel de complejidad, debido a la naturaleza de los datos y las relaciones entre las variables; ademas compara distintas formas de modelar. Es un libro de cabecera, aproveche este ejercicio para conseguirlo, enriquecer el contexto y aprender sobre otras estrategias de análisis.

## Problema 3. Riqueza de especies de anfibios en la superficie central de América del Sur.

En Base a los datos correspondientes al gradiente latitudinal de diversidad de aves anfibios y las variables ambientales relacionadas a las hipotesis planteadas (18 variables totales, ver enunciados de guias anteriores), se quiere responder:

- ¿La riqueza de especies de anfibios (Amphibian.Richness) responde a la temperatura y a la precipitacion en la porcion central de Ameérica del Sur? (Annual.Mean.Temperature y Mean.Annual.Precipitation) ¿Cómo es esa relacion?.
- Indique cual es la variable dependiente o respuesta. ¿Cual es su potencial distribucion de probabilidades? ¿Cales son las variables explicatorias? ¿De que tipo son?.
- Explore como es la relacion entre las variables.
- Plantee los modelos aditivos y con interaccion.
- ¿Que modelo le parece que describe mejor a la Riqueza de especies?.
- Ajuste el modelo.
- Valide el modelo.

## Problema 4. Efecto del herbicida glifosato sobre la fecundidad de aranias

Las aranias son depredadores importantes de varias plagas agricolas y desempenian un papel importante como indicadores de disturbio del ecosistema. En Argentina, el cultivo de soja ha aumentado desde la introduccion de la soja transgenica resistente al glifosato. Esta expansion produjo un aumento en el uso del glifosato, un herbicida de amplio espectro, cuyo efecto sobre la fisiologia de los artropodos es muy poco conocida. En los cultivos transgenicos de soja de la provincia de Buenos Aires, *Alpaida veniliae* (Araneae, Araneidae) es una de las aranias tejedoras mas abundantes. El proposito de este trabajo fue estudiar los efectos del glifosato sobre algunos atributos reproductivos de *A. veniliae*, en laboratorio. Para ello, hembras fecundadas fueron criadas en frascos de vidrio individuales y alimentadas con moscas, que previamente habian sido tratadas con dosis distintas de glifosato (2) o con solvente. Se utilizaron 10 hembras por tratamiento. Para cada hembra se registro la fecundidad (numero de huevos) y la fertilidad (numero de crias). No se observo un efecto letal del glifosato. Los resultados se encuentran en el archivo *arania.csv*.

- Indique cual es la variable dependiente o respuesta. ¿Cual es su potencial distribucion de probabilidades? ¿Cual es la variable explicatoria? ¿De que tipo es?.
- Describa grafica y estadisticamente los datos.
- Plantee el modelo.
- Pruebe los supuestos del modelo. ¿Se verifican?.
- Ajuste el modelo. Calcule las predicciones en escala de la variable respuesta para los tres tratamientos.
- Valide el modelo.
- Compare los tres tratamientos y concluya en relacion al efecto del glifosato sobre la fecundidad de *A. veniliae*.

*Basado en Benamú, M. A., Schneider, M. I., & Sánchez, N. E. (2010). Effects of the herbicide glyphosate on biological attributes of Alpaida veniliae (Araneae, Araneidae), in laboratory. Chemosphere, 78(7), 871-876.*

## Problema 5.

En base al problema del atropellamiento de anfibios presentado, se quiere modelar la Riqueza de anfibios ("S.RICH") en funcion del porcentaje de espacio abierto ("OPEN.L").

- Plantee el modelo teorico en parametros y en el contexto de la experiencia.
- Realice un procedimiento adecuado para modelar los datos.
- ¿Como se modifica la riqueza de anfibios en funcion del espacio abierto?.