Classwork 14b: Ramsey RESET Simulation with `map(.x, .f)`

In CW13b, you wrote a function to take any model and dataset and to conduct the Ramsey RESET test for functional misspecification on it. In this classwork, you will use that function to run an experiment to learn a little more about when you can expect the test to work well and when you might not.

```
library(tidyverse)
library(gapminder)
students <- read_csv("https://raw.githubusercontent.com/cobriant/students_dataset/main/students.csv")
```

1. Define your function `ramsey`.

```
# ramsey <- function(__) {
#   __
# }
```

2. Generate a (fake) dataset with 3 variables: `x`, `z`, and `y` and 100 observations.

`x` and `z` should be independent random variables that are random uniform between 0 and 10. `y` should be `1 + .4 * x + .3 * z - .04 * x^2 - .04 * z^2 + .03 * x * z + u`, where `u` is N(0, 1). Don't give the tibble a name. Since it has random elements, each time you run the code, the values will be different. That's the behavior we want.

```
#?runif

# tibble(
#   x = runif(__),
#   z = runif(__),
#   y = 1 + .4 * x + .3 * z - .04 * x^2 - .04 * z^2 + .03 * x * z + rnorm(n = 100)
# )
```

3. Take your tibble from question 2 and visualize a scatterplot of `x` and `y`.

Also estimate the misspecified model `y ~ x + z`. Running the code multiple times, does it seem like the coefficients on x and z are unbiased?

```
# tibble(
#   x = runif(__),
#   z = runif(__),
#   y = 1 + .4 * x + .3 * z - .04 * x^2 - .04 * z^2 + .03 * x * z + rnorm(n = 100)
# ) %>%
#   __
```

4. Use your function `ramsey` to conduct the Ramsey RESET test on your tibble.

Running the code several times, does it seem to successfully detect the presence of possible nonlinear relationships most of the time?

```
# tibble(
#   x = __,
```

```
#   z = __,
#   y = __ + rnorm(n = 100)
# ) %>%
#   ramsey("y ~ x + z")
```

In statistics, a type 1 error is a false positive: if you reject the null when it's actually true in the population, you have a type 1 error. A type 2 error is a false negative: if you fail to reject the null when it's actually false, you have a type 2 error.

5. If `ramsey` outputs a p.value greater than .05 in the previous question, has a type 1 error occurred, has a type 2 error occurred, or has no error occurred?

6. Next, we'll use `map(.x, .f)` to do a simulation where we run the code in question 4 100 times and count the number of times the Ramsey test rejects the null hypothesis.

We want to use map to run the function `.f` 100 times and output the results of each iteration as a tibble, so we'll use `map_dfr` (dfr stands for data frame in rows). `.x` will be the vector `1:100`, just making sure we run the function 100 times. `.f` should begin with `function(...)`: the dot-dot-dot lets that function accept the element of `.x` as an argument, but it throws it away instead of being unpacked and used inside the function body.

```
# map_dfr(
#   1:100,
#   function(...) {
#     tibble(
#       x = __,
#       z = __,
#       y = __ + rnorm(n = 100)
#     ) %>%
#       ramsey("y ~ x + z")
#   }
# ) %>%
#   count(__)
```

7. Experiment with the data generating process to explore when you can expect the Ramsey test to be more or less reliable.

Is the Ramsey test more or less reliable for smaller sample sizes? Is the Ramsey test more or less reliable when var(u) is larger? Create an example where the Ramsey test works at least 90% of the time. ## Extra Credit: instead of using `count()` at the end of the experiment, plot p-values from each iteration using geom_density. Is the Ramsey test more or less reliable for smaller sample sizes? Compare multiple sample sizes in one plot and use `fill = n`. Draw a vertical line for the p-value of .05.