ETC 2420/5242 Lab 5 2017

Di Cook

Week 5

Purpose

This lab is to practice fitting linear models.

Reading

- Read the material on fitting regression models in Statistics online textbook, Diez, Barr, Cetinkaya-Rundel.
- Read the code in the lecture notes of the second class from Week 4.
- Watch the movie Hans Rosling's TED talk

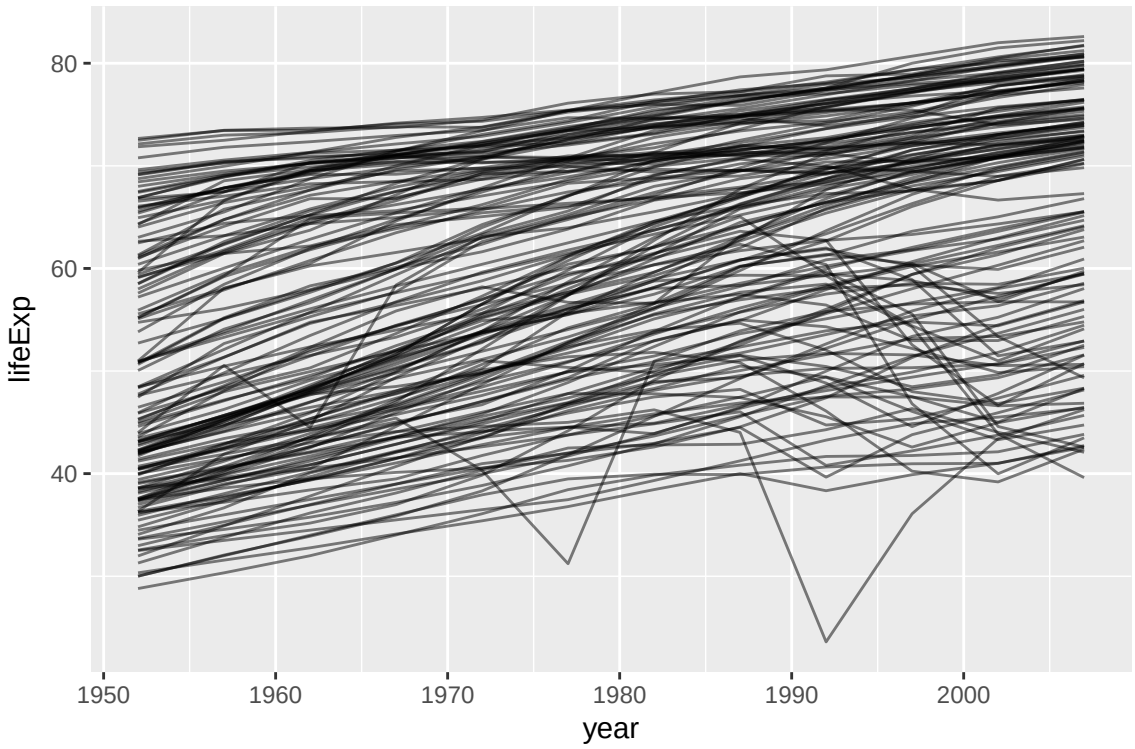
Warmup exercises

- We are going to take a look at the gapminder data used in the first lab.
- The data has demographics of life expectancy and GDP per capita for 142 countries reported every 5 years between 1952 and 2007.

```
library("tidyverse")
library("gapminder")
glimpse(gapminder)
# Observations: 1,704
# Variables: 6
# $ country   <fctr> Afghanistan, Afghanistan, Afghanistan, Afghanistan,...
# $ continent <fctr> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asi...
# $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992...
# $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.8...
# $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 1488...
# $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 78...
```

- How would you describe the following plot?

```
ggplot(data=gapminder, aes(x=year, y=lifeExp, group=country)) +
  geom_line(alpha=0.5)
```

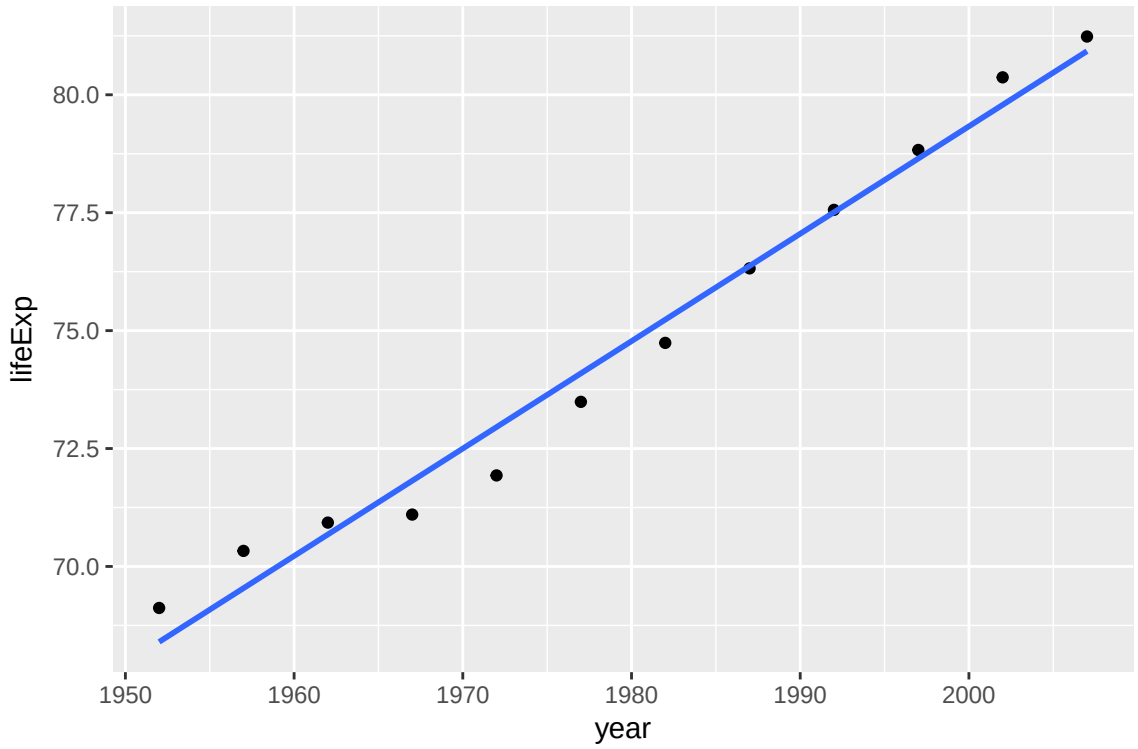


- 1950 is the first year, so for model fitting we are going to shift year to begin in 1950, makes interpretability easier.

```
gapminder2 <- gapminder %>% mutate(year1950 = year-1950)
```

- Then let's fit a model for Australia

```
oz <- gapminder2 %>% filter(country=="Australia")
head(oz)
# # A tibble: 6 x 7
#     country continent  year lifeExp      pop gdpPercap year1950
#      <fctr>    <fctr> <int>   <dbl>    <int>     <dbl>    <dbl>
# 1 Australia   Oceania  1952   69.12  8691212  10039.60        2
# 2 Australia   Oceania  1957   70.33  9712569  10949.65        7
# 3 Australia   Oceania  1962   70.93 10794968  12217.23       12
# 4 Australia   Oceania  1967   71.10 11872264  14526.12       17
# 5 Australia   Oceania  1972   71.93 13177000  16788.63       22
# 6 Australia   Oceania  1977   73.49 14074100  18334.20       27
ggplot(data=oz, aes(x=year, y=lifeExp)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```




```
oz_lm <- lm(lifeExp~year1950, data=oz)
oz_lm
#
# Call:
# lm(formula = lifeExp ~ year1950, data = oz)
#
# Coefficients:
# (Intercept)     year1950
#     67.9451       0.2277
```

- Interpret the model. (This means explain how life expectancy changes over years, since 1950, using the parameter estimates of the model.)
- What was the average life expectancy in 1950?
- What was the average life expectancy in 2000?
- By how much did average life expectancy change over those 50 years?
- We can get various diagnostics out for the model with the `broom` package: the parameter estimates and their significance, the goodness of fit statistics, and model diagnostics.

```
library("broom")
oz_coef <- tidy(oz_lm)
oz_coef
#          term   estimate  std.error statistic      p.value
# 1 (Intercept) 67.9450653 0.35475797  191.5251 3.700841e-19
# 2    year1950  0.2277238 0.01037958   21.9396 8.667222e-10
oz_fit <- glance(oz_lm)
oz_fit
#   r.squared adj.r.squared     sigma statistic      p.value df    logLik
# 1 0.9796477     0.9776125 0.6206086  481.3459 8.667222e-10  2 -10.20868
#        AIC      BIC deviance df.residual
```

```
# 1 26.41735 27.87207  3.85155          10
oz_diag <- augment(oz_lm)
oz_diag
#    lifeExp year1950  .fitted   .se.fit      .resid       .hat    .sigma
# 1   69.120        2 68.40051 0.3370035  0.71948718 0.29487179 0.5885398
# 2   70.330        7 69.53913 0.2943424  0.79086830 0.22494172 0.5816212
# 3   70.930       12 70.67775 0.2551280  0.25224942 0.16899767 0.6476436
# 4   71.100       17 71.81637 0.2212012 -0.71636946 0.12703963 0.6021888
# 5   71.930       22 72.95499 0.1953366 -1.02498834 0.09906760 0.5462421
# 6   73.490       27 74.09361 0.1810238 -0.60360723 0.08508159 0.6194376
# 7   74.740       32 75.23223 0.1810238 -0.49222611 0.08508159 0.6312890
# 8   76.320       37 76.37084 0.1953366 -0.05084499 0.09906760 0.6539352
# 9   77.560       42 77.50946 0.2212012  0.05053613 0.12703963 0.6539304
# 10  78.830       47 78.64808 0.2551280  0.18191725 0.16899767 0.6507881
# 11  80.370       52 79.78670 0.2943424  0.58329837 0.22494172 0.6157712
# 12  81.235       57 80.92532 0.3370035  0.30967949 0.29487179 0.6425249
#         .cooksd  .std.resid
# 1  0.3985452403  1.38061071
# 2  0.3040493566  1.44750213
# 3  0.0202148732  0.44587310
# 4  0.1110602709 -1.23544103
# 5  0.1664636939 -1.74002323
# 6  0.0480744299 -1.01682325
# 7  0.0319694063 -0.82919310
# 8  0.0004096170 -0.08631460
# 9  0.0005526989  0.08715392
# 10 0.0105137692  0.32155479
# 11 0.1653930580  1.06759321
# 12 0.0738339980  0.59423827
```

- What column of the diagnostics contains the (a) fitted values, (b) residuals?
- Look up what the `.hat` column contains. Write a couple of sentences about the meaning of this statistic. Plot the values for Australia, as a dotplot. They should all be pretty similar because there are no influential points for this model.

Question 1

The plot of all the countries was really hard to explain. It was very messy. Trying to say something about how life expectancy has changed around the globe was hard.

Now we are going to fit a simple linear model separately to every country. And use the model fits to simplify the patterns across the globe, in order to be able to explain the changes in life expectancy.

This code will compute the models for you:

```
library("purrr")
by_country <- gapminder2 %>%
  select(country, year1950, lifeExp, continent) %>%
  group_by(country, continent) %>%
  nest()
by_country <- by_country %>%
  mutate(
    model = purrr::map(data, ~ lm(lifeExp ~ year1950,
                                  data = .))
```

```
  )
country_coefs <- by_country %>%
  unnest(model %>% purrr::map(broom::tidy))
country_coefs <- country_coefs %>%
  select(country, continent, term, estimate) %>%
  spread(term, estimate)
kable(head(country_coefs))
```

country	continent	(Intercept)	year1950
Afghanistan	Asia	29.35664	0.2753287
Albania	Europe	58.55976	0.3346832
Algeria	Africa	42.23641	0.5692797
Angola	Africa	31.70797	0.2093399
Argentina	Americas	62.22502	0.2317084
Australia	Oceania	67.94507	0.2277238

Here is simpler code using a `for` loop to compute the slope and intercept for each country.

```
n <- length(table(gapminder2$country))
country_coefs <- data.frame(country=gapminder2$country[seq(1, 1704, 12)],
                  continent=gapminder2$continent[seq(1, 1704, 12)],
                  intercept=rep(0,n),
                  year1950=rep(0,n))
for (i in 1:n) {
  sub <- gapminder2 %>% filter(country==country_coefs$country[i])
  sub_lm <- lm(lifeExp~year1950, data=sub)
  sub_lm_coefs <- coefficients(sub_lm)
  country_coefs$intercept[i] <- sub_lm_coefs[1]
  country_coefs$year1950[i] <- sub_lm_coefs[2]
}
kable(head(country_coefs))
```

country	continent	intercept	year1950
Afghanistan	Asia	29.35664	0.2753287
Albania	Europe	58.55976	0.3346832
Algeria	Africa	42.23641	0.5692797
Angola	Africa	31.70797	0.2093399
Argentina	Americas	62.22502	0.2317084
Australia	Oceania	67.94507	0.2277238

- Pick your favorite country (other than Australia). Find the parameter estimates from the `country_coefs` data frame. Do a hand-sketch of the fitted model.

Question 2

a. Make a scatterplot of the linear model estimates for each country, slope vs intercept. Colour the points by continent.

b. Statistically summarise the relationship between intercept and slope, using words like no association, positive linear association, negative linear association, weak, moderate, strong, outliers, clusters.

c. Do you see a difference between continents? If so, explain what you see.

d. What does it mean for a country to have a high intercept, e.g. 70?

e. What does it mean for a country to have a high slope, e.g. 0.7?

f. Make the plot interactive using the `plotly` package, and find out which countries had a negative slope.

Question 3

Now we are going to examine the fit for each country. We might expect that a linear model is a better fit for some countries and not so good for other countries. Here is the code to extract the model diagnostics for each country's model.

```
country_fit <- by_country %>%
  unnest(model %>%
           purrr::map(broom::glance))
```

Or you can use a `for` loop to compute this:

```
n <- length(unique(gapminder2$country))
country_fit <- data.frame(country=gapminder2$country[seq(1, 1704, 12)],
                  continent=gapminder2$continent[seq(1, 1704, 12)],
                  intercept=rep(0,n),
                  year1950=rep(0,n),
                  r.squared=rep(0,n))
for (i in 1:n) {
  sub <- gapminder2 %>% filter(country==country_fit$country[i])
  sub_lm <- lm(lifeExp~year1950, data=sub)
  sub_lm_fit <- coefficients(sub_lm)
  country_fit$intercept[i] <- sub_lm_coefs[1]
  country_fit$year1950[i] <- sub_lm_coefs[2]
  country_fit$r.squared[i] <- summary(sub_lm)$r.squared
}
kable(head(country_fit))
```

country	continent	intercept	year1950	r.squared
Afghanistan	Asia	55.40729	-0.093021	0.9477123
Albania	Europe	55.40729	-0.093021	0.9105778
Algeria	Africa	55.40729	-0.093021	0.9851172
Angola	Africa	55.40729	-0.093021	0.8878146
Argentina	Americas	55.40729	-0.093021	0.9955681
Australia	Oceania	55.40729	-0.093021	0.9796477

a. Plot the R^2 values as a histogram.

b. Statistically describe the distribution using words like symmetric, skewed left, skewed right, unimodal, bimodal, multimodal, outliers.

c. What do you learn about the model fit across all 142 countries?

Question 4

a. Examine the countries with the worst fit, countries with $R^2 < 0.45$, by making scatterplots of the data, with the linear model overlaid.

b. Each of these countries has a big dip in their life expectancy during the time of the study. Explain these using world history and current affairs information. (Feel free to google for news stories.)

c. Use the statistics to investigate something of your choice related to life expectancy across the world over those 50 years.

TURN IN

- Your `.Rmd` and `.html` file
- Make sure your group members are listed as authors, one person per group will turn in the report

Resources

- Statistics online textbook, Diez, Barr, Cetinkaya-Rundel.
- Watch the movie Hans Rosling's TED talk