Biodiversity patterns and their causes in mammals

Science School in Quantitative Ecology

Antonin Machac

Biodiversity has fascinated biologists ever since Humboldt, Darwin and Wallace began exploring and doumenting the natural world. One of the most striking features of biodiversity is how unevenly it is distributed across geographic space. Understanding the causes of these large-scale patterns is one of the central endeavours of biology.

The most famous of these patterns is the **latitudinal diversity gradient** (LDG): species richness peaks in the tropics and declines towards the poles. Knowledge of this gradient, and the mechanisms behind it, is essential for ecology, conservation and predicting how biodiversity will respond to global changes. Despite decades of progress, however, we still do not have a compelling explanation. Instead, multiple theories have been proposed, each invoking different mechanisms generating the gradient, from which we can derive diagnostic predictions about the shape of the richness–environment relationships.

Mammals are an ideal organismal group for adressing this challenge. They are globally distributed, well surveyed, ecologically diverse and economically highly important. Analyzing their species richness across grid cells worldwide allows us to confront competing theories with empirical data.

Today, we will explore the diversity patterns of mammals, using R exercises, to demonstrate key concepts in biodiversity research and model comparison (AIC, Akaike's weights, p-value). First, we will upload the necessary libraries (ape, geiger, etc.).

```
library (terra)
library(rasterVis)
library(dplyr)
library(ggplot2)
library(patchwork)
```

Now, lets upload and explore the data. We will be using a data frame, which stores information on the number of species, latitude, environmental temperature, precipitation and productivity (NPP) within 1x1 degree grid cells covering the world. Each row of the data frame, therefore, is a grid cell, for which we have species richness and its environmental predictors.

The data frame will be uploaded from your working directory. Working directory is a folder on your computer where you have stored the data frame (called "mammal_data.txt"). This folder should also include the map template (file called "bio01_1degr.asc"). If you are unsure where to find your working directory folder, type "getwd()" into your R console.

```
mamm <- read.delim("mammal_data.txt")
```

```
summary (mamm)
```

```
##      sprich           latit            temp             prec
##  Min.   :  0.00   Min.   :-59.5   Min.   :-261.00   Min.   :   0.0
##  1st Qu.: 19.00   1st Qu.:-22.5   1st Qu.: -41.00   1st Qu.: 259.0
##  Median : 23.00   Median : 15.0   Median :  87.00   Median : 494.0
##  Mean   : 27.76   Mean   : 15.0   Mean   :  81.46   Mean   : 710.5
##  3rd Qu.: 27.00   3rd Qu.: 52.5   3rd Qu.: 227.00   3rd Qu.: 943.0
##  Max.   :198.00   Max.   : 89.5   Max.   : 307.00   Max.   :3200.0
##                                   NA's   :37784     NA's   :37778
##       npp
##  Min.   :    0
##  1st Qu.: 1890
##  Median : 3694
##  Mean   : 4492
##  3rd Qu.: 6200
##  Max.   :22916
##  NA's   :41915
```

Now that we have the data uploaded, lets plot them onto the map. Namely, we will transfer the species richness data onto the raster (blank).

```
blank <- terra::rast("blank.asc")
blank[] <- mamm$sprich
```

We can now plot the map. If you try to plot the raster but encounter an *error*, check whether the raster has the coordinate system specified. If not, specify it manually and set it to "WGS84", as shown below.

```
crs(blank) <- "+proj=longlat +datum=WGS84 +no_defs"
levelplot(blank, par.settings=magmaTheme, contour=F,margin=F, main="species richness of mammals")
```

species richness of mammals
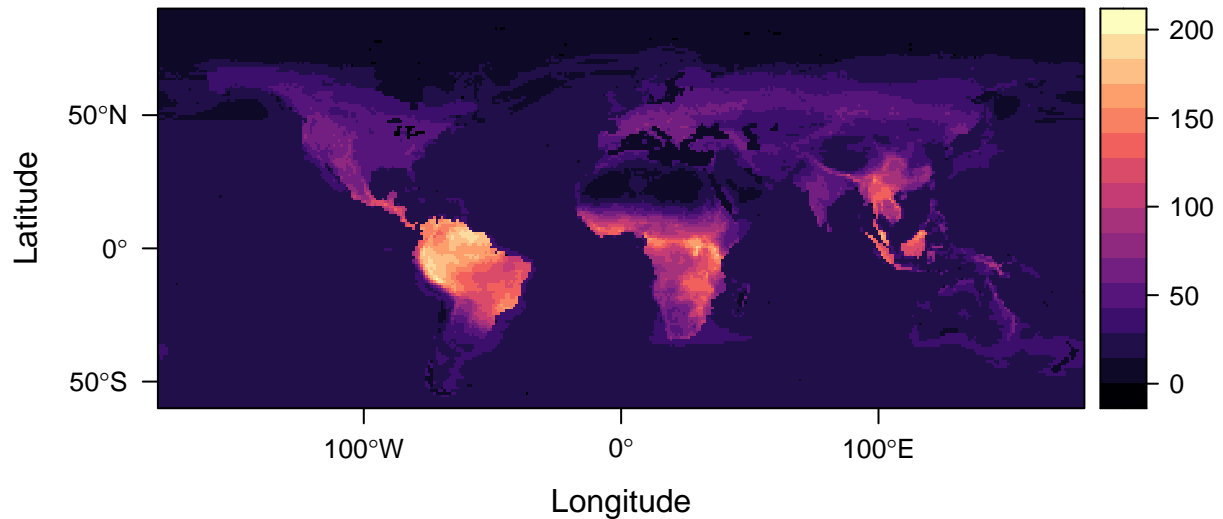
Questions: (1) What is the highest concentration of mammals within a region (i.e. number of species within a 1x1 degree grid cell)? (2) Where is the diversity of mammals generally high, where low? (3) Can you find places within similar environmental conditions, similar latitude, but with very different diversity of mammals? (4) What could cause these differences?

Diversity patterns in mammals

We first examined species richness two-dimensionally across a map (latitude × longitude). To test mechanisms more rigorously, we now collapse space to a one-dimensional gradient and model richness as a function of a single predictor (e.g., temperature, precipitation). As a starting point, we reconstruct the classic latitudinal diversity gradient (LDG) in mammals.

```
plot (sprich ~ latit, mamm, pch = 19, col = "darkblue", cex = 0.25)
abline (v=0, lwd=2)
```
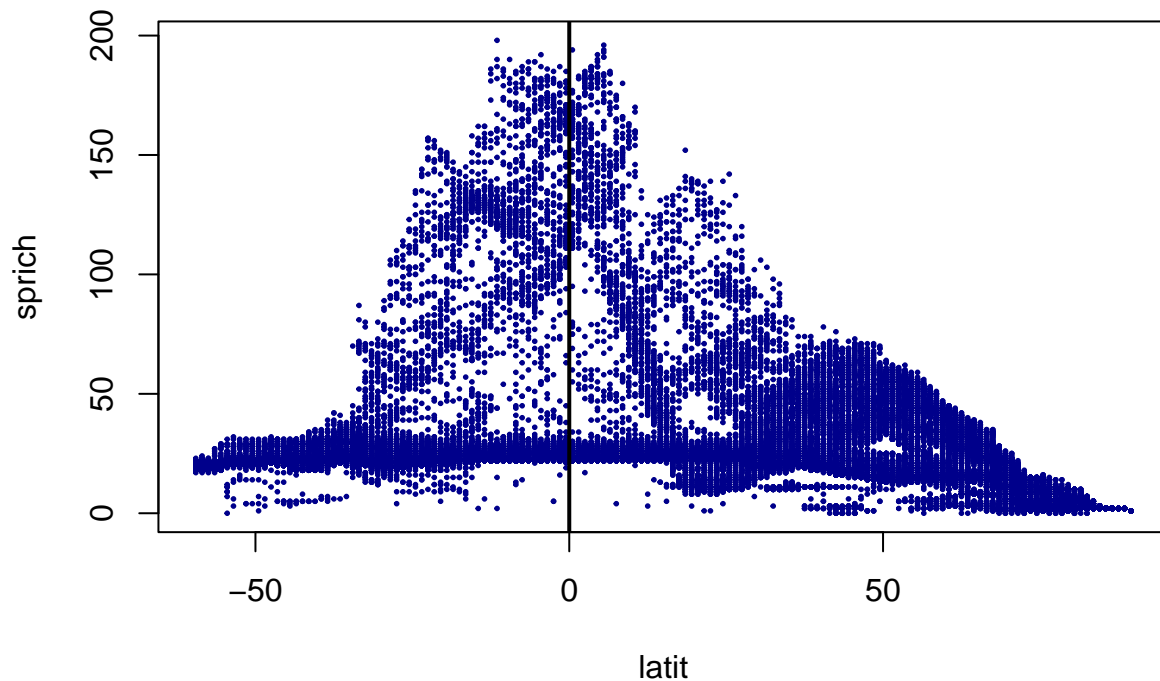
```
plot (sprich ~ abs(latit), mamm, pch = 19, col = "darkred", cex = 0.25)
```

Explaining the diversity patterns

Although hundreds of explanations have been proposed to explain large-scale diversity gradients, only a small subset has matured into compelling theories. A **theory** connects multiple mechanisms into a general process, makes quantitative **predictions**, and can be **empirically tested** and potentially falsified across organismal and geographic systems. In other words, a theory goes beyond noting a correlation (e.g. "richness is higher where it's warm") and formulates a testable model of why and how the diversity pattern arises.

Three of the prominent theories are **(1) Kinetic theory, (2) Metabolic theory, (3) Energy Limits**. All three share a common idea that energy availability influences the rates at which species originate, persist, or coexist. All three produce explicit, testable mathematical forms relating richness to environmental gradients. However, the three theories differ in the underlying process they emphasize. Kinetic theory stresses temperature-driven metabolic and evolutionary rates. Metabolic theory stresses the geometry of resource capture and energy flux, scaling. Energy limitation stresses niche saturation and environmental constraints at high energy levels.

Consequently, the **theories diverge in their predictions** about the diversity-environment relationships. The predicted relationships can be fitted into empirical data, such that we can rigorously statistically assess which relationships better corresponds with diversity patterns in a given organismal group (e.g. mammals).

Note that the diversity patterns and their environmental predictors are always **measured with an error**. We do not know exactly how many species occur in a given grid cell, what is the productivity in this grid cell, some cells are better explored (e.g. Europe) than others (e.g. Amazonia). This means that we have only estimates, which are getting increasingly accurate, as we gather more empirical data on species occurrences,

satellite imagery, weather station measurements, and as we develop better climatic models. Considering these limitations, we **do not really know what is the shape of the latitudinal diversity gradient** (e.g. linear, exponential, power law) and we need to test the shape statistically, using empirical data.

Theory	Core idea	Predicted relationship (richness vs. environment)
Kinetic theory	Higher temperature accelerates metabolism, speciation, and persistence, which produces more species.	linear relationship on a semi-log plot: `log(S) ~ temperature`.
Metabolic theory	Richness scales as a power of available energy and productivity due to fractal resource use.	linear relationship on a log–log plot: `log(S) ~ log (NPP)`.
Energy limitation	Energy boosts richness at low levels but other factors cap it at high energy levels.	Saturating Michaelis–Menten curve: S rises steeply then levels off at Smax.

To test these predictions, we need to format the data. Given that the tested relationships are supposed to be studied after log-transforming species richness and/or its predictors, we need to remove zero and missing values.

```r
#### Remove NAs and zeros, which is needed for the log-transformations
dat <- mamm %>%
  select(sprich, temp, npp) %>%
  filter(is.finite(sprich), is.finite(temp), is.finite(npp)) %>%
  mutate(
    sprich = as.numeric(sprich),
    temp   = as.numeric(temp),
    npp    = as.numeric(npp)
  )

### For log models, we further need strictly positive values
dat_log <- dat %>% filter(sprich > 0, npp > 0)

# Optional: scale NPP so the saturating nls is numerically stable
dat_sat <- dat %>%
  mutate(npp_s = npp / max(npp, na.rm = TRUE))
```

Now, we can fit the predicted relationships. We will basically find such **parameters of the theorized functions (e.g. exponential, power law) that maximize the fit of these functions to the empirical data** (e.g. species richness and temperature). The fitting can be accomplished using simple linear models (after log-transformation) or non-linear optimization, which fits a given function into the data and finds parameter values for the function (e.g Smax) that maximize its fit (e.g. maximizes the likelihood of the data). This optimized fit is then reported and compared to the alternative fits (e.g. linear fit). The optimization is accomplished using the **"nls"** function.

```r
### (A) Kinetic theory
###     log(S) ~ Temperature
###     semi-log relationship

m_exp <- lm(log(sprich) ~ temp, data = dat_log)

### (B) Metabolic theory
###     log(S) ~ log(NPP)
```

```r
###      power law log-log relationship

m_pow <- lm(log(sprich) ~ log(npp), data = dat_log)

### (C) Energy theory
###      S = Smax * (1 - exp(-b * NPP_scaled))
###      Michaelis-Menten-like saturating relationship

# Provide simple starting values; increase maxiter if needed
m_sat <- nls(
 sprich ~ Smax * (1 - exp(-b * npp_s)),  ### specify the equation to be fitted
 data   = dat_sat,
 start  = list(Smax = max(dat_sat$sprich, na.rm = TRUE) * 1.05, b = 2), ### specify starting values
 control = list(maxiter = 200, warnOnly = TRUE) ### setup for the optimization
 )
```

Model comparison, p-values, AIC and Akaike's weights

Model comparison evaluates how well competing statistical models explain a dataset. In frequentist statistics, fit is often assessed by test statistics and **p-values**, which measure whether a model term improves fit relative to a null hypothesis. Broader model selection frameworks use information criteria such as **AIC**, which balances goodness of fit (log-likelihood, R^2) against model complexity. Akaike weights then convert AIC differences into normalized probabilities.

Measures like R^2 or likelihood values describe how much variation a model explains and thus its absolute fit, whereas AIC and weights reflect relative fit or the ranking of the models with respect to each other. Bayesian approaches achieve a similar goal with posterior probabilities of the model, whereby our prior beliefs are updated based on the newly compiled data. Bayesian inference represents an extension of likelihood inference. Identifying the best model is essential for prediction as well.

```r
### OUTPUT

summary(m_exp)
```

```
##
## Call:
## lm(formula = log(sprich) ~ temp, data = dat_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6005 -0.2297  0.0910  0.3237  1.2463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.5275646  0.0058208  606.03   <2e-16 ***
## temp        0.0031690  0.0000358   88.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5276 on 12050 degrees of freedom
## Multiple R-squared:  0.3941, Adjusted R-squared:  0.394
## F-statistic:  7837 on 1 and 12050 DF,  p-value: < 2.2e-16
```

```
summary(m_pow)
```

```
## 
## Call:
## lm(formula = log(sprich) ~ log(npp), data = dat_log)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1399 -0.3446 -0.0326  0.4002  2.7584
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.544039   0.044149   12.32   <2e-16 ***
## log(npp)    0.405542   0.005432   74.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5605 on 12050 degrees of freedom
## Multiple R-squared:  0.3163, Adjusted R-squared:  0.3162
## F-statistic:  5575 on 1 and 12050 DF,  p-value: < 2.2e-16
```

```
summary(m_sat)
```

```
## 
## Formula: sprich ~ Smax * (1 - exp(-b * npp_s))
## 
## Parameters:
##       Estimate Std. Error t value Pr(>|t|)
## Smax 132.72867    2.24032   59.24   <2e-16 ***
## b      3.17134    0.08795   36.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 30.91 on 12059 degrees of freedom
## 
## Number of iterations to convergence: 4
## Achieved convergence tolerance: 1.565e-06
```

```
### COMPARE THE AIC VALUES

AIC(m_exp); AIC(m_pow); AIC(m_sat)
```

```
## [1] 18794.42
```

```
## [1] 20249.75
```

```
## [1] 116995
```

Questions: (1) Which of the three theories explains most of the variation in the diversity of mammals? (2) Compare the p-values produced by the three models, what do these p-values mean, and would it make sense to compare them to each other? (3) To identify the better supported theory, we can use either p-values, or

7

AIC values, or both. Which approach would you prefer? (4) For the AIC values, do we prefer models with higher or lower AIC? (5) What does the p-value mean? It is "p" for probability, but it is a probability of what?

There are many ways to compare the models. One way, which is convenient, is based on the **Akaike's weights**. These weights turn AIC values into normalized probabilities. They capture the probability that a given model is the true model, under the assumption that the true model is included among the tested models (which of course might not be the case). Conveniently, the weights express relative support, ranging in values from zero to one, for alternative models, given the data.

Formally, for model i with AIC_i:

$$\Delta_i = AIC_i - \min(AIC), \quad w_i = \frac{\exp(-\Delta_i/2)}{\sum_{j=1}^{R} \exp(-\Delta_j/2)}$$

where R is the number of models. The weights w_i sum to 1 and are used for model comparison or model averaging. Hirotugu Akaike (1927–2009) was a Japanese statistician who created AIC and originally devised it while working on weather-forecasting models at Japan's Meteorological Research Institute.

```
### CALCULATE AKAIKE'S WEIGHTS

aic_exp <-  AIC(m_exp)
aic_pow <-  AIC(m_pow)
aic_sat <- AIC(m_sat)

# Put into a data frame
aic_df <- data.frame(
  model = c("Exponential (Temp)", "Power law (NPP)", "Saturating (NPP)"),
  AIC   = c(aic_exp, aic_pow, aic_sat)
  )

# Compute deltaAIC
aic_df$delta <- aic_df$AIC - min(aic_df$AIC)

# Compute Akaike weights
aic_df$weight <- exp(-0.5 * aic_df$delta)
aic_df$weight <- aic_df$weight / sum(aic_df$weight)

print(aic_df)
```

```
##                 model       AIC    delta        weight
## 1 Exponential (Temp)  18794.42     0.00  1.000000e+00
## 2    Power law (NPP)  20249.75  1455.33  9.531675e-317
## 3   Saturating (NPP) 116995.03 98200.61  0.000000e+00
```

```
# barplot of the weights
library(ggplot2)
ggplot(aic_df, aes(x = reorder(model, -weight), y = weight, fill = model)) +
  geom_col() +
  labs(x = "Model", y = "Akaike weight", title = "Model support (AIC weights)") +
  theme_minimal() +
  theme(legend.position = "none")
```

Model support (AIC weights)
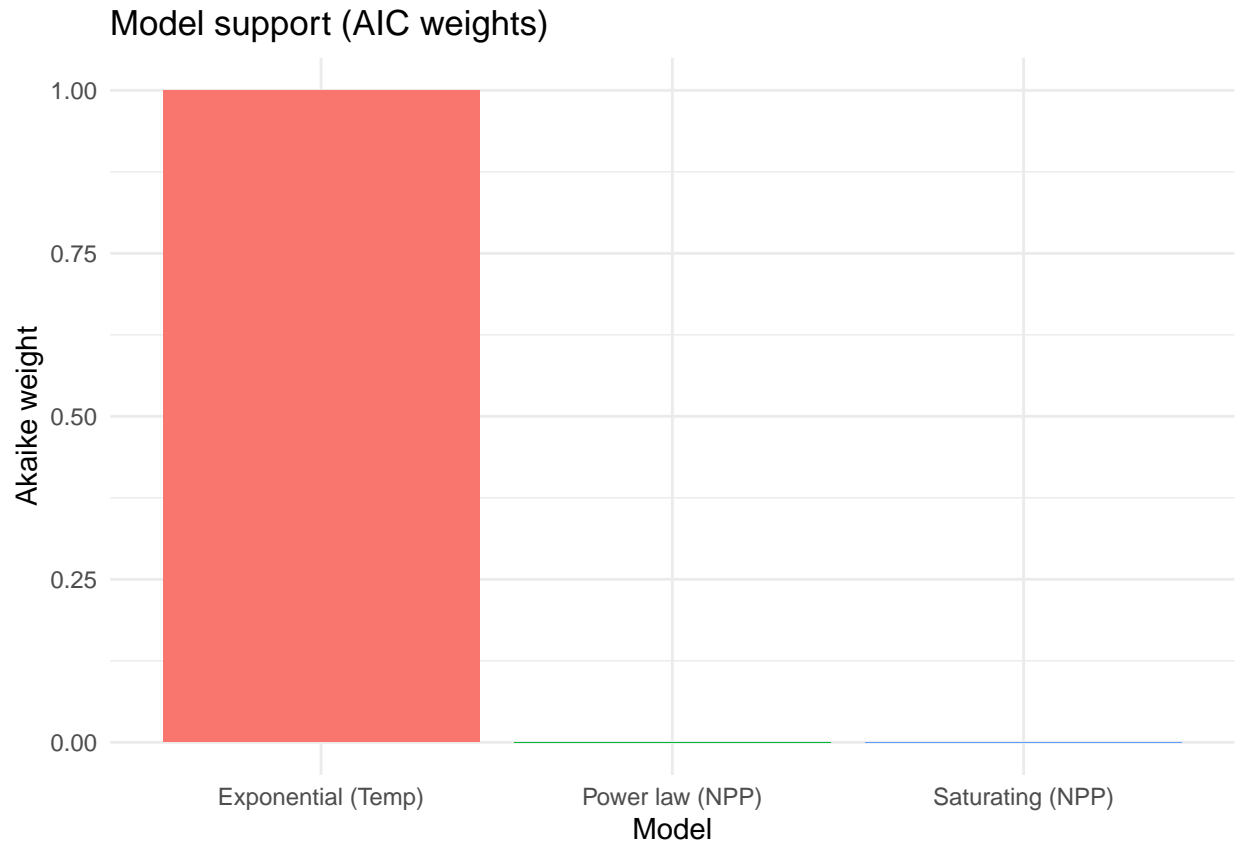
Predictions, relative and absolute fit of the model

Even the **relatively best fitting model** may still be an absolutely **poor description** of the data. It is therefore essential to examine the absolute fit as well as the relative ranking. After fitting each model, generate predicted values across the observed range and plot these predictions against the empirical data. This allows you to see directly how well each model captures the pattern and whether systematic deviations remain, signalling the need for a different functional form or additional predictors. This also applies to predictions.

We can have the best fitting model, which still fails to predict the data, if the model fit is very low in absolute terms. In fact, many of the real models that we use not only in biology, but also in economics, weather forecasting, provide rather poor predictions. Even poor predictions, however, are much better than random guess, biased expert opinions, or even no prediction at all. Consequently, predictive models save taxpayers millions of euros every year.

```r
### PLOT THE PREDICTIONS TO COMPARE THE MODELS

library(ggplot2)
library(patchwork)    # install.packages(patchwork, ggplot2) if needed

# 1) Compute R² for each model

# For linear models, we can use summary()
r2_exp <- summary(m_exp)$r.squared
r2_pow <- summary(m_pow)$r.squared
```

```r
# For nls, compute pseudo-R² = 1 - SSE/SST
sse <- sum(residuals(m_sat)^2)
sst <- sum((dat_sat$sprich - mean(dat_sat$sprich))^2)
r2_sat <- 1 - sse/sst

# 2) Generate predicted curves as before

new_temp <- data.frame(temp = seq(min(dat_log$temp), max(dat_log$temp), length.out = 200))
  pred_exp_curve <- data.frame(
  temp = new_temp$temp,
  sprich_pred = exp(predict(m_exp, newdata = new_temp))
  )

new_npp <- data.frame(npp = seq(min(dat_log$npp), max(dat_log$npp), length.out = 200))
  pred_pow_curve <- data.frame(
  npp = new_npp$npp,
  sprich_pred = exp(predict(m_pow, newdata = new_npp))
  )

new_npp_sat <- data.frame(npp_s = seq(min(dat_sat$npp_s), max(dat_sat$npp_s), length.out = 200))
  pred_sat_curve <- data.frame(
  npp_s = new_npp_sat$npp_s,
  npp   = new_npp_sat$npp_s * max(dat$npp, na.rm = TRUE),
  sprich_pred = predict(m_sat, newdata = new_npp_sat)
  )

# 3) Build ggplots with R² annotation

p_exp <- ggplot(dat_log, aes(x = temp, y = sprich)) +
  geom_point(alpha = 0.4) +
  geom_line(data = pred_exp_curve, aes(x = temp, y = sprich_pred),
        color = "blue", linewidth = 1) +
        annotate("text", x = Inf, y = Inf,
        label = paste0("R² = ", round(r2_exp, 3)),
        hjust = 1.1, vjust = 2, size = 4) +
        labs(x = "Temperature (°C)", y = "Richness", title = "Kinetic / Exponential")

p_pow <- ggplot(dat_log, aes(x = npp, y = sprich)) +
  geom_point(alpha = 0.4) +
  geom_line(data = pred_pow_curve, aes(x = npp, y = sprich_pred),
        color = "red", linewidth = 1) +
        annotate("text", x = Inf, y = Inf,
        label = paste0("R² = ", round(r2_pow, 3)),
        hjust = 1.1, vjust = 2, size = 4) +
        labs(x = "NPP", y = "Richness", title = "Metabolic Power-law")

p_sat <- ggplot(dat_sat, aes(x = npp, y = sprich)) +
  geom_point(alpha = 0.4) +
  geom_line(data = pred_sat_curve, aes(x = npp, y = sprich_pred),
        color = "darkgreen", linewidth = 1) +
        annotate("text", x = Inf, y = Inf,
        label = paste0("R² = ", round(r2_sat, 3)),
        hjust = 1.1, vjust = 2, size = 4) +
```
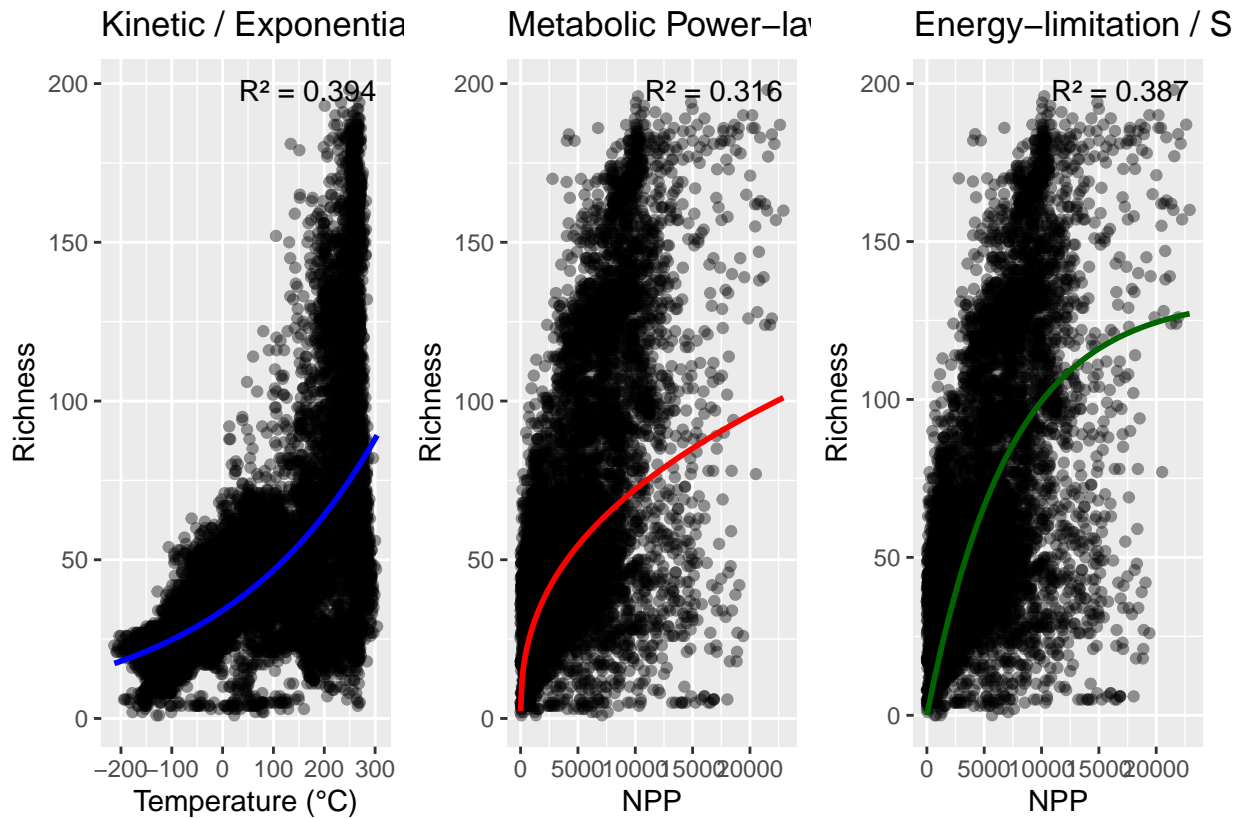
```
              labs(x = "NPP", y = "Richness", title = "Energy-limitation / Saturating")

# 4) Combine plots side by side

(p_exp | p_pow | p_sat)
```


Question: (1) Comment on the fit of the three models, which theory do you think provides the best explanation for diversity patterns in mammals? (2) Is this best of the three theories actually good? Does the theory explain the empirical patterns sufficiently well, or is it just the best of three poor explanations? (3) How could the theory be extended to produce an even better explanation, what is missing from the theory? What other factors likely contribute to diversity patterns in mammals?