Statictical Modeling and Advanced Regression Analyses

R Tutorials

Holger Sennhenn-Reulen
Northwest German Forest Research Institute (NW-FVA), Germany.

October 14, 2024

Contents

1 Software

We use the statistical software environment *R* (R Core Team, 2024), and R add-on packages *ggplot2* (Wickham, 2016).

This document is produced using *Quarto* (Allaire et al., 2024).

1.1 Organize R Session

```
rm(list = ls())
library("ggplot2")
```

2 Linear Regression Model

2.1 Data Simulation

Data are simulated according to the equations given in the lecture slides[1]:

```
set.seed(123)
N <- 500
df <- data.frame(x_1 = runif(n = N),
                 x_2 = runif(n = N))
(beta_0 <- rnorm(n = 1, mean = 1, sd = .1))
```

```
[1] 0.9398107
```

```
(beta_x_1 <- rnorm(n = 1, mean = 1, sd = .1))
```

```
[1] 0.9006301
```

```
(beta_x_2 <- rnorm(n = 1, mean = -.5, sd = .1))
```

```
[1] -0.3973215
```

```
(sigma <- rgamma(n = 1, shape = 1, rate = 4))
```

```
[1] 0.293026
```

[1] For two covariates `x_1` and `x_2`.

```
df$mu <- beta_0 + beta_x_1 * df$x_1 + beta_x_2 * df$x_2
df$y <- df$mu + rnorm(n = N, mean = 0, sd = sigma)
```

2.1.1 Visualisations

```
ggplot(data = df, aes(x = x_1, y = x_2)) +
  geom_point()
```

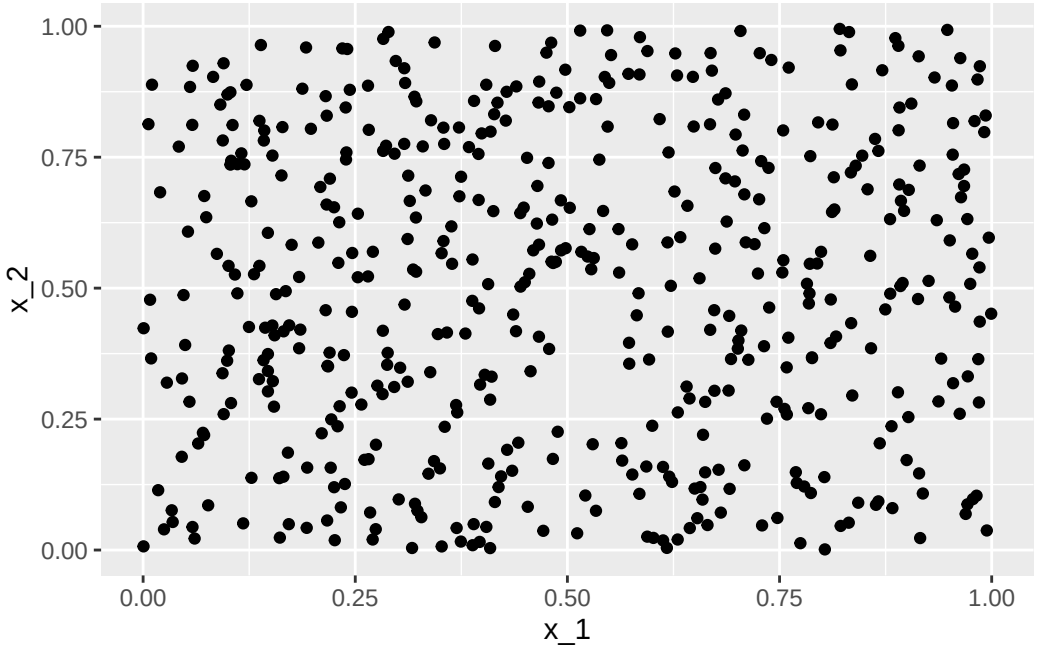

Figure 1: Scatterplot of the two simulated covariates `x_1` and `x_2` - each from the uniform distribution between 0 and 1.

```
ggplot(data = df, aes(x = x_1, y = mu, color = x_2)) +
  geom_point()
```

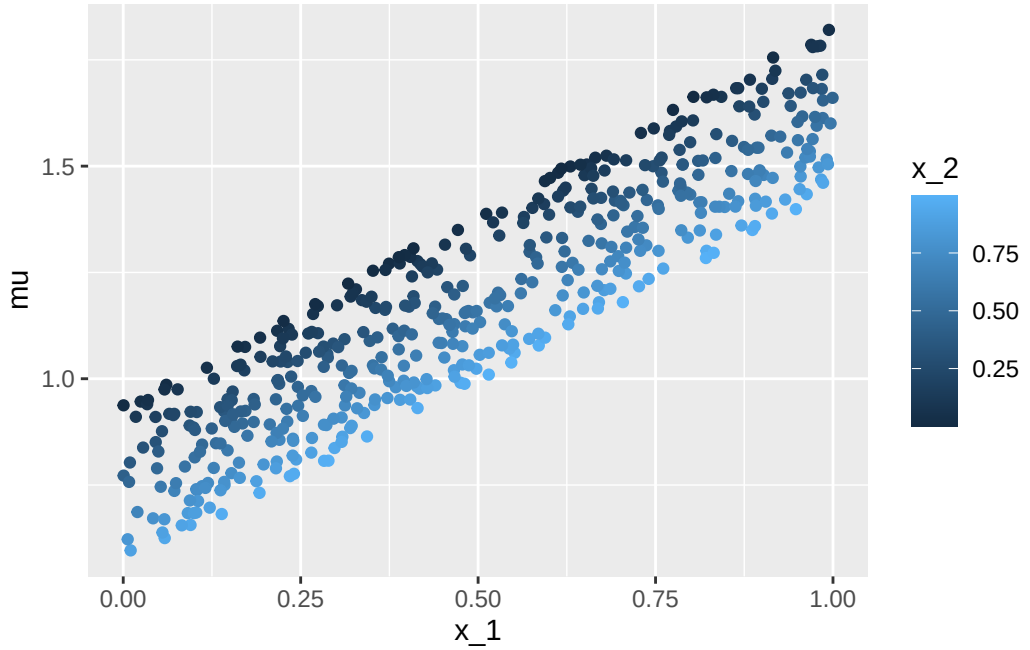

Figure 2: Scatterplot of covariate `x_1` with response `y` - each individual observation is coloured according to the second covariate `x_2`.

```
ggplot(data = df, aes(x = x_2, y = mu, color = x_1)) +
  geom_point()
```

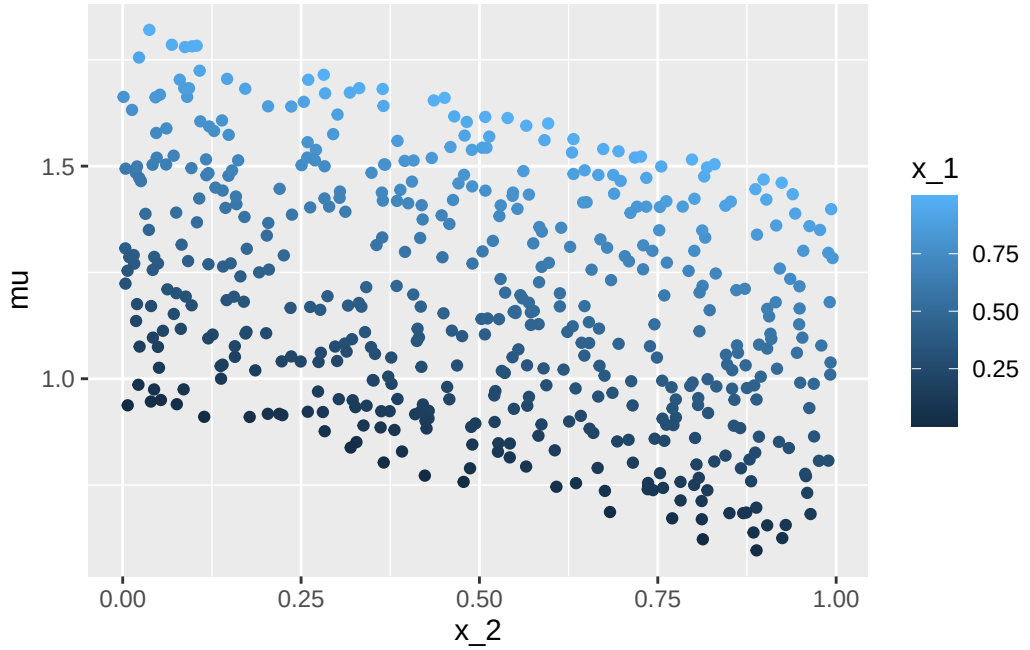

Figure 3: Scatterplot of covariate `x_2` with response `y` - each individual observation is coloured according to the first covariate `x_1`.

```
ggplot(data = df, aes(x = x_1, y = x_2, color = mu)) +
  geom_point()
```

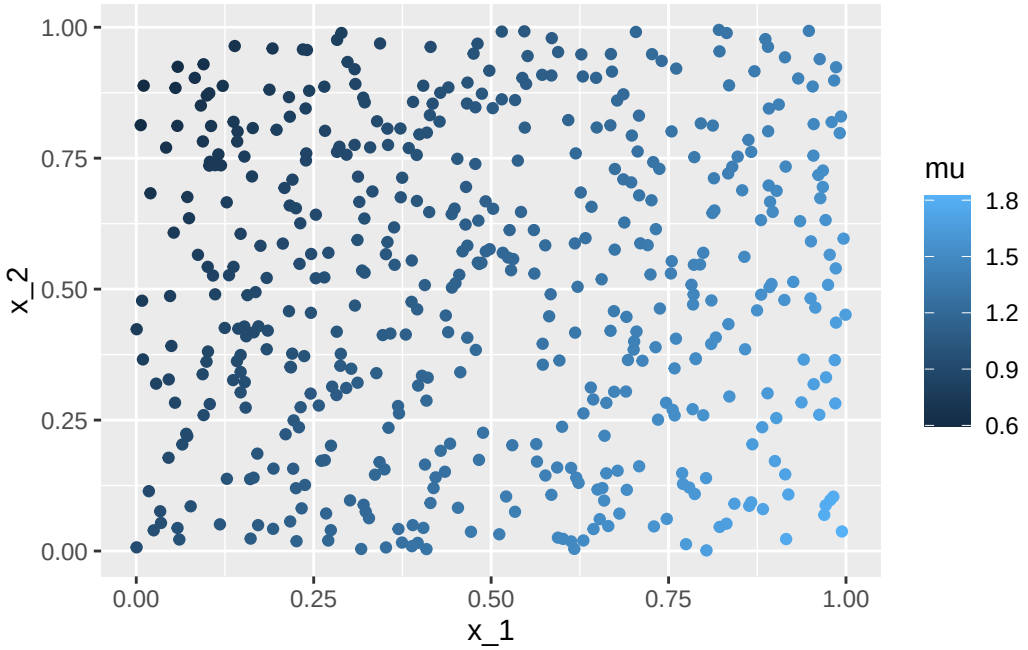

Figure 4: Scatterplot of the two simulated covariates `x_1` and `x_2` - each individual observation is coloured according to the underlying true conditional expectation `mu`.

```
ggplot(data = df, aes(x = x_1, y = x_2, color = y)) +
  geom_point()
```

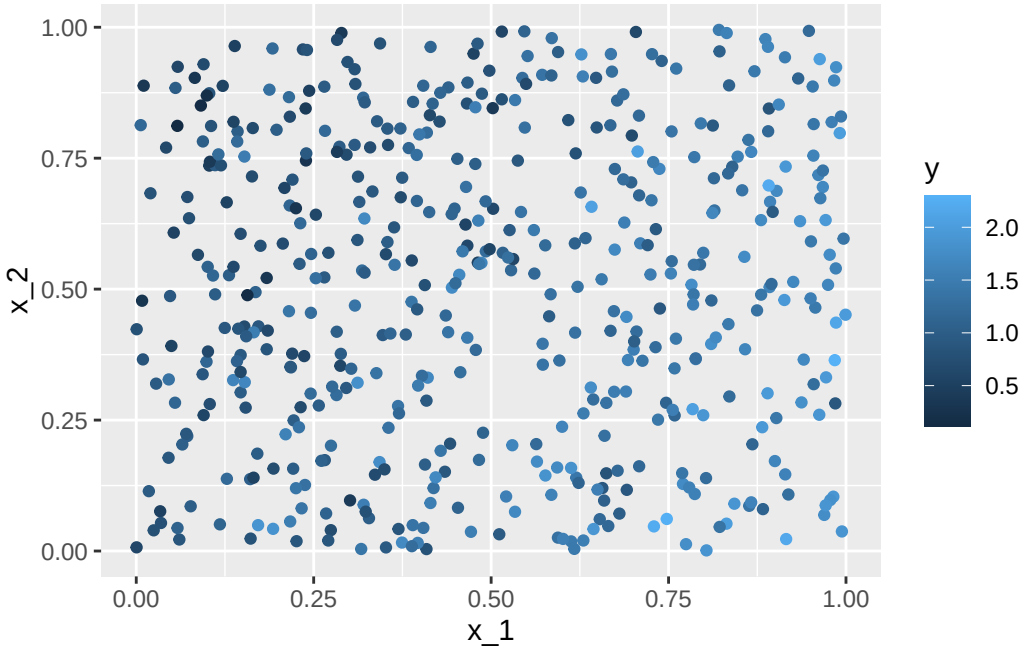

Figure 5: Scatterplot of the two simulated covariates `x_1` and `x_2` - each individual observation is coloured according to the response `y`.

2.2 Modeling

The basic R command for (frequentist) estimation of the parameters of a linear regression model is a call to the function `lm`:

```
Call:
lm(formula = y ~ x_1 + x_2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82082 -0.19805  0.00329  0.19051  0.81138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.91291    0.03448  26.476  < 2e-16 ***
x_1          0.91533    0.04668  19.610  < 2e-16 ***
x_2         -0.36218    0.04566  -7.933 1.43e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2963 on 497 degrees of freedom
Multiple R-squared:  0.4674,    Adjusted R-squared:  0.4652
F-statistic:   218 on 2 and 497 DF,  p-value: < 2.2e-16
```

2.2.1 Visualisations

```
nd <- data.frame(x_1 = seq(0, 1, by = .1),
                 x_2 = .5)
nd$mu <- predict(m, newdata = nd)
ggplot(data = df, aes(x = x_1, y = mu, color = x_2)) +
  geom_point() +
  geom_line(data = nd, aes(x = x_1, y = mu, color = x_2))
```

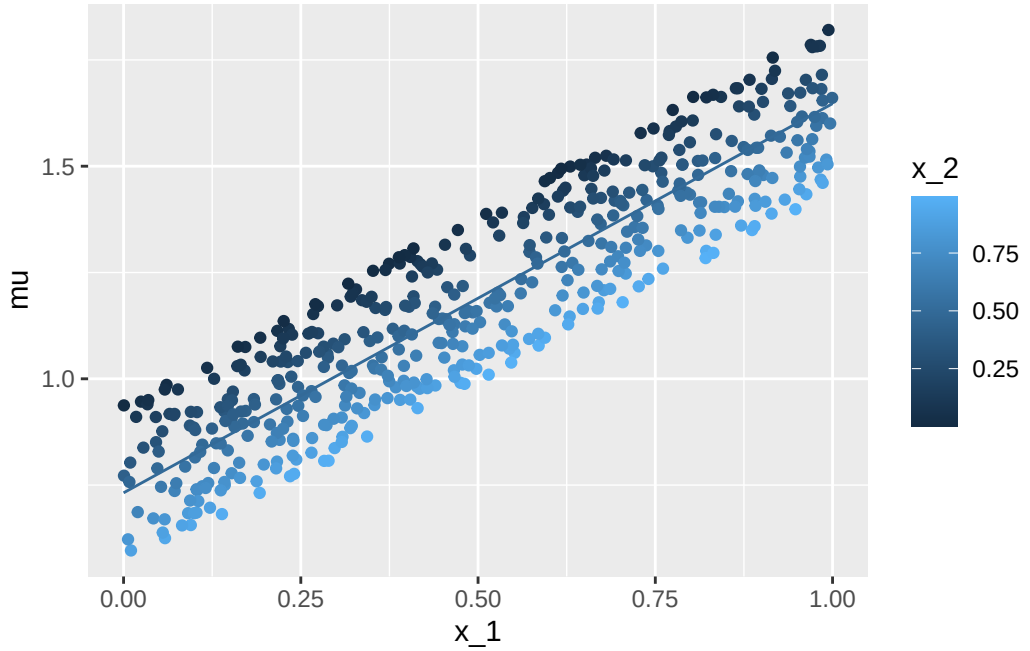

Figure 6: Scatterplot of covariate `x_1` with the true conditional expectation `mu` - each individual observation is coloured according to the second covariate `x_2`. The line gives the point estimation for the conditional expectation with the second covariate `x_2` fixed to 0.5.

```
nd <- data.frame(expand.grid('x_1' = seq(0, 1, by = .1),
                             'x_2' = seq(0, 1, by = .1)))
nd$mu <- predict(m, newdata = nd)
ggplot(data = df, aes(x = x_1, y = mu, color = x_2)) +
```

```
geom_point() +
geom_line(data = nd, aes(x = x_1, y = mu, color = x_2, group = x_2))
```

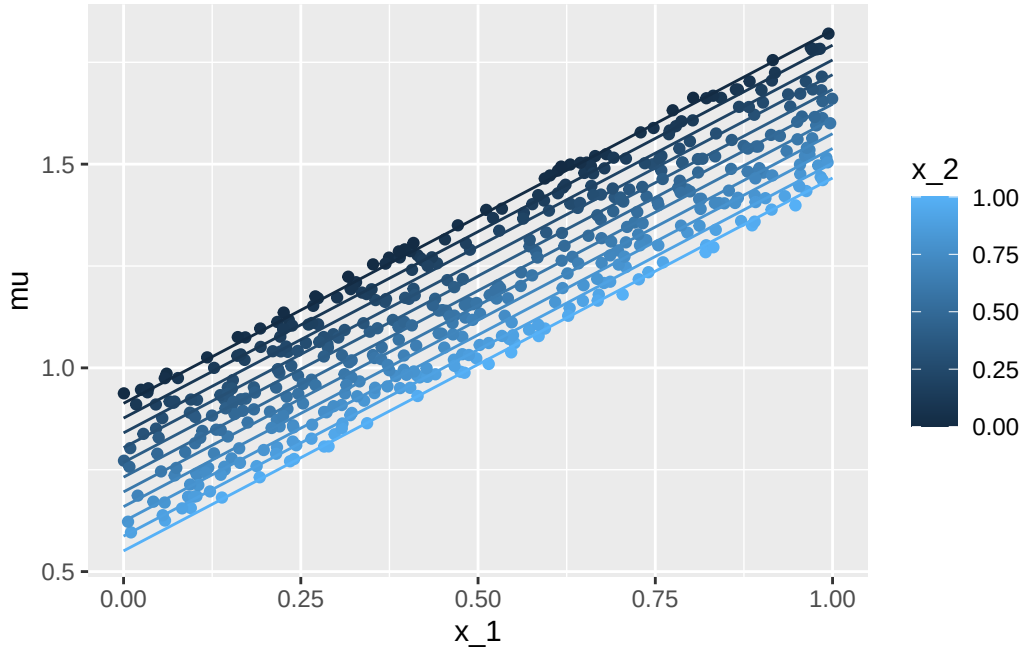

Figure 7: Scatterplot of covariate `x_1` with the true conditional expectation `mu` - each individual observation is coloured according to the second covariate `x_2`. The lines give the point estimation for the conditional expectation with the second covariate `x_2` taking on values between 0 and 1 (at steps of 0.1).

2.3 Add-Ons

2.3.1 Add-On Linear Model: A) Stancode

2.3.1.1 Stan Users Guide

Probabilistic Programming Languages such as *Stan* (Carpenter et al., 2017) allow to `plug together` the single parts of a statistical regression model[2]:

The following Stan-code is published here in the Stan users guide:

```
data {
  int<lower=0> N;
  vector[N] x;
```

[2] Which is actually pretty 'readable' if you get used to the structure for a simple model such the linear regression model.

```
  vector[N] y;
}
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;
}
model {
  y ~ normal(alpha + beta * x, sigma);
}
```

2.3.1.2 Stancode generated by calling `brms::brm`

The R add-on package *brms* (Bürkner, 2017, 2018) allows to implent advanced regression models without being an expert in 'Stan-programming'.

Here is the Stan-code that is implemented by 'brms' for our linear regression model example:

```
brms::make_stancode(brms::bf(y ~ x_1 + x_2, center = F), data = df)
```

```
// generated with brms 2.21.0
functions {
}
data {
  int<lower=1> N;  // total number of observations
  vector[N] Y;  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  int prior_only;  // should the likelihood be ignored?
}
transformed data {
}
parameters {
  vector[K] b;  // regression coefficients
  real<lower=0> sigma;  // dispersion parameter
}
transformed parameters {
  real lprior = 0;  // prior contributions to the log posterior
  lprior += student_t_lpdf(sigma | 3, 0, 2.5)
    - 1 * student_t_lccdf(0 | 3, 0, 2.5);
}
model {
  // likelihood including constants
```

```
  if (!prior_only) {
    target += normal_id_glm_lpdf(Y | X, 0, b, sigma);
  }
  // priors including constants
  target += lprior;
}
generated quantities {
}
```

2.3.2 Add-On Linear Model: B) Posterior predictive check: an introduction 'by hand'

Having an `lm` object already, it is rather straightforward to get posterior samples by using function `sim` from the *arm* (Gelman & Su, 2024) package:

```
library("arm")


S <- sim(m)
str(S)
```

```
Formal class 'sim' [package "arm"] with 2 slots
  ..@ coef : num [1:100, 1:3] 0.882 1.014 0.904 0.978 0.958 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : NULL
  .. .. ..$ : chr [1:3] "(Intercept)" "x_1" "x_2"
  ..@ sigma: num [1:100] 0.323 0.303 0.292 0.309 0.29 ...
```

```
S <- cbind(S@coef, 'sigma' = S@sigma)
head(S)
```

```
     (Intercept)       x_1        x_2     sigma
[1,]   0.8816414 0.9245094 -0.3362733 0.3227662
[2,]   1.0139849 0.7317948 -0.3398411 0.3033703
[3,]   0.9037042 0.9155575 -0.3506924 0.2922883
[4,]   0.9776909 0.8392790 -0.3845609 0.3090220
[5,]   0.9579213 0.8977625 -0.4284596 0.2900632
[6,]   0.9549211 0.8478278 -0.3937226 0.3094227
```

Predict the response for the covariate data as provided by the original data-frame `df` - here only by using the first posterior sample:

```
s <- 1
S[s, ]
```

```
(Intercept)         x_1         x_2       sigma
  0.8816414   0.9245094  -0.3362733   0.3227662
```

```
mu_s <- S[s, '(Intercept)'] + S[s, 'x_1'] * df$x_1 + S[s, 'x_2'] * df$x_2
y_s <- rnorm(n = nrow(df), mean = mu_s, sd = S[s, 'sigma'])
pp <- rbind(data.frame(y = df$y, source = "original", s = s),
            data.frame(y = y_s, source = "predicted", s = s))
ggplot(data = pp, aes(x = y, fill = source)) +
  geom_histogram(alpha = .5, position = "identity")
```

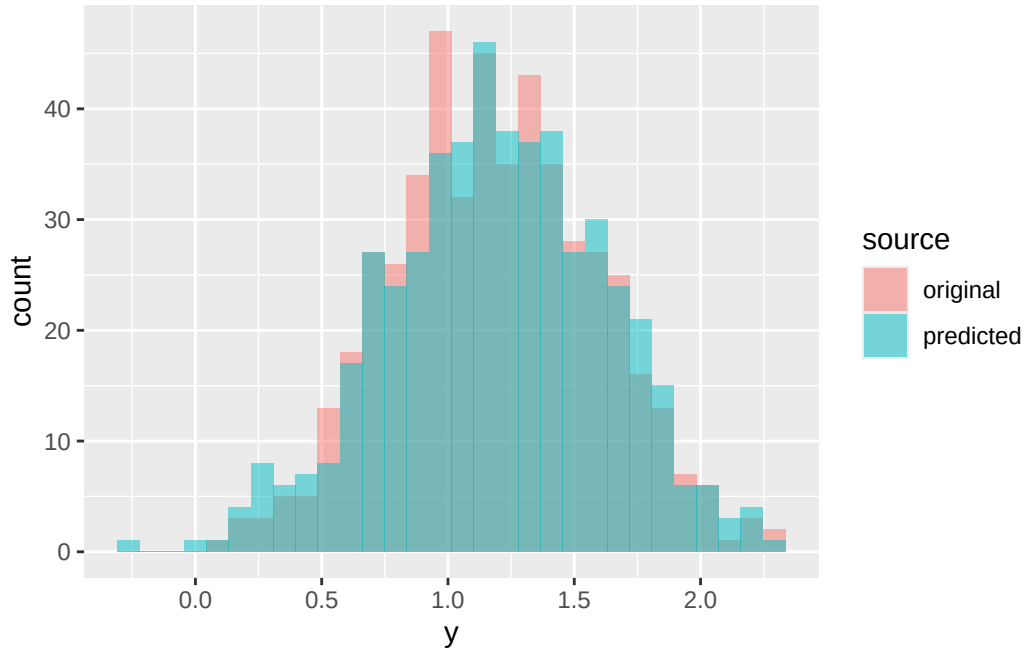

Figure 8: Histogram of the original and the posterior predicted response sample.

Now let's repeat the same for 9 different posterior samples:

```
pp <- NULL
for (s in 1:9) {
  mu_s <- S[s, '(Intercept)'] + S[s, 'x_1'] * df$x_1 + S[s, 'x_2'] * df$x_2
  y_s <- rnorm(n = nrow(df), mean = mu_s, sd = S[s, 'sigma'])
  pp <- rbind(pp,
              data.frame(y = df$y, source = "original", s = s),
```

```
                data.frame(y = y_s, source = "predicted", s = s))
}
ggplot(data = pp, aes(x = y, fill = source)) +
  geom_histogram(alpha = .5, position = "identity") +
  facet_wrap(~ s)
```

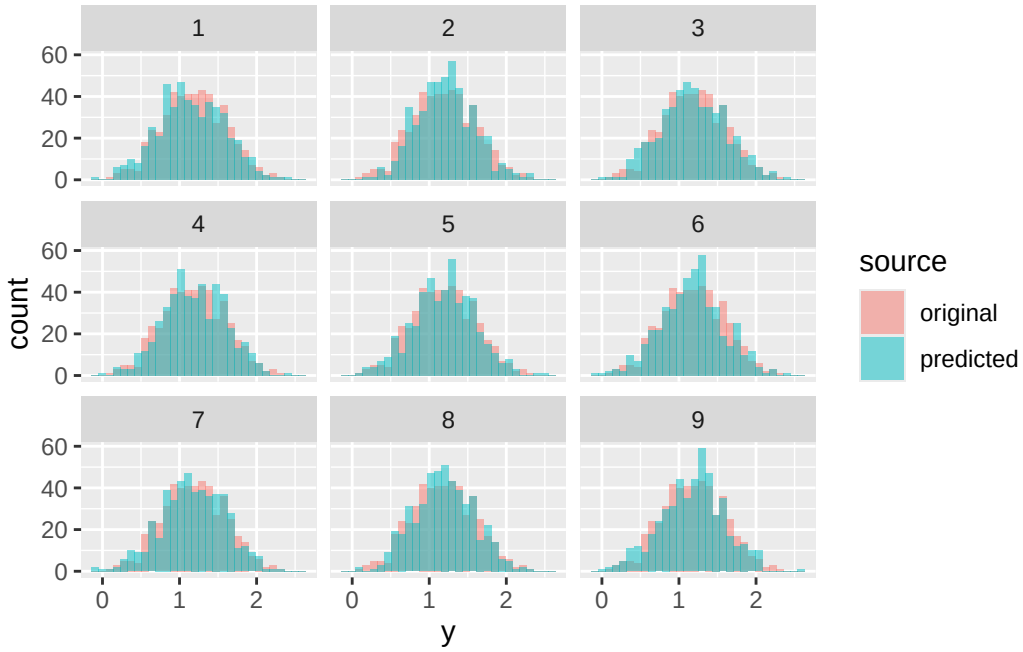

Figure 9: Histogram of the original and the posterior predicted response sample.

```
ggplot(data = pp, aes(x = y, fill = source)) +
  geom_density(alpha = .5, position = "identity") +
  facet_wrap(~ s)
```

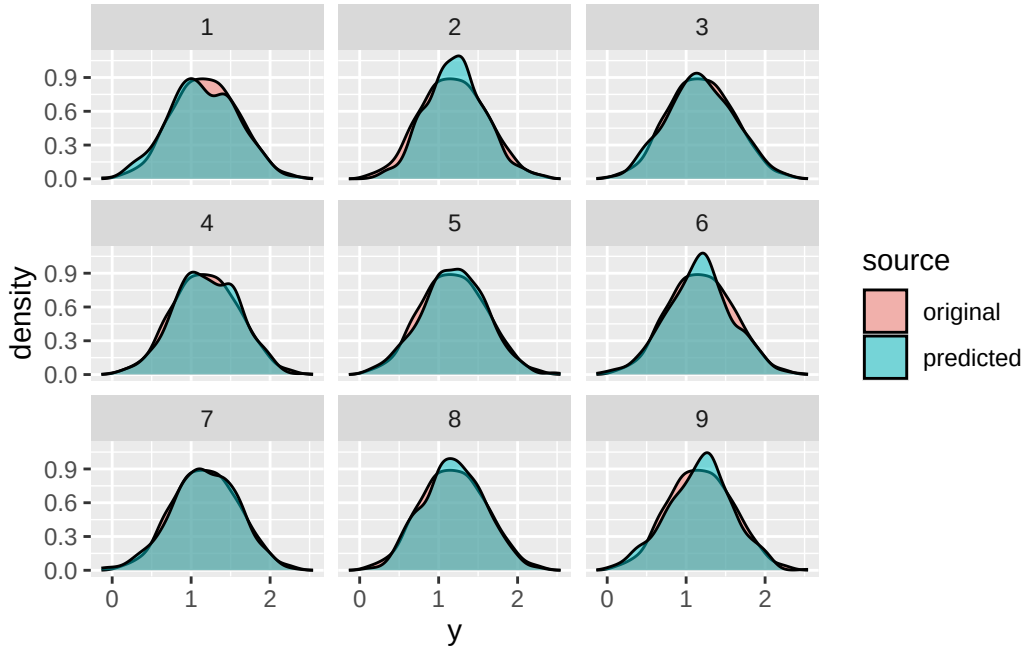

Figure 10: The same as in Figure 9, but now using kernel density visualisations.

```
ggplot(data = pp, aes(x = y, colour = source)) +
  stat_ecdf() +
  facet_wrap(~ s)
```

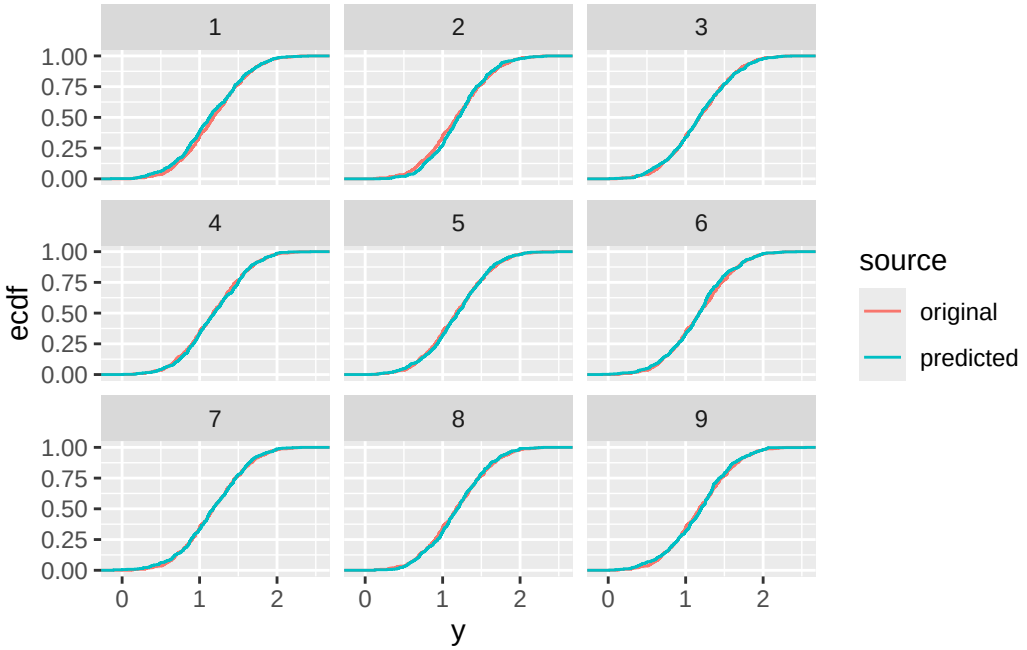

Figure 11: The same as in Figure 9 or Figure 10, but now using empirical cumulative density function visualisations.

```
ggplot(data = subset(pp, source == "predicted"),
       aes(x = y, group = s)) +
  geom_density(position = "identity", fill = NA, colour = "grey") +
  geom_density(data = subset(pp, source == "original" & s == 1),
               aes(x = y), linewidth = 1)
```

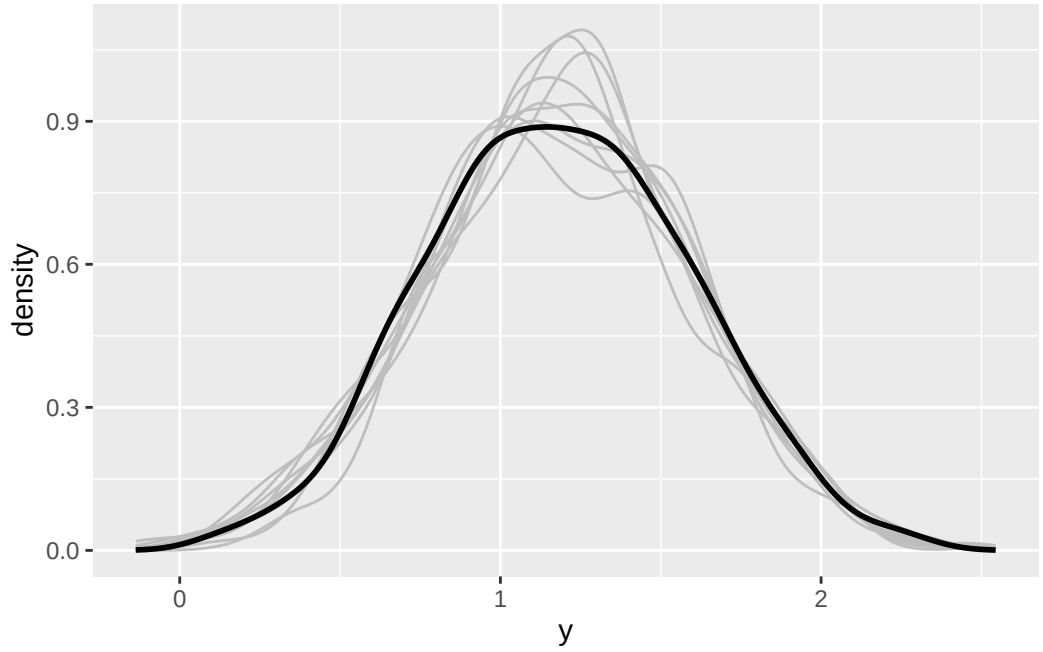

Figure 12: The same as in Figure 12, but now within one plotting window: This visualisation is what `brms::pp_check` will produce if applied on a `brm` object.

References

Allaire, J. J., Teague, C., Scheidegger, C., Xie, Y., & Dervieux, C. (2024). *Quarto (Version 1.4.553)*. https://doi.org/10.5281/zenodo.5960048

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*, 1–32. https://doi.org/10.18637/jss.v076.i01

Gelman, A., & Su, Y.-S. (2024). *Arm: Data analysis using regression and multilevel/hierarchical models*. https://CRAN.R-project.org/package=arm

R Core Team. (2024). *R: A Language and Environment for Statistical Computing (Version 4.4.1)*. R Foundation for Statistical Computing.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org