Análisis de Regresión (2021-3)

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Especialización en Estadística Aplicada

Prof. Sébastien Lozano Forero (slozanof@libertadores.edu.co)

Ejemplos de Modelos Lineales Múltiple

Tabla de contenidos

Ejemplo 1

La base de datos `marketing` se encuentra alojada en el paquete `datarium`. Dicho paquete no se encuentra en la distribución estándar de R (es decir, no está disponible en el CRAI), pero sí está disponible en GitHub. La misma es una base que contiene información sobre ventas basados en gastos en publicidad

- **youtube**: Inversión en publicidad en Youtube
- **facebook**: Inversión en publicidad en Facebook
- **newspaper**: Inversión en publicidad en periódicos
- **sales**: Total de ventas

In [ ]:
```
# install.packages("devtools")
library(usethis)
library(devtools)
# devtools::install_github("kassambara/datarium")
```

```
data("marketing", package = "datarium")
head(marketing, 4)
```

In [ ]:
```
modelo <- lm(sales ~ youtube + facebook + newspaper, data = marketing)
summary(modelo)
```

In [ ]:
```
summary(modelo)$coefficient
```

In [ ]:
```
modelo2  <- lm(sales ~ youtube + facebook, data = marketing)
summary(modelo2)
```

Ejemplo 2

El salario académico de nueve meses de 2008-09 para profesores asistentes, profesores asociados y profesores en una universidad en EE.UU. Los datos se recopilaron como parte del esfuerzo continuo de la administración de la universidad para monitorear las diferencias salariales entre los miembros de la facultad, independiente de género.

- **rank (I1)**: factor con niveles AssocProf, AsstProf, Prof
- **discipline (I2)**: factor con niveles A (Departamento “Teórico”) o B (Departamento "Aplicado").
- **yrs.since.phd (I3)**: Años desde el fin del PhD.
- **yrs.service (I4)**: Años de servicio.
- **sex (I5)**: Género
- **salary (D)**: Salario del noveno mes (dólares)

In [ ]:
```
library(tidyverse)  # data loading, manipulation and plotting
library(carData)    # Salary dataset
library(broom)      # tidy model output
```

Visión detallada y general del estado de la base de datos

In [ ]:
```
glimpse(Salaries)
```

In [ ]:
```
ggplot(Salaries, aes(x = sex, y = salary, color = sex)) +
    geom_boxplot()+
    geom_point(size = 2, position = position_jitter(width = 0.2)) +
    stat_summary(fun.y = mean, geom = "point", shape = 20, size = 6, color = "blue")+
    theme_classic() +
    facet_grid(.~rank)
```

Antes de procesar el análisis detallado, ajustemos primero un modelo de regresión lineal simple en el que predecimos el salario en función de la categoría de género. Para verificar los niveles actuales de sexo, podemos usar la función `levels()` y proporcionar el nombre de la columna. Por defecto, R trata el primer nivel como nivel de referencia (aquí femenino).

In [ ]:
```
levels(Salaries$sex)
```

Vamos a crear el modelo saturado (con todas las variables). Usualmente no es el mejor enfoque, pero da una idea general de la relación entre las variables

In [ ]:
```
lm_total <- lm(salary~., data = Salaries)
summary(lm_total)
```

Para escapar del problema de la multicolinealidad (correlación entre variables independientes) y para filtrar las variables / características esenciales de un gran conjunto de variables, generalmente se realiza una regresión escalonada. El lenguaje R ofrece forward, backwards y ambos tipos de regresión paso a paso (stepwise). Se puede ajustar una regresión por pasos hacia atrás utilizando la función `step()` proporcionando el objeto del modelo inicial y el argumento de dirección. El proceso comienza con el ajuste inicial de todas las variables y luego, con cada iteración, comienza a eliminar las variables una a una si la variable no mejora el ajuste del modelo. La métrica AIC se utiliza para comprobar la mejora del ajuste del modelo.

In [ ]:
```
step(lm_total, direction = "forward")
```

Aquí, puede ver que eliminando la variable sexo del modelo, apenas provoca alguna mejora en el valor de AIC. Ahora, ajustando el modelo recomendado

In [ ]:
```
lm_step_backward <- lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
                       data = Salaries)
summary(lm_step_backward)
```

Por lo tanto, podemos probar una hipótesis de cuánto aumenta o disminuye el salario promedio para quienes tienen años de servicio de 20 a 40 años y de 40 a 60 años en comparación con los de 0 a 20 años (referencia). Para ese propósito, necesitamos crear tres categorías 0-20, 20-40, 40-60 años donde agrupamos la variable continua, es decir, los valores de años de servicio.. Discretización de la variable `Salaries`

In [ ]:
```
Salaries_mod <- Salaries %>%
    mutate(service_time_cat = case_when(
        between(yrs.service, 0, 20)~"upto20",
        between(yrs.service, 20, 40)~"20_40",
        between(yrs.service, 40, 60)~"40_60"))
```

El siguiente paso es convertir la nueva variable en categórica. Ahora bien, si revisamos los niveles podemos observar que los niveles no están en el orden correcto.

In [ ]:
```
Salaries_mod$service_time_cat <- as.factor(Salaries_mod$service_time_cat)
levels(Salaries_mod$service_time_cat)
```

Por tanto, se deberán organizar.

In [ ]:
```
Salaries_mod$service_time_cat <- relevel(Salaries_mod$service_time_cat, ref = "upto20")
levels(Salaries_mod$service_time_cat)
```

Conteos generales del cruce de variables

In [ ]:
```
table(Salaries_mod$service_time_cat)
```

In [ ]:
```
lm4 <- lm(formula = salary ~ rank + discipline + yrs.since.phd + sex + service_time_cat, data = Salaries_mod)
summary(lm4)
```

La tabla de coeficientes del modelo mostró que a medida que aumenta el tiempo de servicio, el salario disminuye (coeficientes negativos) en comparación con los 0-20 años de servicio. En comparación con la categoría de años de servicio de 0 a 20 años, una persona que tiene entre 20 y 40 años de servicio recibe en promedio 8905,1 $ menos de salario, de manera similar, una persona que tiene entre 40 y 60 años de servicio gana 16710,4 $ menos de salario.

Para el ajuste del modelo final

In [ ]:
```
step(lm4, direction = "backward")
```

In [ ]:
```
lm5 <- lm(formula = salary ~ rank + discipline + yrs.since.phd + service_time_cat,
          data = Salaries_mod)
summary(lm5)
```

Ejemplo 3 (OMS: Tendencias y Análisis del sucidio a nivel mundial)

La mayoría de los datos utilizados en este análisis se obtuvieron de la Organización Mundial de la Salud.

Notas de limpieza de datos

- 7 países eliminados (<= 3 años de datos en total)
- Se eliminaron los datos de 2016 (pocos países tenían alguno, a los que sí les faltaban datos a menudo)
- Se eliminó el HDI debido a que faltaban 2/3 datos
- El continente se agregó al conjunto de datos mediante el paquete de código de país
- África tiene muy pocos países que proporcionan datos sobre suicidios

1. Datos

In [ ]:
```
install.packages("rworldmap")

library(tidyverse) # general
library(ggalt) # dumbbell plots
library(countrycode) # continent
library(rworldmap) # quick country-level heat maps
library(gridExtra) # plots
library(broom) # significant trends within countries

theme_set(theme_light())


# 1) Import & data cleaning
```

```r
data <- read_csv("master.csv")

# glimpse(data) # will tidy up these variable names

# sum(is.na(data$`HDI for year`)) # remove, > 2/3 missing, not useable

# table(data$age, data$generation) # don't like this variable

data <- data %>%
  select(-c(`HDI for year`, `suicides/100k pop`)) %>%
  rename(gdp_for_year = `gdp_for_year ($)`,
         gdp_per_capita = `gdp_per_capita ($)`,
         country_year = `country-year`) %>%
  as.data.frame()



# 2) OTHER ISSUES

# a) this SHOULD give 12 rows for every county-year combination (6 age bands * 2 genders):

# data %>%
#   group_by(country_year) %>%
#   count() %>%
#   filter(n != 12) # note: there appears to be an issue with 2016 data
# not only are there few countries with data, but those that do have data are incomplete

data <- data %>%
  filter(year != 2016) %>% # I therefore exclude 2016 data
  select(-country_year)


# b) excluding countries with <= 3 years of data:

minimum_years <- data %>%
  group_by(country) %>%
  summarize(rows = n(),
            years = rows / 12) %>%
  arrange(years)

data <- data %>%
```

```r
  filter(!(country %in% head(minimum_years$country, 7)))


# no other major data issues found yet



# 3) TIDYING DATAFRAME
data$age <- gsub(" years", "", data$age)
data$sex <- ifelse(data$sex == "male", "Male", "Female")


# getting continent data:
data$continent <- countrycode(sourcevar = data[, "country"],
                              origin = "country.name",
                              destination = "continent")

# Nominal factors
data_nominal <- c('country', 'sex', 'continent')
data[data_nominal] <- lapply(data[data_nominal], function(x){factor(x)})


# Making age ordinal
data$age <- factor(data$age,
                   ordered = T,
                   levels = c("5-14",
                              "15-24",
                              "25-34",
                              "35-54",
                              "55-74",
                              "75+"))

# Making generation ordinal
data$generation <- factor(data$generation,
                   ordered = T,
                   levels = c("G.I. Generation",
                              "Silent",
                              "Boomers",
                              "Generation X",
                              "Millenials",
                              "Generation Z"))
```

```
data <- as_tibble(data)


# the global rate over the time period will be useful:

global_average <- (sum(as.numeric(data$suicides_no)) / sum(as.numeric(data$population))) * 100000

# view the finalized data
glimpse(data)
```

2. Análisis Global

2.1 Análisis General

In [ ]:

```
data %>%
  group_by(year) %>%
  summarize(population = sum(population),
            suicides = sum(suicides_no),
            suicides_per_100k = (suicides / population) * 100000) %>%
  ggplot(aes(x = year, y = suicides_per_100k)) +
  geom_line(col = "deepskyblue3", size = 1) +
  geom_point(col = "deepskyblue3", size = 2) +
  geom_hline(yintercept = global_average, linetype = 2, color = "grey35", size = 1) +
  labs(title = "Suicidios globales (por 100k)",
       subtitle = "Tendencia en el tiempo, 1985 - 2015.",
       x = "Año",
       y = "Suicidios por 100k") +
  scale_x_continuous(breaks = seq(1985, 2015, 2)) +
  scale_y_continuous(breaks = seq(10, 20))
```

Algunos hechos que destacan:

- La tasa máxima de suicidios fue de 15,3 muertes por cada 100.000 en 1995
- Disminuyó de manera constante, a 11.5 por 100k en 2015 (~ 25% de disminución)
- Las tarifas solo ahora están volviendo a las tarifas anteriores a los 90
- Datos limitados en la década de 1980, por lo que es difícil decir si la tasa era realmente representativa de la población mundial.

2.2 Por continente

In [ ]:

```
continent <- data %>%
  group_by(continent) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  arrange(suicide_per_100k)

continent$continent <- factor(continent$continent, ordered = T, levels = continent$continent)

continent_plot <- ggplot(continent, aes(x = continent, y = suicide_per_100k, fill = continent)) +
  geom_bar(stat = "identity") +
  labs(title = "Suicidios globales (por 100k), por Continente",
  x = "Continente",
  y = "Suicidios por 100k",
  fill = "Continente") +
  theme(legend.position = "none", title = element_text(size = 10)) +
  scale_y_continuous(breaks = seq(0, 20, 1), minor_breaks = F)


continent_time <- data %>%
  group_by(year, continent) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000)

continent_time$continent <- factor(continent_time$continent, ordered = T, levels = continent$continent)

continent_time_plot <- ggplot(continent_time, aes(x = year, y = suicide_per_100k, col = factor(continent))) +
  facet_grid(continent ~ ., scales = "free_y") +
  geom_line() +
  geom_point() +
  labs(title = "Tendencia en el tiempo, por continente",
       x = "Año",
       y = "Suicidios por 100k",
       color = "Continente") +
  theme(legend.position = "none", title = element_text(size = 10)) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F)

grid.arrange(continent_plot, continent_time_plot, ncol = 2)
```

Algunos hechos que destacan

- Tasa europea más alta en general, pero ha disminuido constantemente ~ 40% desde 1995
- La tasa europea para 2015 similar a Asia y Oceanía
- La línea de tendencia para África se debe a la mala calidad de los datos: solo 3 países han proporcionado datos
- Las tendencias de Oceanía y América son más preocupantes

2.3 Por sexo

In [ ]:

```
sex_plot <- data %>%
  group_by(sex) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
ggplot(aes(x = sex, y = suicide_per_100k, fill = sex)) +
  geom_bar(stat = "identity") +
  labs(title = "Global suicides (per 100k), by Sex",
       x = "Sex",
       y = "Suicides per 100k") +
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0, 25), minor_breaks = F)

### with time
sex_time_plot <- data %>%
  group_by(year, sex) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = factor(sex))) +
  facet_grid(sex ~ ., scales = "free_y") +
  geom_line() +
  geom_point() +
  labs(title = "Trends Over Time, by Sex",
       x = "Year",
       y = "Suicides per 100k",
       color = "Sex") +
  theme(legend.position = "none") +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F)

grid.arrange(sex_plot, sex_time_plot, ncol = 2)
```

Algunos hechos que destacan:

- A nivel mundial, la tasa de suicidio de los hombres ha sido ~ 3,5 veces mayor para los hombres

- Las tasas de suicidio de hombres y mujeres alcanzaron su punto máximo en 1995, disminuyendo desde
- Esta proporción de 3,5: 1 (hombre: mujer) se ha mantenido relativamente constante desde mediados de los 90
- Sin embargo, durante los años 80, esta proporción era tan baja como 2,7: 1 (hombre: mujer)

2.4 Por edad

In [ ]:

```
age_plot <- data %>%
  group_by(age) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = age, y = suicide_per_100k, fill = age)) +
  geom_bar(stat = "identity") +
  labs(title = "Global suicides per 100k, by Age",
       x = "Age",
       y = "Suicides per 100k") +
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0, 30, 1), minor_breaks = F)

### with time
age_time_plot <- data %>%
  group_by(year, age) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = age)) +
  facet_grid(age ~ ., scales = "free_y") +
  geom_line() +
  geom_point() +
  labs(title = "Trends Over Time, by Age",
       x = "Year",
       y = "Suicides per 100k",
       color = "Age") +
  theme(legend.position = "none") +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F)


grid.arrange(age_plot, age_time_plot, ncol = 2)
```

Algunos hechos que destacan:

- A nivel mundial, la probabilidad de suicidio aumenta con la edad
- Desde 1995, la tasa de suicidio para todas las personas mayores de 15 años ha ido disminuyendo linealmente

- La tasa de suicidios de las personas mayores de 75 años se ha reducido en más del 50% desde 1990
- La tasa de suicidios en la categoría "5-14" permanece prácticamente estática y pequeña (<1 por 100.000 por año)

2.5 Por País

2.5.1 General

In [ ]:

```
country <- data %>%
  group_by(country, continent) %>%
  summarize(n = n(),
            suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  arrange(desc(suicide_per_100k))

country$country <- factor(country$country,
                          ordered = T,
                          levels = rev(country$country))

ggplot(country, aes(x = country, y = suicide_per_100k, fill = continent)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = global_average, linetype = 2, color = "grey35", size = 1) +
  labs(title = "Global suicides per 100k, by Country",
       x = "Country",
       y = "Suicides per 100k",
       fill = "Continent") +
  coord_flip() +
  scale_y_continuous(breaks = seq(0, 45, 2)) +
  theme(legend.position = "bottom")
```

Algunos hechos que destacan:

- La tasa de Lituania ha sido la más alta por un amplio margen:> 41 suicidios por cada 100.000 (por año)
- Destacan países europeos con tasas altas, pocos con tasas bajas

Es importante tener en cuenta que, por estos motivos, es posible que mirar las cifras a nivel mundial / continental no sea realmente representativo del mundo / continente.

La comparación de las tasas brutas de suicidio de los países también puede dar lugar a algunos problemas: la definición de suicidio (y la fiabilidad de que una muerte se registra como suicidio) probablemente variará entre países.

Sin embargo, es probable que las tendencias a lo largo del tiempo (dentro de los países) sean confiables.

2.5.2 Paises con tendencia

Interesa saber cómo está cambiando la tasa de suicidios con el tiempo en cada país. En lugar de visualizar las tasas de los 93 países a lo largo del tiempo, ajusto una regresión lineal simple a los datos de cada país, se extraen aquellos con un $p-$valor de "año" (corregido para comparaciones múltiples) de <0.05.

En otras palabras: a medida que pasa el tiempo, se buscan países donde la tasa de suicidios aumenta o disminuye linealmente con el tiempo. Estos pueden ordenarse por rango por su coeficiente de "año", que sería su tasa de cambio a medida que pasa el tiempo.

In [ ]:
```
country_year <- data %>%
  group_by(country, year) %>%
  summarize(suicides = sum(suicides_no),
            population = sum(population),
            suicide_per_100k = (suicides / population) * 100000,
            gdp_per_capita = mean(gdp_per_capita))


country_year_trends <- country_year %>%
  ungroup() %>%
  nest(-country) %>% # format: country, rest of data (in list column)
  mutate(model = map(data, ~ lm(suicide_per_100k ~ year, data = .)), # for each item in 'data', fit a linear model
         tidied = map(model, tidy)) %>% # tidy each of these into dataframe format - call this list 'tidied'
  unnest(tidied)

country_year_sig_trends <- country_year_trends %>%
  filter(term == "year") %>%
  mutate(p.adjusted = p.adjust(p.value, method = "holm")) %>%
  filter(p.adjusted < .05) %>%
  arrange(estimate)

country_year_sig_trends$country <- factor(country_year_sig_trends$country,
                                          ordered = T,
                                          levels = country_year_sig_trends$country)
```

In [ ]:
```
# plot 1
ggplot(country_year_sig_trends, aes(x=country, y=estimate, col = estimate)) +
```

```
  geom_point(stat='identity', size = 4) +
  geom_hline(yintercept = 0, col = "grey", size = 1) +
  scale_color_gradient(low = "green", high = "red") +
  geom_segment(aes(y = 0,
                   x = country,
                   yend = estimate,
                   xend = country), size = 1) +
  labs(title="Change per year (Suicides per 100k)",
       subtitle="Of countries with significant trends (p < 0.05)",
       x = "Country", y = "Change Per Year (Suicides per 100k)") +
  scale_y_continuous(breaks = seq(-2, 2, 0.2), limits = c(-1.5, 1.5)) +
  theme(legend.position = "none") +
  coord_flip()
```

Algunos hechos que destacan

- ~ 1/2 de todos los países, las tasas de suicidio están cambiando linealmente a medida que pasa el tiempo
- 32 (2/3) de estos 48 países están disminuyendo
- En general, esto es una imagen positiva.

In [ ]:

```
### Lets look at those countries with the steepest increasing trends

top12_increasing <- tail(country_year_sig_trends$country, 12)

country_year %>%
  filter(country %in% top12_increasing) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ country) +
  theme(legend.position = "none") +
  labs(title="12 Steepest Increasing Trends",
       subtitle="Of countries with significant trends (p < 0.05)",
       x = "Year",
       y = "Suicides per 100k")
```

Algunos hechos que destacan

- Corea del Sur muestra la tendencia más preocupante: un aumento en el suicidio de 0,931 personas (por cada 100 mil, por año), el aumento más pronunciado a nivel mundial
- Guyana es similar, con + 0.925 personas (por 100k, por año)
- Entre 1998 y 1999 (5,3 a 24,8), la tasa de Guyana aumentó en ~ 365%
- Los datos históricos de Guyana parecen cuestionables: es conocido por tasas de suicidio muy altas, pero el salto parece poco probable (¿tal vez cambió la forma en que clasificaron el suicidio?)

In [ ]:
```
### Now those with the steepest decreasing trend

top12_decreasing <- head(country_year_sig_trends$country, 12)

country_year %>%
  filter(country %in% top12_decreasing) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ country) +
  theme(legend.position = "none") +
  labs(title="12 Steepest Decreasing Trends",
       subtitle="Of countries with significant trends (p < 0.05)",
       x = "Year",
       y = "Suicides per 100k")
```

Algunos hechos que destacan:

- Estonia muestra la tendencia más positiva: cada año, aproximadamente 1,31 personas menos (por cada 100.000) se suicidan, la disminución más pronunciada a nivel mundial
- Entre 1995 y 2015, esto se reduce de 43,8 a 15,7 por 100.000 (por año), una disminución del 64%.
- La tendencia de la Federación de Rusia es interesante, que recién comenzó a descender en 2002. Desde entonces ha disminuido en aproximadamente un 50%.

2.6 Diferencias de género por continente

In [ ]:
```
data %>%
  group_by(continent, sex) %>%
  summarize(n = n(),
```

```
            suicides = sum(as.numeric(suicides_no)),
            population = sum(as.numeric(population)),
            suicide_per_100k = (suicides / population) * 100000) %>%
  ggplot(aes(x = continent, y = suicide_per_100k, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_hline(yintercept = global_average, linetype = 2, color = "grey35", size = 1) +
  labs(title = "Gender Disparity, by Continent",
   x = "Continent",
   y = "Suicides per 100k",
   fill = "Sex") +
  coord_flip()
```

Algunos hechos que destacan

- Los hombres europeos tuvieron el mayor riesgo entre 1985 y 2015, con ~ 30 suicidios (por cada 100.000, por año)
- Asia tuvo la sobrerrepresentación más pequeña de suicidios masculinos: la tasa fue ~ 2.5 veces más alta para los hombres
- Comparativamente, la tasa de Europa fue ~ 3,9 veces más alta para los hombres

2.7 Diferencias de género por país

In [ ]:

```
country_long <- data %>%
  group_by(country, continent) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  mutate(sex = "OVERALL")

### by country, continent, sex

sex_country_long <- data %>%
  group_by(country, continent, sex) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000)


sex_country_wide <- sex_country_long %>%
  spread(sex, suicide_per_100k) %>%
  arrange(Male - Female)


sex_country_wide$country <- factor(sex_country_wide$country,
                                   ordered = T,
```

```
                                        levels = sex_country_wide$country)

sex_country_long$country <- factor(sex_country_long$country,
                                   ordered = T,
                                   levels = sex_country_wide$country) # using the same order



### this graph shows us how the disparity between deaths varies across gender for every country
# it also has the overall blended death rate - generally countries with a higher death rate have a higher disparity
# this is because, if suicide is more likely in a country, the disparity between men and women is amplified

ggplot(sex_country_wide, aes(y = country, color = sex)) +
  geom_dumbbell(aes(x=Female, xend=Male), color = "grey", size = 1) +
  geom_point(data = sex_country_long, aes(x = suicide_per_100k), size = 3) +
  geom_point(data = country_long, aes(x = suicide_per_100k)) +
  geom_vline(xintercept = global_average, linetype = 2, color = "grey35", size = 1) +
  theme(axis.text.y = element_text(size = 8),
        legend.position = c(0.85, 0.2)) +
  scale_x_continuous(breaks = seq(0, 80, 10)) +
  labs(title = "Gender Disparity, by Continent & Country",
       subtitle = "Ordered by difference in deaths per 100k.",
       x = "Suicides per 100k",
       y = "Country",
       color = "Sex")
```

Algunos hechos que destacan:

- La sobrerrepresentación de hombres en las muertes por suicidio parece ser universal y se puede observar en diferentes grados en cada país.
- Mientras que las mujeres tienen más probabilidades de sufrir depresión y pensamientos suicidas, los hombres tienen más probabilidades de morir por suicidio.
- Esto se conoce como la paradoja de género en la conducta suicida.

2.8 Diferencias de edad por continente

In [ ]:

```
data %>%
  group_by(continent, age) %>%
  summarize(n = n(),
```

```
            suicides = sum(as.numeric(suicides_no)),
            population = sum(as.numeric(population)),
            suicide_per_100k = (suicides / population) * 100000) %>%
  ggplot(aes(x = continent, y = suicide_per_100k, fill = age)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_hline(yintercept = global_average, linetype = 2, color = "grey35", size = 1) +
  labs(title = "Age Disparity, by Continent",
       x = "Continent",
       y = "Suicides per 100k",
       fill = "Age")
```

Algunos hechos que destacan:

- Para las Américas, Asia y Europa (que constituyen la mayor parte del conjunto de datos), la tasa de suicidio aumenta con la edad
- Las tasas de Oceanía y África son más altas para las personas de 25 a 34 años

2.9 Cuándo el país se hace más rico, ¿qué pasa con la tasa de suicidio?

Depende del país: para casi todos los países, existe una alta correlación entre el año y el pib per cápita, es decir, a medida que pasa el tiempo, el pib per cápita aumenta linealmente.

Al calcular las correlaciones de Pearson entre "año" y "PIB per cápita" dentro de cada país, se obtienen los resultados:

La correlación media fue 0,878, lo que indica una relación lineal positiva muy fuerte.

Básicamente, esto significa que mirar dentro de un país y preguntar "¿un aumento en el clima (por persona) tiene un efecto en la tasa de suicidio" es bastante similar a preguntar "la tasa de suicidio de un país aumenta a medida que pasa el tiempo?".

Esto fue se resolvió antes, **¡depende del país!**. Algunos países aumentan con el tiempo, la mayoría disminuye.

Una pregunta ligeramente diferente a continuación.

2.10 ¿Países ricos tienen tasas de suicidio más altas?

En lugar de mirar las tendencias dentro de los países, se toma cada país y se calcula su PIB medio (per cápita) a lo largo de todos los años en los que hay datos disponibles. Luego se mide cómo esto se relaciona con la tasa de suicidios del país durante todos esos años.

El resultado final es un punto de datos por país, destinado a dar una idea general de la riqueza de un país y su tasa de suicidios.

In [ ]:

```
country_mean_gdp <- data %>%
  group_by(country, continent) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000,
            gdp_per_capita = mean(gdp_per_capita))

ggplot(country_mean_gdp, aes(x = gdp_per_capita, y = suicide_per_100k, col = continent)) +
  geom_point() +
  scale_x_continuous(labels=scales::dollar_format(prefix="$"), breaks = seq(0, 70000, 10000)) +
  labs(title = "Correlation between GDP (per capita) and Suicides per 100k",
       subtitle = "Plot containing every country",
       x = "GDP (per capita)",
       y = "Suicides per 100k",
       col = "Continent")
```

Hay bastantes países de alto leverage y residuales que podrían tener un impacto significativo en el ajuste de la línea de regresión (por ejemplo, Lituania, arriba a la izquierda). Se identificarán usando Distancia de Cook, excluyendo aquellos países con un valor de CooksD superior a $4/n$.

Se evalúan las estadísticas de este modelo (con los valores atípicos eliminados) a continuación.

In [ ]:

```
model1 <- lm(suicide_per_100k ~ gdp_per_capita, data = country_mean_gdp)

gdp_suicide_no_outliers <- model1 %>%
  augment() %>%
  arrange(desc(.cooksd)) %>%
  filter(.cooksd < 4/nrow(.)) %>% # removes 5/93 countries
  inner_join(country_mean_gdp, by = c("suicide_per_100k", "gdp_per_capita")) %>%
  select(country, continent, gdp_per_capita, suicide_per_100k)

model2 <- lm(suicide_per_100k ~ gdp_per_capita, data = gdp_suicide_no_outliers)

summary(model2)
```

El valor p del modelo es 0.0288 <0.05. Esto significa que podemos rechazar la hipótesis de que el PIB de un país (per cápita) no tiene asociación con su tasa de suicidio (por 100 mil).

El R-cuadrado es 0.0544, por lo que el PIB (per cápita) explica muy poco de la variación en la tasa de suicidios en general.

¿Qué significa todo esto?

Existe una relación lineal positiva débil pero significativa: los países más ricos están asociados con tasas más altas de suicidio, pero esta es una relación débil que se puede ver en el gráfico siguiente.

In [ ]:
```
ggplot(gdp_suicide_no_outliers, aes(x = gdp_per_capita, y = suicide_per_100k, col = continent)) +
  geom_point() +
  geom_smooth(method = "lm", aes(group = 1)) +
  scale_x_continuous(labels=scales::dollar_format(prefix="$"), breaks = seq(0, 70000, 10000)) +
  labs(title = "Correlation between GDP (per capita) and Suicides per 100k",
       subtitle = "Plot with high CooksD countries removed (5/93 total)",
       x = "GDP (per capita)",
       y = "Suicides per 100k",
       col = "Continent") +
  theme(legend.position = "none")
```

El modelo que se ajusta es el siguiente

Suicides = Suicidios por 100k habitantes GDP = PIB per capita (en miles de USD)

$$Suicides = 8.7718 + 0.1115 * GDP$$

Esto significa que, a nivel de país y durante el período de tiempo de este análisis (1985-2015), un aumento del PIB (per cápita) de $ 8,967 se asoció con 1 suicidio adicional, por cada 100.000 personas, por año.

3. Comparando el Reino Unido, Irlanda, Estados Unidos, Francia y Dinamarca

Creo que sería útil comparar algunos países que la gente podría pensar que son similares al Reino Unido (cultural, legal y económicamente).

3.1. Tendencia General

In [ ]:
```
data_filtered <- data %>%
  filter(country %in% c("United Kingdom",
                        "Ireland",
                        "United States",
                        "France",
                        "Denmark"))
```

```
data_filtered %>%
  group_by(country, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  labs(title = "UK, Ireland, US, France & Denmark",
       subtitle = "Suicides per 100k population, 1985 - 2015",
       x = "Year",
       y = "Suicides per 100k",
       col = "Country")
```

Algunos hechos que destacan

- La tasa de suicidios del Reino Unido ha sido consistentemente más baja desde 1990, y se ha mantenido bastante estática desde ~ 1995
- Francia ha tenido históricamente la tasa más alta, pero ahora es aproximadamente igual a Estados Unidos
- Estados Unidos tiene la tendencia más preocupante, aumentando linealmente en ~ 1/3 desde 2000

3.2. Por sexo

In [ ]:

```
data_filtered %>%
  group_by(country, sex, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  facet_wrap(~ sex, scales = "free_y", nrow = 2) +
  labs(title = "UK, Ireland, US, France & Denmark",
       subtitle = "Suicides per 100k population, 1985 - 2015",
       x = "Year",
       y = "Suicides per 100k",
       col = "Country")
```

Algunos hechos que destacan:

- Para el Reino Unido, no hay un aumento obvio en la tasa de suicidios para los hombres que no se pueda observar en igual medida en las mujeres.
- Nuevamente, en el caso de hombres y mujeres, Francia ha disminuido hasta llegar a ser aproximadamente igual a EE. UU. En 2015
- Las diferentes líneas de tendencia para hombres y mujeres en Irlanda son inusuales: en 1990, la tasa de hombres aumenta, pero no se puede observar lo mismo para las mujeres.

3.3. Únicamente entre 2010 y 2015

Considerando los anteriores hallazgos, estamos realmente más interesados en los datos de los últimos años (Francia, por ejemplo, ha cambiado mucho), por lo que se restringe el período de tiempo a 2010 en adelante.

3.3.1. Proporción de suicidios que son de hombres

In [ ]:

```
t1 <- data_filtered %>%
  filter(year >= 2010) %>%
  group_by(sex) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000)

global_male_proportion <- t1$suicide_per_100k[2] / sum(t1$suicide_per_100k)


t2 <- data_filtered %>%
  filter(year >= 2010, continent == "Europe") %>%
  group_by(sex) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000)

european_male_proportion <- t2$suicide_per_100k[2] / sum(t2$suicide_per_100k)


data_filtered %>%
  filter(year >= 2010) %>%
  group_by(country, sex) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = country, y = suicide_per_100k, fill = sex)) +
  geom_bar(position = "fill", stat = "identity") +
  geom_hline(yintercept = global_male_proportion) +
  geom_hline(yintercept = european_male_proportion, col = "blue") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proportion of suicides that were Male & Female",
```

```
        subtitle = "2010 - 2015 only, with reference lines for Europe (blue) & Globally (black)",
        x = "Country",
        y = "",
        fill = "Sex")
```

Algunos hechos que destacan;

- Patrón similar al observado a lo largo del análisis: los hombres representan ~ 75% de las muertes por suicidio
- La mayor proporción se encuentra en Irlanda: 81,7% hombres
- La proporción más baja es para Dinamarca: 73,5% hombres

3.3.2. Tasas por edad

In [ ]:

```
data_filtered %>%
  filter(year >= 2010) %>%
  group_by(country, age) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = country, y = suicide_per_100k, fill = age)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Age ",
       subtitle = "2010 - 2015 only",
       x = "Country",
       y = "Suicides per 100k",
       fill = "Age")
```

Algunos hechos que destacan

- Existe una gran diferencia en la "tendencia" de las tasas de suicidio ya que la edad varía dentro de cada país
- La tasa de suicidio aumenta con la edad en Francia, Dinamarca y EE. UU. (En menor medida)
- Los de 35 a 54 años tienen mayor riesgo en Irlanda y el Reino Unido, que siguen más cerca de una distribución gaussiana

3.3.2. Tasas por genero para varias edades

In [ ]:

```
data_filtered %>%
  filter(year >= 2010) %>%
  group_by(country, sex, age) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = age, y = suicide_per_100k, fill = country)) +
```

```
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ sex, scales = "free_x") +
  labs(title = "Age Disparity, by Country",
       subtitle = "2010 - 2015 only",
       x = "Age",
       y = "Suicides per 100k",
       fill = "Country") +
  coord_flip() +
  theme(legend.position = "bottom")
```

Algunos hechos que destacan

- En los EE. UU., La tasa de suicidios de los hombres sigue aumentando con la edad, pero la tasa de mujeres disminuye en la vejez
- Esta extraña disparidad solo está presente en los EE. UU. Y tengo curiosidad por saber por qué ocurre
- El Reino Unido tiene la tasa de suicidio más baja o la segunda más baja de todos los grupos de edad y sexo

3.4. Tasas por genero para varias edades

Existe una gran preocupación en Reino Unido con respecto a los problemas de salud mental y el suicidio de los hombres jóvenes y de mediana edad. Se restringi el análisis a solo hombres con ddades "15-24", "25-34" y "35-54". Básicamente, observamos si existen tendencias preocupantes. Tener otros países aquí para comparar será útil y ayudará a proporcionar una perspectiva en el análisis.

3.4.1. Hombres entre 15 y 64

In [ ]:

```
data_filtered %>%
  filter(age %in% c("15-24", "25-34", "35-54"), sex == "Male") %>%
  group_by(country, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  labs(title = "UK, Ireland, US, France & Denmark",
       subtitle = "Suicides per 100k population, 1985 - 2015",
       x = "Year",
       y = "Suicides per 100k",
       col = "Country")
```

Algunos hechos que destacan

- La tendencia de Irlanda durante la década de 1990 fue muy preocupante
- Pasó de 14 (por 100k, por año) a 33,3 entre 1988 y 1998, un aumento del 138%
- Nuevamente, EE. UU. Muestra la tendencia actual más obvia y preocupante
- Comparativamente, para los hombres jóvenes y de mediana edad, el Reino Unido parece bastante plano a lo largo del tiempo.

3.4.2. Hombres entre 15 y 64

In [ ]:

```
data_filtered %>%
  filter(age %in% c("15-24", "25-34", "35-54"), sex == "Male") %>%
  group_by(country, age, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  facet_wrap(~ age, nrow = 3, scales = "free_y") +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  labs(title = "UK, Ireland, US, France & Denmark",
       subtitle = "Suicides per 100k population, 1985 - 2015",
       x = "Year",
       y = "Suicides per 100k",
       col = "Country")
```

Algunos hechos que destacan

- En el caso de los hombres del Reino Unido, solo la categoría "35-54" parece estar aumentando, con un ligero aumento (~ 10-15%) durante la última década.
- Las tasas del Reino Unido para hombres en las categorías "15-24" y "25-34" parecen planas y ligeramente decrecientes, respectivamente
- Ha sido bastante difícil describir estas tendencias del Reino Unido, lo que creo que es algo positivo. La comparación con otros países definitivamente ayuda con la perspectiva.

Conclusiones Generales

- Las tasas de suicidio están disminuyendo a nivel mundial.

- De los países que muestran claras tendencias lineales a lo largo del tiempo, 2/3 están disminuyendo.
- En promedio, la tasa de suicidios aumenta con la edad.
- Esto sigue siendo cierto cuando se controla por continente en América, Asia y Europa, pero no para África y Oceanía.
- Existe una débil relación positiva entre el PIB de un país (per cápita) y la tasa de suicidios.
- La tasa de suicidio más alta jamás registrada en un grupo demográfico (durante 1 año) es 225 (por cada 100.000 habitantes).
- Existe una representación excesiva de hombres en las muertes por suicidio en todos los niveles de análisis (a nivel mundial, a nivel de continente y de país). A nivel mundial, la tasa masculina es ~ 3,5 veces mayor.

In [ ]: