Un modelo con distribución binomial

Ahora trabajaremos con estos datos de 'Weevil_damage.csv' que puedes importar desde el repositorio. Podemos examinar si el daño al pino silvestre por los gorgojos (una variable binaria, 1/0) varía según el bloque en el que se encuentran los árboles. ¿Puedes imaginar que diferentes bloques representan diferentes poblaciones de pino silvestre, y quizás algunos de ellos serán particularmente vulnerables a los gorgojos? Debido a la naturaleza binaria de la variable de respuesta (verdadero o falso), aquí es apropiado un modelo binomial.

```
Weevil_damage <- read.csv("Weevil_damage.csv")

# Haz el bloque un factor

# Corre el modelo

weevil.m <- glm(damage_T_F ~ block, family = binomial, data = Weevil_damage)
summary(weevil.m)
```

Echa un vistazo a la salida de resumen. Parece que la probabilidad de que un pino sufra daños por gorgojos varía significativamente según el bloque en el que se encuentra el árbol. Las estimaciones que ves no son tan sencillas de interpretar como las de los modelos lineales, donde la estimación representa el cambio en Y por un cambio en 1 unidad de X, porque los modelos binomiales son un tipo de regresión logística que se basa en relaciones impares logarítmicas, pero No entraremos en detalles aquí. Las estimaciones más grandes aún significan una mayor influencia de sus variables, ¡solo ten en cuenta que no es una relación lineal! Y por tanto, no obtendrás un valor de R cuadrado para evaluar la bondad del ajuste del modelo, pero puedes obtenerlo observando la diferencia entre la desviación nula (variabilidad explicada por un modelo nulo, por ejemplo, glm (daños_T_F ~ 1)) y la desviación residual, por ejemplo la cantidad de variabilidad que queda después de haber explicado algo con su variable explicativa. En resumen, cuanto mayor sea la reducción en la desviación, mejor será el trabajo de su modelo para explicar una relación.

Otro ejemplo desmenuzado

Vamos a trabajar con el conjunto de datos de cobertura vegetal a largo plazo de 'Toolik Lake Field Station'. Estos datos son de composición de plantas recopilados durante cuatro años en cinco sitios a lo largo del tiempo en la tundra ártica en el norte de Alaska. Una pregunta simple que podríamos hacernos con estos datos es: ¿cómo ha cambiado la riqueza de especies en estas parcelas a lo largo del tiempo?

Pregunta 1: ¿Cómo ha cambiado la riqueza de especies de plantas con el tiempo en el lago Toolik?

Una vez que hemos descifrado nuestra pregunta de investigación, tenemos que descifrar nuestra hipótesis. Para llegar a una hipótesis, necesitamos aprender algo sobre este sistema.

Asumamos la siguiente hipótesis: la riqueza de especies de plantas está aumentando con el tiempo. Podríamos esperar esto, ya que estas parcelas de tundra podrían estar experimentando un calentamiento y el calentamiento podría conducir a una mayor riqueza de especies de plantas en las comunidades de plantas de tundra.

Hipótesis 1: La riqueza de especies de plantas ha aumentado con el tiempo en el lago Toolik.

Ahora que tenemos una hipótesis, es una buena práctica escribir también una hipótesis nula. ¿Cuáles son las hipótesis que estamos probando entre aquí? Por ejemplo, una hipótesis nula para estos datos y esta pregunta podría ser:

Hipótesis nula: La riqueza de especies de plantas no ha cambiado con el tiempo en el lago Toolik.

También podríamos tener una hipótesis alternativa:

Hipótesis 2: La riqueza de especies de plantas ha disminuido con el tiempo en el lago Toolik.

Toolik Lake Station está en Alaska, un lugar que se ha estado calentando a un ritmo mayor que el resto del mundo, por lo que también podríamos preguntarnos cómo influye la temperatura en las comunidades de plantas allí, en particular, en su riqueza. Entonces, planteamos una segunda pregunta: Pregunta 2: ¿Cómo influye la temperatura media anual en la riqueza de especies de plantas?

Hipótesis 1: Las temperaturas más altas se corresponden con una mayor riqueza de especies.

¿En qué se diferencian las preguntas 1 y 2?

Modelos de detección: Cuando preguntamos cómo ha cambiado la riqueza de especies de plantas a lo largo del tiempo, estamos interesados en detectar el cambio. Queremos saber qué sucedió con las comunidades de plantas en Toolik Lake, pero no estamos probando nada sobre por qué ocurrieron tales cambios en la riqueza de especies (y tal vez no hubo cambios con el tiempo).

Modelos de atribución: Cuando preguntamos cómo la temperatura influye en la riqueza de especies de plantas, buscamos atribuir los cambios que hemos visto a un factor específico, en este caso, la temperatura. Los modelos de atribución a menudo son el siguiente paso de un modelo de detección. Primero, queremos saber qué sucedió, luego tratamos de averiguar por qué sucedió. Por ejemplo, si encontramos una fuerte relación positiva entre la temperatura y la riqueza de especies (p. ej., a medida que aumenta la temperatura, también aumenta la riqueza de especies), es probable que la temperatura sea uno de los impulsores de los cambios a escala local en la riqueza de especies.

Por ahora, esto debería ser suficiente configuración para que podamos avanzar con nuestros modelos, pero recuerda siempre comenzar con la pregunta primero cuando realices cualquier proyecto de investigación y análisis estadístico.

Pensando en nuestros datos

Hay diferentes pruebas estadísticas que podríamos usar para realizar nuestros análisis y el tipo de prueba estadística que usamos depende de la pregunta y el tipo de datos que tenemos para probar nuestra pregunta de investigación. Como ya hemos pensado un poco en nuestra pregunta, pensemos ahora en nuestros datos. ¿Qué tipo de datos estamos tratando aquí?

Nuestros datos consisten en la cobertura de especies de plantas medida a lo largo de cuatro años

```
library(tidyverse)
library(lme4)  # modelos jerárquicos
library(sjPlot)  # visualizar las salidas de los modelos
library(glmmTMB) # Modelos lineales jeraquizados

toolik_plants <- read.csv("toolik_plants.csv")
head(toolik_plants)
str(toolik_plants)
```

Site y Species son del tipo de carácter (texto, compuesto de letras), son nombres y los trataremos como variables categóricas. Year, Cover, Mean.Temp y SD.Temp son datos numéricos y continuos, son números. La cobertura muestra la cobertura relativa (de 1) para diferentes especies de plantas, Mean.Temp es la temperatura media anual en Toolik Lake Station y SD.Temp es la desviación estándar de la temperatura media anual. Luego, tenemos Tratamiento, otra variable categórica que se refiere a diferentes tratamientos químicos, p.ej. algunas parcelas recibieron nitrógeno extra, otras fósforo extra. Finalmente, tenemos Block y Plot, que brindan información más detallada sobre dónde se tomaron las medidas.

Los números de las parcelas están actualmente codificados como números (num) - 1, 2,…8, convirtiéndolo en una variable numérica. Deberíamos convertirlos en una variable categórica, ya que al igual que Sitio y Bloque, los números representan las diferentes categorías, no los datos de conteo reales.

En R, podemos usar el tipo de factor para denotar un vector/columna como datos categóricos.

```
# Un truco nuevo:
toolik_plants <-
  toolik_plants %>%
  mutate(across(c(Site, Block, Plot), as.factor))

str(toolik_plants)
```

Ahora, pensemos en las distribuciones de los datos. Nuestra estructura de datos es un poco como una muñeca rusa, así que comencemos a analizar capa por capa.

```r
unique(toolik_plants$Site)

length(unique(toolik_plants$Site))
```

tenemos 6 sitios: (06MAT, DH, MAT, MNT and SAG).

```r
# Agrupemos por sitio

toolik_plants %>% group_by(Site) %>%
  summarise(block.n = length(unique(Block)))
```

Dentro de cada sitio, hay diferentes números de bloques: algunos sitios tienen tres bloques de muestra, otros tienen cuatro o cinco.

```r
toolik_plants %>% group_by(Block) %>%
  summarise(plot.n = length(unique(Plot)))
```

Dentro de cada bloque, hay ocho parcelas más pequeñas.

```r
unique(toolik_plants$Year)
```

Hay cuatro años de datos desde 2008 hasta 2012.

¿Cuántas especies están representadas en este conjunto de datos? Usemos algo de código para resolver esto. Usando las funciones únicas y de longitud, podemos contar cuántas especies hay en el conjunto de datos como un todo.

```r
length(unique(toolik_plants$Species))
```

Hay 129 especies diferentes, pero ¿son todas realmente especies? Siempre es una buena idea ver qué se esconde detrás de los números, así podemos imprimir las especies para ver qué tipo de especies son.

```r
unique(toolik_plants$Species)
```

Algunas categorías de plantas son solo musgo y liquen y pueden ser especies diferentes o más de una especie, pero para los fines del tutorial, podemos contarlas como una sola especie. Sin embargo, hay otros registros que definitivamente no son especies: basura, desnudo (refiriéndose al suelo desnudo), cubierta de madera, tubo, agujero, rastro de campañol, removido, excremento de campañol, hongos, agua, caca de caribú, rocas, hongos, caca de caribú, arena para animales, caca de campañol, caca de campañol, Unk?.

Filtraremos los registros que no necesitamos usando la función de filtro del paquete dplyr.

```r
toolik_plants <- toolik_plants %>%
  filter(!Species %in% c("Woody cover", "Tube",
                         "Hole", "Vole trail",
                         "removed", "vole turds",
                         "Mushrooms", "Water",
                         "Caribou poop", "Rocks",
                         "mushroom", "caribou poop",
```

```
                          "animal litter", "vole poop",
                          "Vole poop", "Unk?"))
```

Veamos cuántas especies tenemos ahora:

```
length(unique(toolik_plants$Species))
```

115 especies. A continuación, podemos calcular cuántas especies se registraron en cada parcela en cada año de muestreo.

```
# Calcular riqueza
toolik_plants <- toolik_plants %>%
  group_by(Year, Site, Block, Plot) %>%
  mutate(Richness = length(unique(Species))) %>%
  ungroup()
```

Para explorar más los datos, podemos hacer un histograma de riqueza de especies.

```
(hist <- ggplot(toolik_plants, aes(x = Richness)) +
    geom_histogram() +
    theme_classic())
```

Hay algunas otras cosas en las que deberíamos pensar. Hay diferentes tipos de datos numéricos aquí. Por ejemplo, los años son números enteros: no podemos tener el año 2000.5.

La cobertura vegetal puede ser cualquier valor que sea positivo, por lo tanto, debe estar entre 0 y 1. Esto lo podemos ver cuando hacemos un histograma de los datos:

```
(hist2 <- ggplot(toolik_plants, aes(x = Relative.Cover)) +
    geom_histogram() +
    theme_classic())
```

Pensando en nuestro diseño experimental

En el conjunto de datos Toolik de cobertura vegetal, tenemos replicación espacial y temporal. La replicación espacial está en tres niveles diferentes: hay múltiples sitios, que tienen múltiples bloques dentro de ellos y cada bloque tiene ocho parcelas. La replicación temporal se refiere a los diferentes años en los que se registró la cobertura vegetal: cuatro años.

¿Qué otros tipos de problemas podríamos tener que considerar?

Autocorrelación espacial

Una de las suposiciones de un modelo es que los puntos de datos son independientes. En realidad, ese es muy raramente el caso. Por ejemplo, las parcelas que están más cerca unas de otras pueden ser más similares, lo que puede o no estar relacionado con algunos de los controladores que estamos probando, p.ej. temperatura.

Autocorrelación temporal

Del mismo modo, es posible que los puntos de datos de un año no sean independientes de los del año anterior. Por ejemplo, si una especie era más abundante en el año 2000, eso también influirá en su abundancia en 2001.

Convierte una pregunta en un modelo

Volvamos a nuestra pregunta original: Pregunta 1: ¿Cómo ha cambiado la riqueza de especies de plantas con el tiempo en el lago Toolik?

¿Cuál es nuestra variable dependiente e independiente aquí? Podríamos escribir nuestro modelo base en palabras:

La riqueza es una función del tiempo.

En R, esto se convierte en el código: riqueza ~ tiempo.

La riqueza es nuestra variable dependiente (respuesta) y el tiempo es nuestra variable independiente (predictora). Este es nuestro modelo base, pero ¿qué otras cosas debemos tener en cuenta? ¿Qué pasaría si solo modeláramos la riqueza como una función del tiempo sin tratar con la otra estructura en nuestros datos? Vamos a averiguarlo.

Conoce los diferentes tipos de modelos

Antes de volver a nuestro conjunto de datos para el que estamos diseñando un modelo, revisemos algunos conceptos básicos de estadísticas.

Aquí hay algunas preguntas a considerar.

```
¿Cuál es la diferencia entre una variable continua y categórica en un modelo
lineal?
¿Cuántas variables puede tener un modelo?
¿Es mejor tener un modelo con cinco variables o un modelo por variable? ¿Cuán
do elegimos variables?
¿Qué es un efecto fijo? ¿Qué es un efecto aleatorio?
¿Cuál es el resultado más importante de la salida de un modelo?
¿Por qué importa qué tipo de modelos usamos?
```

Modelos lineales generales

Modelo sin efectos aleatorios:

```
plant_m <- lm(Richness ~ I(Year-2007), data = toolik_plants)
summary(plant_m)
```

Observa cómo hemos transformado la columna Año: I (Año - 2007) significa que el año 2008 se convertirá en el Año 1; luego, su modelo estima la riqueza en el primer, segundo, etc., año del período del estudio. De lo contrario, si hubiéramos mantenido los años como 2008, 2009,…, el modelo habría estimado la riqueza muy atrás en el pasado, comenzando desde el Año 1, Año 2… Año 1550 hasta 2012. Esto haría que la magnitud de las estimaciones fuera equivocada.

Puedes experimentar para ver qué sucede si solo agregamos el Año: ¡de repente, la pendiente del cambio de especies sube a los cientos!

Suposiciones hechas:

```
Los datos se distribuyen normalmente.
Los puntos de datos son independientes entre sí.
La relación entre las variables que estamos estudiando es en realidad lineal.
```

¿Crees que se cumplen los supuestos de un modelo lineal general para nuestras preguntas y datos? ¡Probablemente no!

A partir de los histogramas, podemos ver que los datos no se distribuyen normalmente y, además, si pensamos en cuáles son los datos, son recuentos de números enteros (número de especies), probablemente un poco sesgados hacia la izquierda. Por estas razones, podría ser adecuada una distribución de Poisson, no una normal.

Sabemos que debido a cómo se configuró el diseño experimental (parcelas dentro de bloques dentro de sitios), los puntos de datos no son independientes entre sí. Si no tomamos en cuenta los efectos de parcela, bloque y nivel de sitio, estamos ignorando por completo la estructura jerárquica de nuestros datos, lo que podría conducir a inferencias incorrectas basadas en resultados de modelos incorrectos.

¿Qué es la convergencia de modelos?

La convergencia del modelo es si el modelo ha funcionado o no, si ha estimado su variable de respuesta (y los efectos aleatorios, los vamos a ver a continuación), básicamente si las matemáticas subyacentes han funcionado o si se han "roto" de alguna manera. Cuando ajustamos modelos más complicados, estamos empujando los límites de las matemáticas subyacentes y las cosas pueden salir mal, por lo que es importante verificar que el modelo realmente funcionó y que las estimaciones que estamos haciendo tienen sentido en el contexto de los datos sin procesar y la pregunta que estamos haciendo/las hipótesis que estamos probando.

La verificación de la convergencia del modelo se puede realizar en diferentes niveles. Con los modelos paramétricos, una buena práctica es verificar las gráficas residuales versus las predichas.

Por ahora, revisemos la gráfica residual versus predicha para nuestro modelo lineal. Mediante el uso de la función 'plot()', podemos trazar los residuos frente a los valores ajustados, un gráfico QQ de residuos estandarizados, un gráfico de ubicación de escala (raíces cuadradas de residuos estandarizados frente a valores ajustados) y un gráfico de residuos frente al apalancamiento que suma bandas correspondientes a las distancias de Cook de 0.5 y 1. Mirar estos gráficos puede ayudar a identificar cualquier valor atípico que tenga una gran influencia y confirmar que el modelo realmente se ha ejecutado, por ejemplo quieres que los puntos de datos en el gráfico Q-Q sigan la línea uno a uno.

```
plot(plant_m
```

Modelos jerárquicos usando lme4

```
library(ggeffects) # Conectar los resultados del modelo con las gráficas
```

Ahora que hemos explorado la idea de un modelo jerárquico, veamos cómo cambia nuestro análisis si incorporamos o no elementos del diseño experimental a la jerarquía de nuestro modelo.

Primero, modelemos con un solo sitio como un efecto aleatorio. Este modelo no incorpora la replicación temporal en los datos ni el hecho de que existen parcelas dentro de bloques dentro de esos sitios.

```
plant_m_plot <- lmer(Richness ~ I(Year-2007) + (1|Site), data = toolik_plants
)
summary(plant_m_plot)
```

En los resultados de salida puedes ver los tamaños del efecto (en la columna "Estimación" en la parte "Efectos fijos" del resumen), un elemento clave de las salidas del modelo. Los tamaños del efecto nos informan sobre las fortalezas de las relaciones que estamos probando. En este modelo, la variable “Año” tiene un efecto de alrededor de -0.7 sobre la “Riqueza”, lo que puede interpretarse como una disminución anual de 0.7 especies.

Sin embargo, todavía no estamos tomando en cuenta las diferentes parcelas y bloques, así que agreguémoslos gradualmente y veamos cómo cambian los resultados.

```
plant_m_plot2 <- lmer(Richness ~ I(Year-2007) + (1|Site/Block), data = toolik
_plants)
summary(plant_m_plot2)
```

¿Han cambiado las estimaciones de los tamaños del efecto?

```
plant_m_plot3 <- lmer(Richness ~ I(Year-2007) + (1|Site/Block/Plot), data = t
oolik_plants)
summary(plant_m_plot3)
```

Este modelo final responde a nuestra pregunta sobre cómo ha cambiado la riqueza de especies de plantas con el tiempo, al mismo tiempo que explica la estructura jerárquica de los datos.

Revisemos el gráfico de "residuos ajustados vs" de nuestro modelo.

```
plot(plant_m_plot3)
```

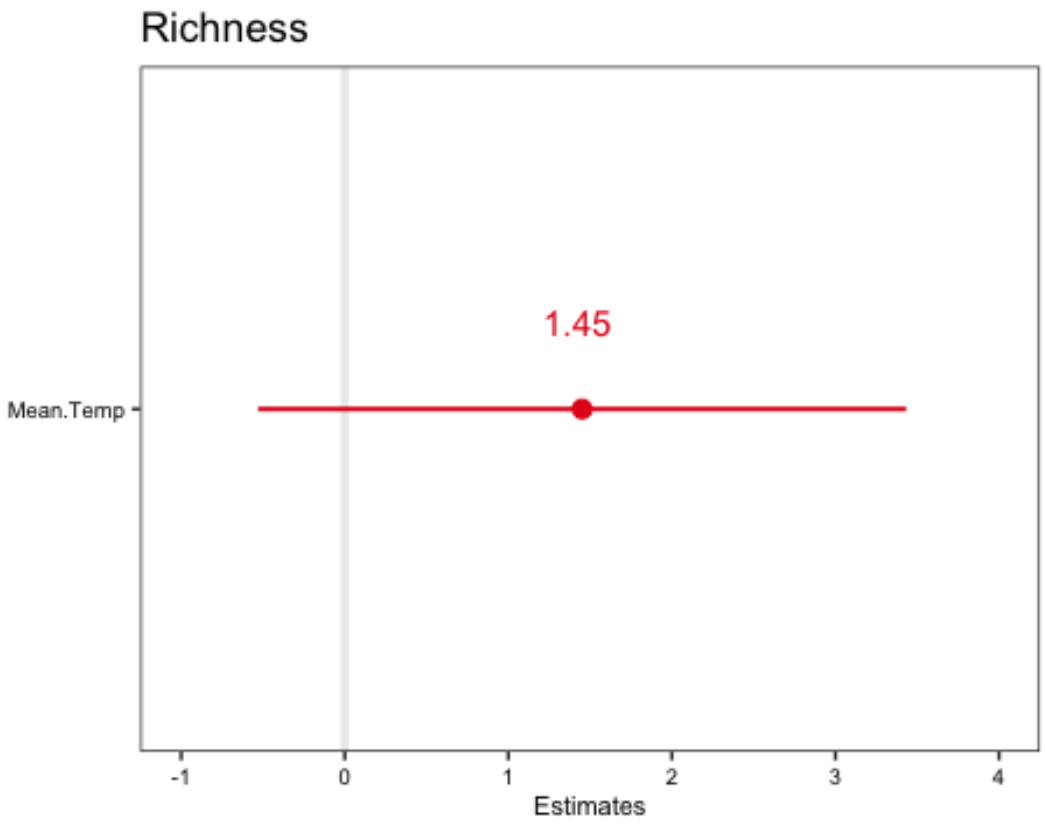
Los puntos de este gráfico están distribuidos uniformemente a ambos lados de la línea horizontal, lo que es una buena señal de que los residuos del modelo no violan los supuestos de los modelos lineales.

Ahora, veamos el efecto de la temperatura media sobre la riqueza. Usaremos la misma estructura jerárquica para efectos aleatorios de sitio/bloque/parcela. Esta vez también agregaremos el año como efecto aleatorio.

```
plant_m_temp <- lmer(Richness ~ Mean.Temp + (1|Site/Block/Plot) + (1|Year),
                     data = toolik_plants)
summary(plant_m_temp)
```

Veamos primero el efecto fijo esta vez:

```
#
(temp.fe.effects <- plot_model(plant_m_temp, show.values = TRUE))
```



El intervalo de confianza muy amplio esta vez sugiere una gran incertidumbre sobre el efecto de la temperatura en la riqueza.

Tamaño del efecto del efecto aleatorio del año Suposiciones hechas:

```
 Los datos se distribuyen normalmente.
 Los puntos de datos son independientes entre sí.
 La relación entre las variables que estamos estudiando es en realidad lineal
 Las parcelas representan la replicación espacial y los años representan la
replicación temporal en nuestros datos.
```

Supuestos no contabilizados:

```
 No hemos tenido en cuenta la autocorrelación espacial en los datos, es
decir, si es más probable que las parcelas ubicadas más cerca muestren respue
stas similares que las parcelas más alejadas.
 No hemos tenido en cuenta la autocorrelación temporal en los datos, es
decir, si la influencia de años anteriores de datos influye en los datos de
un año determinado.
```

Pendientes aleatorias versus intersecciones aleatorias lme4

Ahora podemos pensar en tener pendientes aleatorias e intersecciones aleatorias. Para nuestra pregunta, ¿cómo influye la temperatura en la riqueza de especies?, podemos permitir que cada parcela tenga su propia relación con la temperatura.

```
plant_m_rs <- lmer(Richness ~ Mean.Temp + (Mean.Temp|Site/Block/Plot) + (1|Ye
ar),
                   data = toolik_plants)

summary(plant_m_rs)
```

Consulta los resultados de resumen y los mensajes que recibimos. Este modelo no converge y no debemos confiar en sus resultados: la estructura del modelo es demasiado complicada para los datos subyacentes, por lo que ahora podemos simplificarla.

Si el código se está ejecutando durante un tiempo, no dudes en hacer clic en el botón "Detener" y continuar con el tutorial, ya que el modelo no convergerá.

```
plant_m_rs <- lmer(Richness ~ Mean.Temp + (Mean.Temp|Site) + (1|Year),
                   data = toolik_plants)
summary(plant_m_rs)
```

Esta vez, el modelo converge, pero ten en cuenta que estamos ignorando la estructura jerárquica debajo de "Sitio" y, por lo tanto, violando la suposición sobre puntos de datos independientes (los datos debajo del nivel "Sitio" en realidad están agrupados). Pero lo usaremos para mostrar cómo son los modelos de pendientes aleatorias.

Para tener una mejor idea de lo que están haciendo las pendientes aleatorias y las intersecciones, podemos visualizar las predicciones del modelo. Usaremos el paquete ggeffects para calcular las predicciones del modelo y trazarlas. Primero, calculamos las predicciones generales para la relación entre la riqueza de especies y la temperatura. Luego, calculamos las predicciones para cada parcela, visualizando así la variación entre parcelas. Ten en cuenta que el segundo gráfico tiene pendientes e intersecciones que varían libremente (es decir, son diferentes para cada gráfico).

```
ggpredict(plant_m_rs, terms = c("Mean.Temp")) %>% plot()
```

```
ggpredict(plant_m_rs, terms = c("Mean.Temp", "Site"), type = "re") %>% plot()
+
  theme(legend.position = "bottom")
```

Una nota importante sobre los gráficos honestos

Curiosamente, las opciones predeterminadas de la función ggpredict() establecen la escala de manera diferente para los ejes y en las dos gráficas. Si solo ves la primera gráfica, a primera vista pensarías que la riqueza de especies aumenta mucho a medida que aumenta la temperatura. Pero toma nota del eje y: en realidad no comienza en cero, por lo que se muestra que la relación es mucho más fuerte de lo que realmente es.

Podemos trazar manualmente las predicciones para superar este problema.

```
predictions <- ggpredict(plant_m_rs, terms = c("Mean.Temp"))

(pred_plot1 <- ggplot(predictions, aes(x, predicted)) +
    geom_line() +
    geom_ribbon(aes(ymin = conf.low, ymax = conf.high), alpha = .1) +
    scale_y_continuous(limits = c(0, 35)) +
    labs(x = "\nMean annual temperature", y = "Predicted species richness\n")
)
```

¡La relación entre la temperatura y la riqueza de especies ya no parece tan fuerte! De hecho, vemos aumentos bastante pequeños en la riqueza de especies a medida que aumenta la temperatura. ¿Qué te dice eso sobre nuestra hipótesis?

Ahora podemos hacer lo mismo, pero esta vez teniendo en cuenta el efecto aleatorio.

```
# Predicciones para cada grupo (cada parcela es un factor aleatorio)
# re = efecto aleatorio

predictions_rs_ri <- ggpredict(plant_m_rs, terms = c("Mean.Temp", "Site"), ty
pe = "re")

(pred_plot2 <- ggplot(predictions_rs_ri, aes(x = x, y = predicted, colour = g
roup)) +
    stat_smooth(method = "lm", se = FALSE)  +
    scale_y_continuous(limits = c(0, 35)) +
    theme(legend.position = "bottom") +
    labs(x = "\nMean annual temperature", y = "Predicted species richness\n")
)

(pred_plot3 <- ggplot(predictions_rs_ri, aes(x = x, y = predicted, colour = g
roup)) +
    stat_smooth(method = "lm", se = FALSE)  +
    theme(legend.position = "bottom") +
    labs(x = "\nMean annual temperature", y = "Predicted species richness\n")
)
```