L06: Intro to Linear Regression

L06: Intro to Linear Regression

Statistics is Everywhere

Regression

Fitting a linear model in R

Add the regression line to the scatter plot using `geom_abline()`

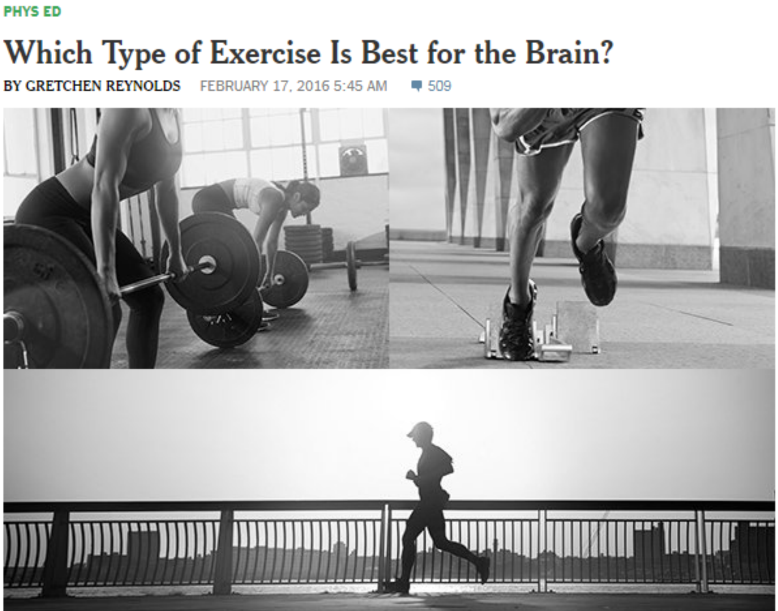
Transforming data

How do outliers affect the line of best fit?

Counfounding

Statistics is Everywhere

Excercise and the Brain



PHYS ED

Which Type of Exercise Is Best for the Brain?

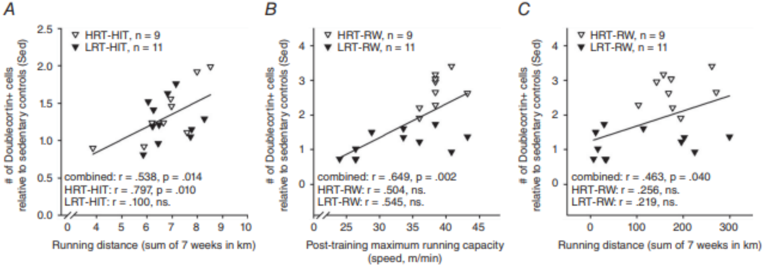
BY GRETCHEN REYNOLDS FEBRUARY 17, 2016 5:45 AM 509

Excercise and the Brain

- from *The New York Times*, February 2016:
  "Some forms of exercise may be much more effective than others at bulking up the brain, according to a remarkable new study in rats. For the first time, scientists compared head-to-head the neurological impacts of different types of exercise: running, weight training and high-intensity interval training. The surprising results suggest that going hard may not be the best option for long-term brain health"

Excercise and the Brain

- from *The Journal of Physiology*



learning objectives

- Introduce linear regression
  - How do we find the line of best fit?
  - What is the slope?
  - What is the intercept?
  - What is the R squared?
- Using R to run a linear regression and add a regression line to a scatter plot
- How do we transform data that do not look linear to make a line?
- How do outliers influence our line of best fit?
- Some Important cautions
  - Association is not causation
  - Do not extrapolate beyond your data
  - Always consider potential confounders in your interpretation
  - Confirm the shape of your data visually

Regression

What is a regression line?

- A straight line that is fitted to data to minimize the distance between the data and the fitted line.
- It is often called the line of best fit.
- It is also called the least-squares regression line (sometimes refered to as *ordinary least squares or ols*) this is because mathmatically, the criteria for choosing this line is based on the sum of squares of the vertical distances from the line. We choose the line that minimizes this sum.

What is a regression line?

Once we have calculated this line, the line of best fit can be used to describe the relationship between the explanatory and response variables.

- Can you fit a line of best fit for non-linear relationships?
- Very important to visualize the relationship first. Why?

Equation of the line of best fit

The line of best fit can be represented by the equation for a line:

$$y = a + bx$$

where a is the intercept and b is the slope.

This equation encodes a lot of useful information

In earlier math classes you may have seen this expressed as:

$$y = mx + b$$

Equation of the line of best fit: the intercept

$$y = a + bx$$

If $x = 0$, the equation says that $y = a$, which is why a is known as the intercept.

Note: Is the value of the intercept always meaningful?

Equation of the line of best fit: the slope

$$y = a + bx$$

b is known as the slope because an increase from x to $x + 1$ is associated with an increase in y by the amount b.

The slope is closely related to the correlation coefficient:

$$b = r\frac{S_y}{S_x}$$

If the correlation coefficient is negative what will be the sign of the b?

Model R squared

The r^2 value or R squared, is the fraction of the variation in the values of y that is explained by the regression of y on x

In a regression where every observation fell exactly on the regression line, the value of r^2 would be 1.

In a linear regression with only one x the r^2 is the square of the correlation coefficient.

Fitting a linear model in R

Fitting a linear model in R

Code template:

```
lm(formula = y ~ x, data = your_dataset)
```

- `lm()` is the function for a linear model.
- The first argument that `lm()` wants is a formula `y ~ x`.
  - `y` is the response variable from your dataset
  - `x` is the explanatory variable
  - be careful with the order of x and y! It is opposite from the default order in ggplot

  ggplot(data,aes(x=your_x, y=your_y))
- The second argument sent to `lm()` is the data set.
  - the default order of declaring the data as the second argument in lm() is different from the ggplot2 and dplyr functions

Why the package broom?

We will pull in a new package here: library(broom) and apply the tidy() function as follows: tidy(your_lm)

- `broom` has functions that make the output from the linear model look clean
- `tidy` is a function from the `broom` package that tidies up the output

Example: Manatee deaths and powerboat purchases

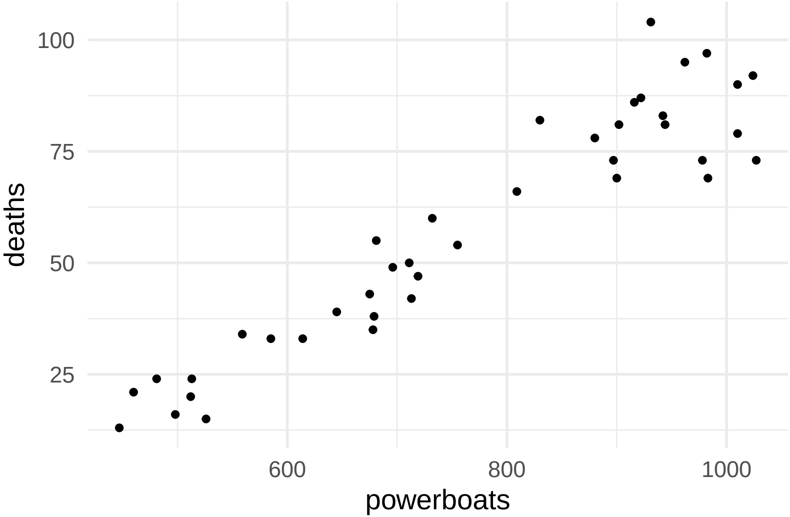
Let's apply the lm() function. Recall the manatee example from our last lecture that examined the relationship between the number of registered powerboats and the number of manatee deaths in Florida between 1977 and 2016.

Recall that the relationship appeared linear when we examined the scatter plot:

```
library(ggplot2)
mana_death1<-ggplot(mana_data, aes(x = powerboats, y = deaths)) +
  geom_point() +
  theme_minimal(base_size = 15)
```

Manatee deaths and powerboat purchases

```
mana_death1
```



lm() of manatee deaths and powerboat purchases

Calculate the line of best fit:

```
mana_lm <- lm(deaths ~ powerboats, mana_data)
library(broom)
tidy(mana_lm)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -46.8     6.03        -7.75 2.43e- 9
## 2 powerboats     0.136   0.00764     17.8  5.21e-20
```

Only pay attention to the term and estimate columns for now.

lm() of manatee deaths and powerboat purchases

Interpret the model output

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -46.8     6.03        -7.75 2.43e- 9
## 2 powerboats     0.136   0.00764     17.8  5.21e-20
```

- Intercept: The predicted number of deaths if there were no powerboats. But the prediction is negative. Why?
- Powerboats: This is the slope. What does the estimated slope for powerboats mean?

Interpreting the slope

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -46.8     6.03        -7.75 2.43e- 9
## 2 powerboats     0.136   0.00764     17.8  5.21e-20
```

- A one unit change in the number of powerboats registered (X 1,000) is associated with an increase of manatee deaths of 0.1358. That is, an increase in the number of powerboats registered by 1,000 is association with 0.1358 more manatee deaths.
- If powerboat registered increased by 100,000 how many more manatee deaths are expected?

Change units

```
mana_data_units<-mana_data%>%mutate(actual_powerboats = powerboats * 1000)
mana_lm_units <- lm(deaths ~ actual_powerboats, mana_data_units)
tidy(mana_lm_units)
```

```
## # A tibble: 2 x 5
##   term              estimate  std.error statistic  p.value
##   <chr>                <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)      -46.8      6.03          -7.75 2.43e- 9
## 2 actual_powerboats   0.000136 0.00000764     17.8  5.21e-20
```

What happened to the slope? To the intercept?

Getting the R-squared from your model

When we run a linear model, the r-squared is also calculated. Here is how to see the r-squared for the manatee data:

```
library(broom)
glance(mana_lm)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BI
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl
## 1     0.893         0.890  8.82      316. 5.21e-20     1  -143.  292.  297
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Focus on:

- Column called `r.squared` values only.
- Interpretation of r-squared: The fraction of the variation in the values of y that is explained by the line of best fit.

Correlation vs R Squared

```
library(dplyr)
mana_cor <- mana_data %>%
  summarize(corr_mana = cor(powerboats, deaths))
mana_cor
```

```
## # A tibble: 1 x 1
##   corr_mana
##       <dbl>
## 1     0.945
```

Correlation vs R Squared

```
glance(mana_lm)%>% pull(r.squared)
```

```
## [1] 0.8926573
```

```
#square the correlation coefficient
.9448054^2
```

```
## [1] 0.8926572
```

Add the regression line to the scatter plot using `geom_abline()`

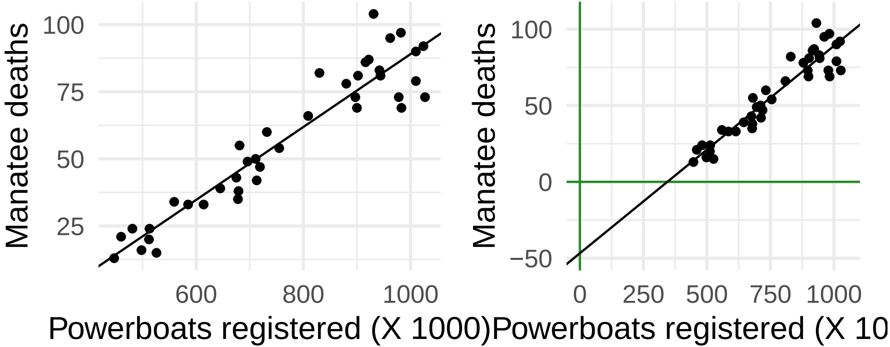
Add the regression line to the scatter plot using `geom_abline()`

We add a statement to our ggplot geom_abline(intercept = your_intercept, slope = your_slope)

so for our manatee data geom_abline(intercept = -46.7520, slope = 0.1358)

Note: by default, ggplot only shows the ploting region that corresponds to the range of data

Add the regression line to the scatter plot using `geom_abline()`



Add the regression line to the scatter plot using `geom_abline()`

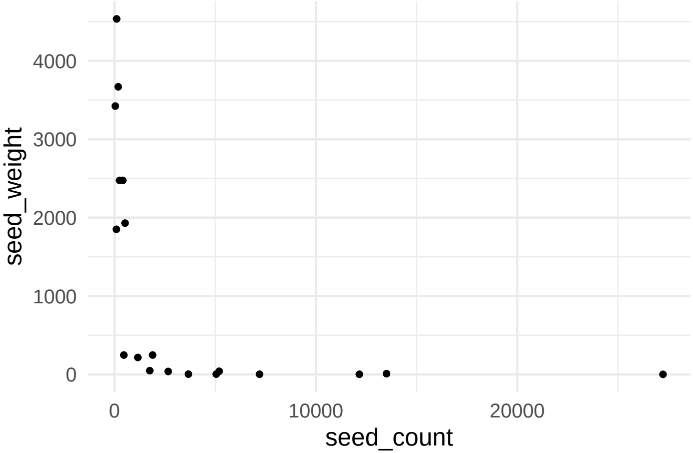
- When we add the line, we can see the intercept estimate. It is where the line of best fit intersects the y axis. Should we interpret it?
  - It is far from the bulk of the data, there is no data near powerboats $= 0$
  - Interpretation would be extrapolation, and is not supported by these data

Transforming data

Transforming data

- Sometimes, the data is transformed to another scale so that the relationship between the transformed x and y is linear
- Table 3.4 in B&M provides data on the mean number of seeds produced in a year by several common tree species and the mean weight (in milligrams) of the seeds produced.

Scatter plot of seed_weight vs. seed_count



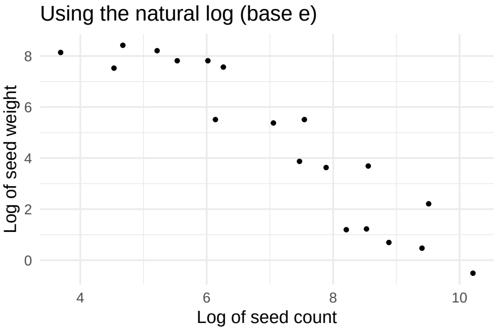
- seed_count and seed_weight both vary widely
- Their relationship is not linear

Investigate the relationship between their logged variables

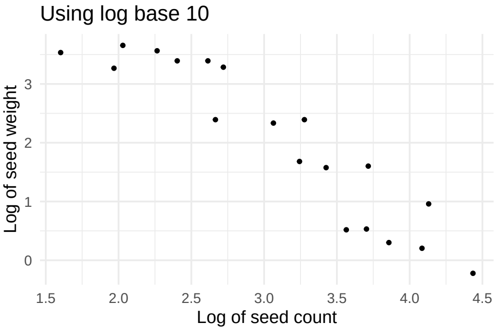
- Add transformed variables to the dataset using `mutate()`.
- We add both log base *e* and log base 10 variables for illustration

```
library(dplyr)
seed_data <- seed_data %>% mutate(log_seed_count = log(seed_count),
                                  log_seed_weight = log(seed_weight),
                                  log_b10_count = log(seed_count, 10),
                                  log_b10_weight = log(seed_weight, 10))
```

Plot transformed data (log base e)



Plot transformed data (log base 10)



- You can use either base 10 or base *e* for class.
- The calculations using base *e* are easier

lm() on the log (base e) variables

```
seed_mod <- lm(log_seed_weight ~ log_seed_count, data = seed_data)
tidy(seed_mod)
```

```
## # A tibble: 2 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       15.5      1.08       14.3 6.37e-11
## 2 log_seed_count    -1.52     0.147     -10.4 9.28e- 9
```

```
glance(seed_mod) %>% pull(r.squared)
```

```
## [1] 0.8631177
```

- Interpret the intercept:
- Interpret the slope:

lm() on the log (base 10) variables

```
seed_mod_b10 <- lm(log_b10_weight ~ log_b10_count, data = seed_data)
tidy(seed_mod_b10)
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       6.73     0.469      14.3 6.37e-11
## 2 log_b10_count    -1.52     0.147     -10.4 9.28e- 9
```

```
glance(seed_mod_b10) %>% pull(r.squared)
```

```
## [1] 0.8631177
```

- What is different from the log base *e* output?

Predictions from `lm()` when using log (base *e*) data

- What seed weight is predicted for a seed count of 2000?
- Worked calculation:

1. Write down the line of best fit:
   $log_e(seed.weight) = 15.49130 - 1.522220 \times log_e(seed.count)$
2. Plug in $seed.count = 2000$ into the line of best fit:
   $log_e(seed.weight) = 15.49130 - 1.522220 \times log_e(2000)$
3. Solve for seed count by exponentiating both sides:

$$seed.weight = exp(15.49130 - 1.522220 \times log_e(2000))$$

(this uses the property that $e^{log_e(x)} = x$)

$$seed.weight = 50.45$$

4. Interpret: Seeds are expected to weigh 50.45 for trees having a seed count of 2000.

How do outliers affect the line of best fit?

How do outliers affect the line of best fit?

To study this, we use data from the Organization for Economic Co-operation and Development (OECD). This dataset was downloaded from http://dx.doi.org/10.1787/888932526084 and contains information on the health expenditure per capita and the GDP per capita for 40 countries.

```
library(readxl)

spending_dat <- read_xlsx("Ch04_Country-healthcare-spending.xlsx",
                          sheet = 2,
                          range = "A7:D47")
```

Have a look

Next, we want to examine the imported data to see if it is how we expect:

```
head(spending_dat)
```

```
## # A tibble: 6 x 4
##   Country   Country.code `Health expenditure per capita` `GDP per capita`
##   <chr>     <chr>                                  <dbl>            <dbl>
## 1 Australia AUS                                     3445            39409
## 2 Austria   AUT                                     4289            38823
## 3 Belgium   BEL                                     3946            36287
## 4 Brazil    BRA                                      943            10427
## 5 Canada    CAN                                     4363            38230
## 6 Chile     CHL                                     1186            14131
```

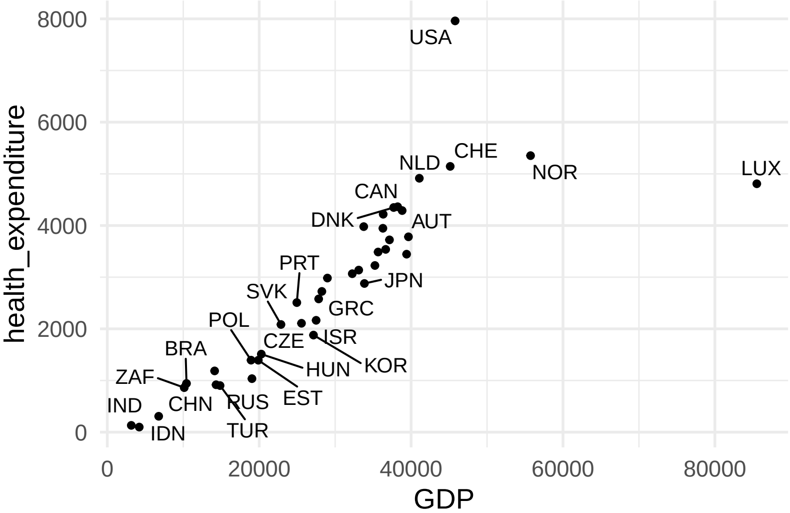
Rename() some variables to use a consistent naming style

If the variable name has spaces, we must use back ticks when referring to it:

```r
library(dplyr)
spending_dat <- spending_dat %>%
  rename(country_code = Country.code,
         health_expenditure = `Health expenditure per capita`, # back ticks
         GDP = `GDP per capita`) # back ticks
```

Examine the relationship

Make a scatter plot of `health_expenditure` (our response variable) vs. each country's level of `GDP`:



Examine the relationship

Is the relationship linear? Which countries are outliers?

Fit a linear model to these data

```
lm(health_expenditure ~ GDP, data = spending_dat)
```

```
##
## Call:
## lm(formula = health_expenditure ~ GDP, data = spending_dat)
##
## Coefficients:
## (Intercept)          GDP
##    44.65623      0.09399
```

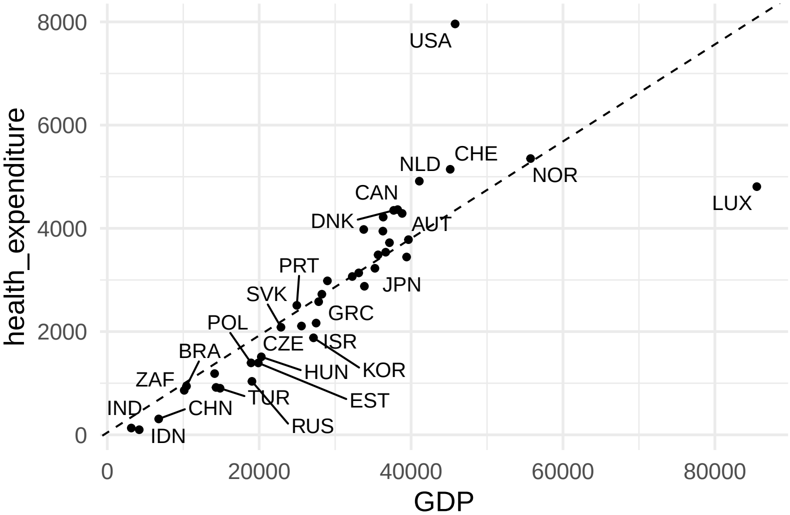
Examine the relationship

Add the regression line to the graph:

```
GDP_withline<-ggplot(spending_dat, aes(x = GDP, y = health_expenditure)) +
  geom_point() +
  geom_text_repel(aes(label = country_code)) + # this adds the country code a
  geom_abline(intercept = 44.65623, slope = 0.09399, lty = 2) +
  theme_minimal(base_size = 15)
```

Examine the relationship

```
## Warning: ggrepel: 15 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Examine the relationship without Luxembourg in the data

Let's see whether removing Luxembourg changes the fit of the line. We can remove Luxembourg using the `filter()` command from `dplyr`:

```
spending_dat_no_LUX <- spending_dat %>% filter(country_code != "LUX")

lm(health_expenditure ~ GDP, data = spending_dat_no_LUX)
```

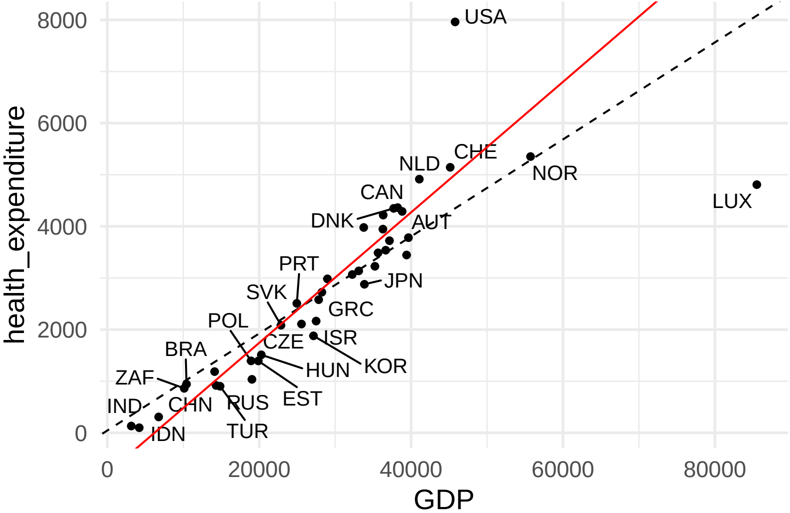
```
##
## Call:
## lm(formula = health_expenditure ~ GDP, data = spending_dat_no_LUX)
##
## Coefficients:
## (Intercept)          GDP
##   -785.1044       0.1264
```

Examine the relationship without Luxembourg in the data

```
GDP_nolux<-ggplot(spending_dat, aes(x = GDP, y = health_expenditure)) + geom_
  geom_text_repel(aes(label = country_code)) +
  geom_abline(intercept = 44.65623, slope = 0.09399, lty = 2) +
  geom_abline(intercept = -785.1044, slope = 0.1264, col = "red") +
  theme_minimal(base_size = 15)
```

Examine the relationship without Luxembourg in the data

```
## Warning: ggrepel: 15 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Examine the relationship without USA in the data

```
spending_dat_no_USA <- spending_dat %>% filter(country_code != "USA")

lm(health_expenditure ~ GDP, data = spending_dat_no_USA)
```

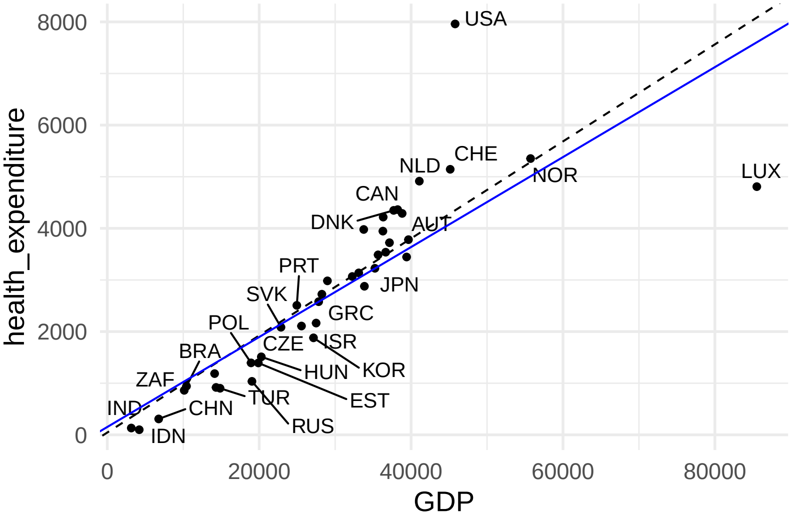
```
##
## Call:
## lm(formula = health_expenditure ~ GDP, data = spending_dat_no_USA)
##
## Coefficients:
## (Intercept)          GDP
##   152.26274      0.08714
```

Examine the relationship without USA in the data

```
GDP_nousa<-ggplot(spending_dat, aes(x = GDP, y = health_expenditure)) + geom_
  geom_text_repel(aes(label = country_code)) +
  geom_abline(intercept = 44.65623, slope = 0.09399, lty = 2) +
  geom_abline(intercept = 152.26274, slope = 0.08714, col = "blue") +
  theme_minimal(base_size = 15)
```

Examine the relationship without USA in the data

```
## Warning: ggrepel: 15 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Examine the relationship without LUX or USA in the data

Let's write the code together to remove both the USA and LUX and see how it affects the fit:

```
spending_dat_no_USA_LUX <- spending_dat %>%

  filter(country_code != "USA" & country_code != "LUX")

#alternatively, you could have written:
spending_dat_no_USA_LUX <- spending_dat %>%

  filter(! country_code %in% c("USA", "LUX"))

#pick the filter command that makes the most sense to you.
```

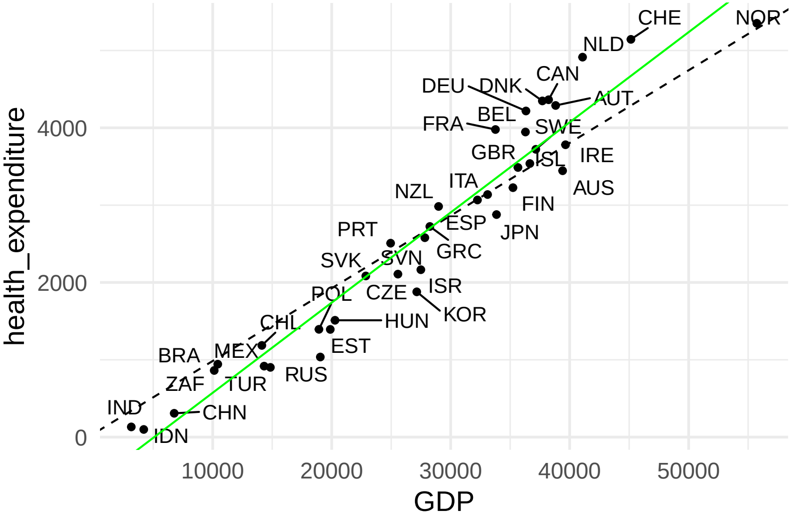
Examine the relationship without LUX or USA in the data

```
lm(health_expenditure ~ GDP, data = spending_dat_no_USA_LUX)
```

```
##
## Call:
## lm(formula = health_expenditure ~ GDP, data = spending_dat_no_USA_LUX)
##
## Coefficients:
## (Intercept)          GDP
##   -592.6973       0.1166
```

```
GDP_noluxnousa<-ggplot(spending_dat_no_USA_LUX, aes(x = GDP, y = health_expen
  geom_text_repel(aes(label = country_code)) +
  geom_abline(intercept = 44.65623, slope = 0.09399, lty = 2) +
  geom_abline(intercept = -592.6973, slope = 0.1166 , col = "green") +
  theme_minimal(base_size = 15)
```

Examine the relationship without LUX or USA in the data



Examine the relationship without LUX or USA in the data

What would happen if USA's point had actually been along the original line of best fit (say at $x = 80000$ and $y = 7500$) and we re-fit the line without USA's point?

Would USA have been an outlier? Would it be considered influential?

But, is it causal?

- Creating a scatter plot and a simple linear model is an important step in many analyses. It allows you to see the relationship between two quantitative variables and estimate the line of best fit.
- Sometimes these relationships will be used to make claims of causality.

Baldi & Moore emphasize that experiments are the best way to study causality. While this is often true, sophisticated causal methods have been developed for the analysis of observational data.

Counfounding

Counfounding

Your book talks about “lurking variables” which Baldi & Moore define as:

> *A variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.*

They also (pg 157) define confounding by saying:

> *Two variables (explanatory or lurking) are confounded when their effects on a response variable cannot be distinguished from each other.*

I strongly disagree with this definition. We will use a different definition in this class.

Definition of Counfounding

A relationship between your variable of interest (exposure, treatment) and your outcome of interest (disease status, health condition etc) is confounded when there is a variable that is associated with both the exposure and outcome, and is not on the causal pathway between the two.

Variables that are on the causal pathway are those that represent a way in which the exposure acts on the outcome. For example, poor cognitive function would be on the causal pathway between lack of sleep and trying to pay for groceries with your library card.

Discussion of Music example from Baldi & Moore

Example 4.7 “Nature, nuture, and lurking variables” presents an advertisement from the Michigan Symphony:

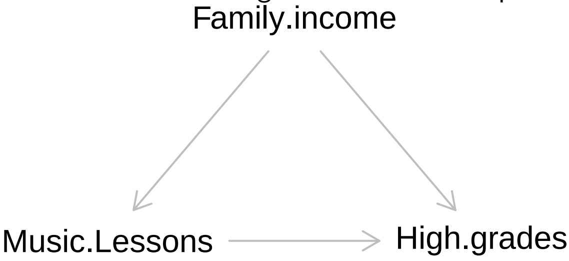
"Question: Which students scored 51 points higher in verbal skills and 39 points higher in math?

Answer: Students who had experience in music."

Marketers often make leading statements that make their product or service sound appealing. The purpose of this ad was to have the target audience impute that music causes higher marks at school because there is an association between enrollment in music and higher marks. However, are students enrolled in music lessons otherwise the same as students not enrolled in music lessons? What else do you expect to differ between these groups of students?

Discussion of some examples from Baldi & Moore

We can encode these differences in a causal diagram. Here is a simple one to demonstrate the concept:



The direction of the arrows from the "Family Income" node makes explicit that we believe family income to be a confounder of the relationship between taking music lessons and achieving higher grades. It means that not only do these children take music lessons, they also come from families with higher incomes, and higher incomes lead to higher grades in other ways. Of course, family income is not the only possible confounder. What are some others?

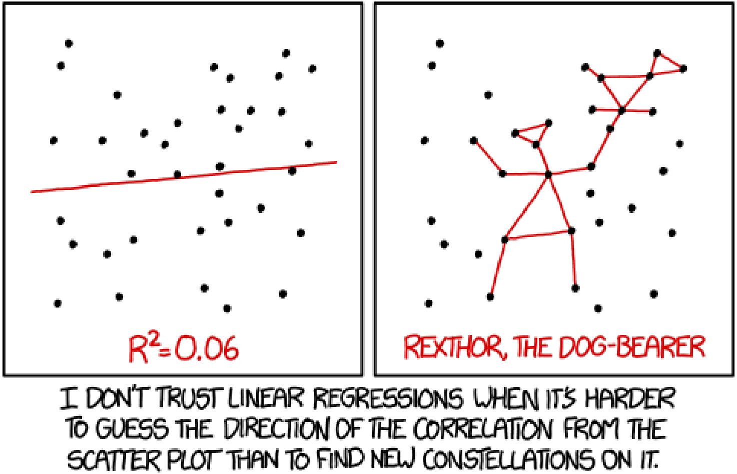
Counfounding

In this course, we don't have time to go into methods that adjust for multiple variables or address how to control for confounding or other types of bias that limit causal interpretations.

However, know that causality can be studied using observational data and relies on clever study designs and oftentimes on advanced methods.

Comic Relief

From xkcd.com


R^2=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.