Essay 2: World Health Data (Solutions Demo)

by Mike DeCrescenzo

April 16, 2019

This document will walk through the code and interpretation of Essay 2.

Prepping R and data: First we attach our packages and import the data.

```
# Packages we will use.
library("tidyverse")
library("here")
library("broom")
library("stargazer")

# here's a fun trick to change the default ggplot theme
theme_set(theme_minimal())
# for themes: https://ggplot2.tidyverse.org/reference/ggtheme.html
# or check out other packages
```

```
# This code will read the data into R *and* trim the variables.
# You can rename variables in the `select()` function.
# It reads like "keep `v9` as the name `life_exp`"
who <- read_csv(here("data", "who2009.csv")) %>%
  select(country, regionname,
         life_exp = v9, inf_mort = v22, hc_work = v159,
         hosp_beds = v168, hc_gdp = v174, pocket_private = v186,
         hc_pc = v192, fertility = v249, gni_pc = v259)
```

1 Linear Regression

```
# create the first figure
ggplot(who, aes(x = inf_mort, y = life_exp)) +
  geom_point(color = "goldenrod") +
  geom_smooth(method = "lm", color = "maroon4", size = 0.5) +
  labs(x = "Infant Mortality\n(Deaths before age 1 per 1,000 live births)",
       y = "Life Expectancy at birth")
```

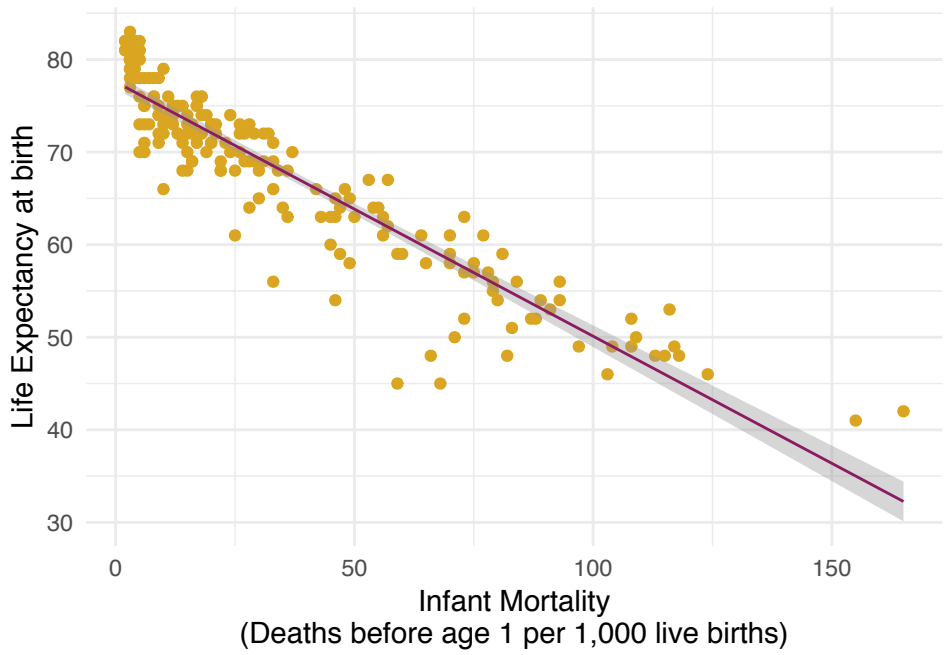

Figure 1: Life Expectancy and Infant Mortality

Figure 1 shows the relationship between life expectancy at birth (y) and the infant mortality rate (x) for countries measured in 2007. The infant mortality rate is measured as the number of deaths for infants less than 1 year of age per 1,000 life births. There is a strong negative relationship between these two variables, meaning that higher infant mortality is associated with lower life expectancy.

```
# first model
lin_mod <- lm(life_exp ~ inf_mort, data = who)
```

Table 1 at the end of the document shows the results of a linear regression of life expectancy on infant mortality (in the first column). The estimated equation is

$$\widehat{\text{lifeExp}} = 77.6 + (-0.275 \times \text{infMort})\,. \tag{1}$$

The constant of 77.6 indicates that if infant mortality were 0, we would predict that life expectancy would be 77.6 *on average*. The coefficient on the infant mortality rate is −0.275. This suggests that as the infant mortality rate increases by 1 death per 1,000 live births, the average life expectancy is predicted to decrease by 0.275 years. In turn, if infant mortality increased by 10 deaths, we would expect a decrease in life expectancy of 2.75 years on average. Assuming a null hypotheses that the true relationship is 0, the probability of observing a relationship this strong by chance would be 1.73×10^{-82}, indicated by the p-value. Because the p-value is so small, we can reject a null hypothesis at conventional levels of statistical significance ($p < .05$).

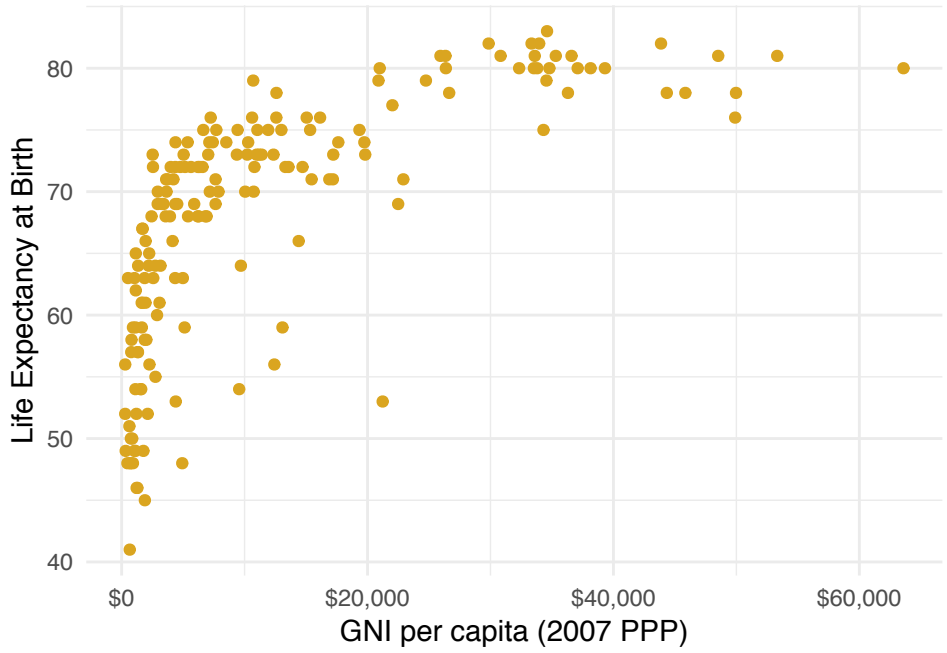

Figure 2: Life Expectancy and GNI per capita (not logged)

2 Nonlinear Regression

```
# without logging.
# BONUS: Transforming the x labels using the "dollar" function
#        from {scales} pkg
ggplot(data = who, aes(x = gni_pc,  y = life_exp)) +
  geom_point(color = "goldenrod") +
  scale_x_continuous(labels = scales::dollar) +
  labs(x = "GNI per capita (2007 PPP)",
       y = "Life Expectancy at Birth")
```

Figure 2 shows life expectancy (y) plotted against gross national income (GNI) per capita (x) measured in U.S. dollars adjusted for international purchasing power parity (PPP). The relationship appears to be nonlinear. As GNI per capita increases, there is a strong initial increase in life expectancy. After a certain point, the relationship begins to weaken. This relationship shows us that as countries become *much* wealthier, their returns on life expectancy begin to diminish on the margins. This pattern is indicative of a logarithmic relationship, which often arise when the x variable is a quantity created by multiplicative processes (such as wealth creation or population growth).[1]

[1] If we plotted a histogram of GNI per capita, we would find a skewed distribution with a "long right tail," which is a common indicator that a variable is the result of a multiplicative process.

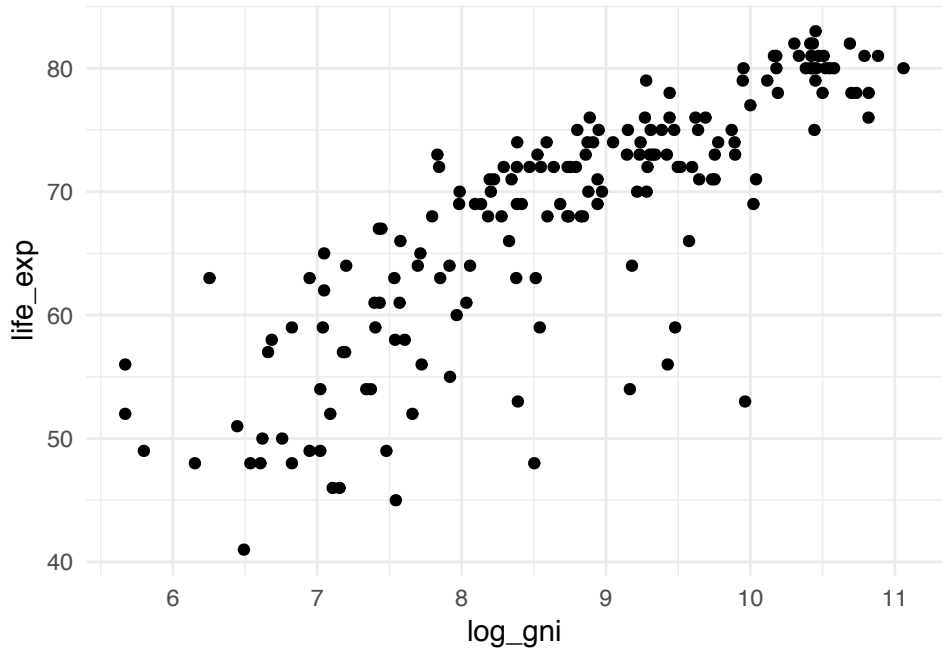

Figure 3: Life Expectancy and (Logged) GNI Per Capita

```
# created logged GNI
who <- who %>%
  mutate(log_gni = log(gni_pc))
```

```
# not making this pretty since it's just for visual inspection
ggplot(data = who, aes(x = log_gni, y = life_exp)) +
  geom_point()
```

In order to estimate the relationship between GNI per capita and life expectancy, we should calculate the log of GNI. Figure 3 shows the relationship between life expectancy and the *log* of GNI per capita. What was a nonlinear relationship became a linear relationship after logging the *x* variable. Because there is a *transformation* of our data that look linear, we can proceed to estimate the linear relationship on the transformed data using a linear regression analysis.

```
# estimate model
log_mod <- lm(life_exp ~ log_gni, data = who)

# generate predictions (yhat ~ log(x))
# compute a conf.interval w/ the std. error of the prediction
# calculate GNI on its original scale for plotting purposes
#   this is: gni = e^(log(gni)), since e^() and log() are inverses!
log_predictions <- augment(log_mod) %>%
```

```
  mutate(MOE = 1.96 * .se.fit,
         conf.low = .fitted - MOE,
         conf.high = .fitted + MOE,
         unlogged_gni = exp(log_gni))
```

When we estimate the linear relationship between life expectancy and logged GNI per capita, the estimated coefficients give us the following equation for the predicted life expectancy:

$$\widehat{\text{lifeExp}} = 11.876 + 6.444 \log(\text{GNIpc}) \tag{2}$$

Table 1 (column 2) also summarizes the results of this regression. The constant tells us that when log(GNIpc) is zero, the average expected life expectancy is just under 12 years. How do we interpret this? For logarithms with any base, $\log_b(1) = 0$, so this is the model's prediction when GNI per capita is only $1, which is beyond the range of our data. The coefficient for log GNI tells us that as the log of GNI increases by 1, life expectancy is predicted to increases by 6.4 years on average. The p-value for the effect of GNI is 1.21×10^{-44}. Because this p-value is far below 0.05, we can again reject a null hypothesis of zero relationship at convention levels of statistical significance, since the p-value tells us the probability of observing a relationship this strong or stronger if the null were true.

```
# plot predictions. Points and line use different Y data,
#   so we have to specify when they should be different
ggplot(log_predictions, aes(x = unlogged_gni, y = life_exp)) +
  geom_ribbon(aes(ymin = conf.low, ymax= conf.high),
              color = "gray", alpha = 0.3) +
  geom_point(color = "goldenrod") +
  geom_line(aes(y = .fitted), color = "maroon4") +
  labs(x = "GNI per capita", y = "Life Expectancy") +
  scale_x_continuous(labels = scales::dollar)
```

Because it is difficult to interpret a one-unit increase in a logged variable, we instead plot the model's predictions over GNI per capita. Figure 4 shows these predictions. The logarithmic model is a decent fit to the data, as it captures the same pattern as before: a strong initial increase in life expectancy as countries get more wealthy, but after a certain amount of wealth, additional wealth doesn't increase life expectancy much more.

3 Multiple Regression

```
hc_mod <- lm(life_exp ~ hc_gdp, data = who)

ggplot(who, aes(x = hc_gdp, y = life_exp)) +
  geom_point() +
```

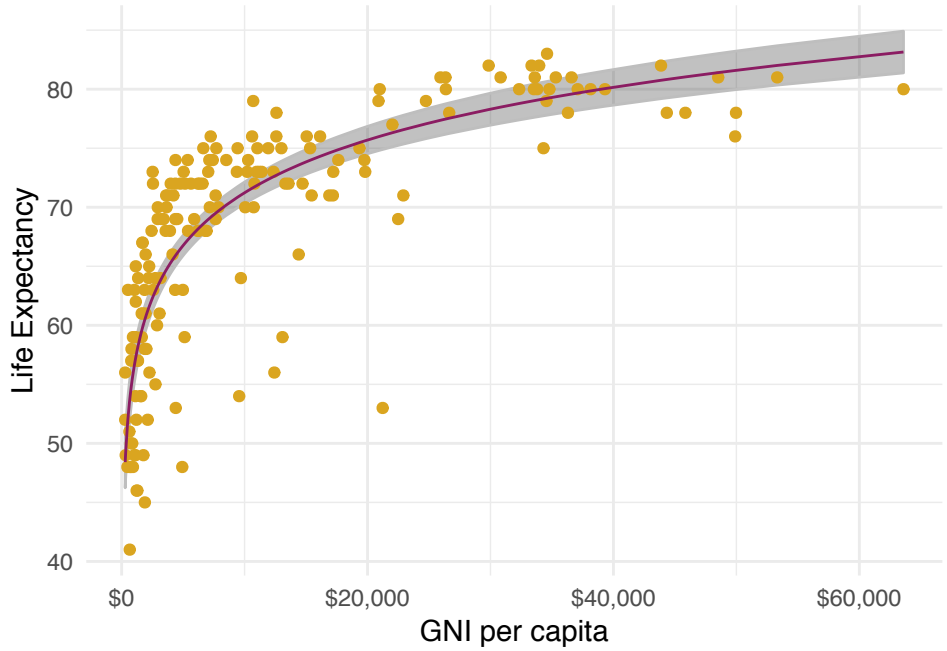

Figure 4: Predicted and actual life expectancy as a function of GNI per capita. Predictions estimated from a regression of y on $\log(x)$

```
  geom_smooth(method = "lm") +
  labs(x = "Health Expenditures (% of 2006 GDP)",
       y = "Life Expectancy")
```

Figure 5 plots life expectancy (y) over health expenditures as a percent of GDP (x). We find only a slight positive relationship that is initially weaker than we might expect. (We summarize this simple regression in the third column of Table 1). Why wouldn't countries that invest in health care have better health outcomes. One reason this could be is that overall national wealth confounds this relationship. Countries that are less wealthy might spend a high share of their GDP on health care, but because they aren't very wealthy to begin with, there isn't ultimately that much money flowing into health care. Meanwhile, countries that are very wealthy might be able to put a lot of resources into health care without it taking up too much of their overall GDP. We should control for national wealth (measured with GNI) and other variables in order to clarify the relationship between the health care share of GDP and life expectancy.

```
multi_mod <- lm(life_exp ~ hc_gdp + log_gni + inf_mort,
                data = who)
```

```
# Put all models into one table.
# omit the F statistic (we never talked about this)
# make pretty variable labels
```

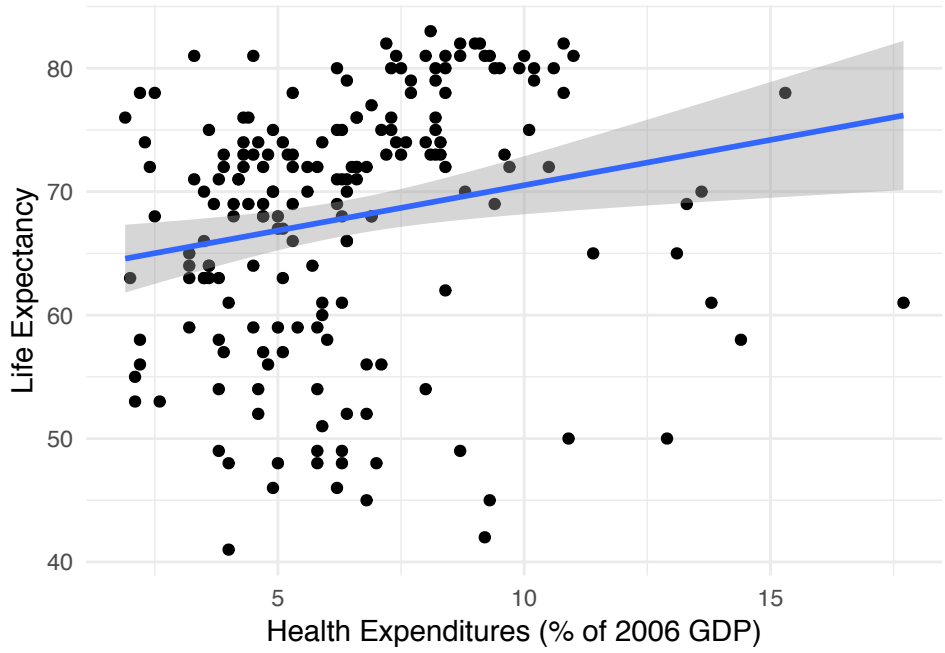

Figure 5: Initial view of life expectancy vs. health care share of GDP

```
stargazer(lin_mod, log_mod, hc_mod, multi_mod,
  type = "text",
  omit.stat = c("F"),
  dep.var.labels = "Life Expectancy",
  covariate.labels =
  c("Inf Mort", "log(GNI)", "Health \\% GDP", "Intercept")
)
```

The final column in Table 1 shows the results of the multiple regression. The estimated equation is

$$\widehat{\text{lifeExp}} = 62.3 + 0.199\text{HealthGDP} + 1.4\log(\text{GNI}) - 0.23\text{InfMort}, \tag{3}$$

which describes the *partial* effect of each variable, holding others constant. The intercept is still outside the range of the data (there are no cases where $\log(GNI) = 0$). As the health share of GDP increases by one percentage point, life expectancy is predicted to increase an average of 0.199 years. This is actually a *weaker* relationship than we initially estimated, where a one-unit increase in health share of GDP predicted an increase in life expectancy of 0.734 years. One-unit increases in log GNI and infant mortality are associated with a 1.4-year increase and a −0.23-year decrease in life expectancy, respectively. The p-values for all variables are below .05, indicating that we would reject the null hypothesis of zero relationship for all predictor variables in the regression.

Table 1:

	Dependent variable:			
	Life Expectancy			
	(1)	(2)	(3)	(4)
Inf Mort	−0.275***			−0.229***
	(0.008)			(0.013)
log(GNI)		6.444***		1.428***
		(0.336)		(0.352)
Health % GDP			0.734***	0.199*
			(0.264)	(0.104)
Intercept	77.578***	11.876***	63.180***	62.370***
	(0.401)	(2.959)	(1.839)	(3.489)
Observations	193	175	191	174
R^2	0.857	0.680	0.039	0.885
Adjusted R^2	0.856	0.678	0.034	0.883
Residual Std. Error	3.881 (df = 191)	5.691 (df = 173)	10.023 (df = 189)	3.430 (df = 170)
Note:				*p<0.1; **p<0.05; ***p<0.01

We can generate predictions from the model by setting our input variables to any values that we want. If we want to isolate the effect of the health share of GDP, we generate a series of predictions that reflect variation in the health share of GDP only; infant mortality and log GNI are held constant at their means. We do this by starting with the original `who` data, setting some variables to be constant, and then using the modified data to generate model predictions from `augment()`.

```
# there are missing data in the original variables
# must use na.rm = TRUE to calculate mean
multi_preds <- who %>%
  mutate(log_gni = mean(log_gni, na.rm = TRUE),
         inf_mort = mean(inf_mort, na.rm = TRUE)) %>%
  select(life_exp, hc_gdp, log_gni, inf_mort) %>%
  augment(multi_mod, newdata = .) %>%
  mutate(MOE = 1.96 * .se.fit,
         conf.low = .fitted - MOE,
         conf.high = .fitted + MOE)
```

```
ggplot(data = who, aes(x = hc_gdp, y = life_exp)) +
  geom_smooth(method = "lm", color = "black",
              size = 0.5) +
  geom_ribbon(data = multi_preds,
              aes(ymin = conf.low, ymax = conf.high),
              fill = "maroon4", alpha = 0.3) +
  geom_line(data = multi_preds, aes(y = .fitted),
            color = "maroon4") +
  labs(x = "Health Share of GDP",
       y = "Predicted Life Expectancy") +
  coord_cartesian(ylim = c(55, 85)) +
  annotate(geom = "text", x = 7, y = 75,
           label = "Simple model") +
  annotate(geom = "text", x = 12, y = 65,
           label = "Multiple regression", color = "maroon4")
```

I plot two sets of predictions alongside one another in Figure 6. First (black) is the prediction from a model where life expectancy is predicted as a function of the health share of GDP *only*. Then I plot the predictions from the multiple regression controlling for log GNI and infant mortality (purple). By plotting the predictions for both models side-by-side, we can see that the multiple regression model estimates a weaker effect of the health share of GDP. In other words, the bigger model thinks that log GNI and infant mortality are better at explaining life expectancy than the health share of GDP, which is why its effect shrinks when we include the other variables in the model despite the fact that the adjusted R^2 value increases from .034 in the simpler model to .883 in the bigger model. (Interestingly, this model explains barely

Figure 6: Predicted effect of health share of GDP, simple vs. multiple regression models

any more variation in life expectancy than the infant-mortality-only model, which has an adjusted R^2 of .856).