Review: Describing Variables

The Effect Book: Chapter 3

Outline

Read: Chapter 4- Describing Relationships

- Conditional Distributions
- Conditional Means
- "Controlling for a variable"
- Relationships in Software

Relationships

- Two variables X and Y are *independent* if learning the value of one tells you nothing about the other
- For example, knowing the outcome of a roulette spin tells you nothing about what the next spin will be
- They are *dependent* if learning the value of one *changes the distribution* of the other
- For example, among all Americans, about 50.8\% are legally female and about 49.2\% are legally male
- But if we *learn that someone's name is Susan*, that distribution will change considerably in favor of female!

Conditional Values

- Another way of saying we *learn the value of one variable* is to say we *condition* on the value of that variable
- "Conditional on someone's name being Susan, the distribution of legal sex is 96% female and 4% male"
- P(Gender | Name = "Susan")
- All the stuff we talked about last time with distributions applies here
- It just applies only to *a portion of the data* (the Susans!)

Conditional Means

- In many fields, including economics, we are very interested in calculating *conditional means*
- i.e. *what is the mean of the conditional distribution?*
- And we want to know this not only for *one* value of the variable we are conditioning on, but *all* the values
- The *population* conditional mean is the *conditional expectation*, or $E(Y|X)$

Conditional Means/Expectations and Relationships

- Talking about how two variables are related is just another way of talking about conditional values (usually, conditional means)
- If X and Y are positively related, that means "at higher values of X, the conditional expectation $E(Y|X)$ is higher"
- If they're negatively related, "at higher values of X, $E(Y|X)$ is lower"
- If their relationship is U-shaped (or similar), "at higher values of X, $E(Y|X)$ changes but not always in the same direction"

Conditional Means and Causal Inference

- *Statistics* is great at telling us how two variables are related
- *Causal inference* is *entirely* about figuring out *which conditional relationships are causal and in what way*
- You cannot infer causality from a relationship alone!
- As I will show, some of the relationships $Y|X$ you will see today may reflect an X that causes Y, or a Y that causes X, or something else entirely!

"Explaining"

- We can also say that $Y|X$ is "the part of Y that is *explained by* X"
- If $E(CoffeeCupsPerDay|\ Occupation = Professor) = 1.79$, and I drink 3 cups per day, then 1.79 of my cups are "explained by" the fact that I'm a professor, and $3 - 1.79 = 1.21$ of my cups are "not explained by" being a professor
- This can extend to multiple variables!
- If $E(CoffeeCupsPerDay|\ Occupation = Professor, Gender = Male) = 2.13$, then 2.13 of my cups are "explained by" occupation and age, and $3 - 2.13 = .83$ of my cups are "not explained by" those two things

"Explaining"

- This is a purely statistical explanation
- This doesn't necessarily mean that 2.13 of my cups are **because** I'm a professor and a man, but rather that 2.13 of my cups are **what would be expected** given what you know about me
- The "explanation why" would be the same *statistical* calculation, but would require us to figure out *which* of those calculations is causal in nature.

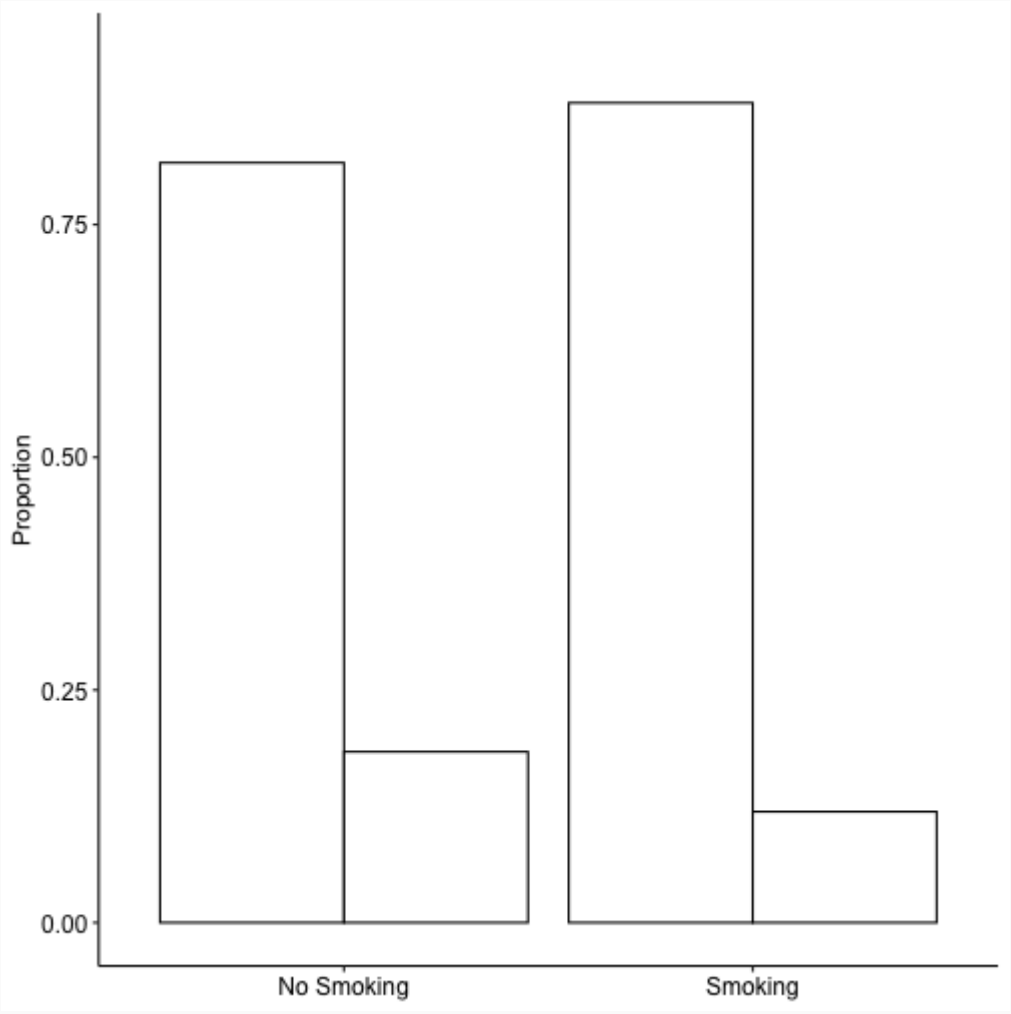
Ways of Demonstrating Conditional Means

- When X only takes a few values, we can just show the distribution of Y conditional on each of the values of X, and compare them to each other. Any distribution-showing method works - density plots, bar graphs...
- Scatter plots show all the data and imply the conditional mean
- For continuous X data, you can calculate the conditional mean $Y|X$ over different *ranges* of X, either splitting X into bins, or doing "local means"
- Regression

Example

- Following the textbook, we'll be using data from Emily Oster's study of the relationship between taking Vitamin E and health outcomes
- (and whether that relationship changes as a result of Vitamin E being briefly recommended, then not!)
- We will start with versions that can be used for *discrete* X values, like "smoking vs. non-smoking"
- Remember, the proportion of a binary variable **is** its mean

Conditional Means: Contrasting Bar Graphs



Conditional Means

- Or we can just show that information in a table

```
oster %>% select(smoke,supplement_vite_single) %>%
          mutate(smoke = as.factor(
            case_when(smoke == 0 ~ 'No Smoking',
                      TRUE ~ 'Smoking'))) %>%
  table() %>% prop.table(margin = 1)
```

```
##              supplement_vite_single
## smoke         No Vitamin E Took Vitamin E
##   No Smoking     0.8164492      0.1835508
##   Smoking        0.8919705      0.1080295
```

Conditional Means

- That was $TookVitaminE|Smoking$. How about $Smoking|TookVitaminE$?

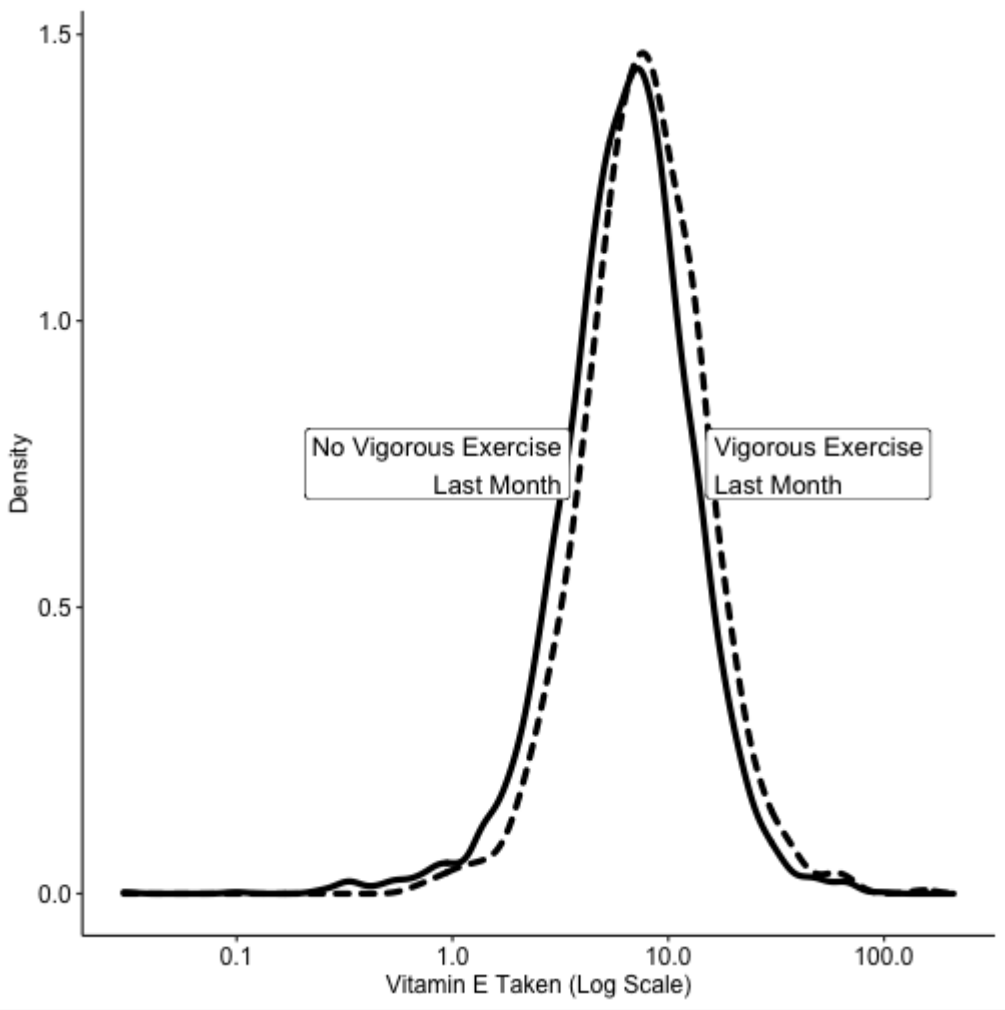
```
oster %>% select(smoke,supplement_vite_single) %>%
          mutate(smoke = as.factor(
            case_when(smoke == 0 ~ 'No Smoking',
                      TRUE ~ 'Smoking'))) %>%
  table() %>% prop.table(margin = 2)
```

```
##              supplement_vite_single
## smoke         No Vitamin E Took Vitamin E
##   No Smoking    0.6999860      0.8124162
##   Smoking       0.3000140      0.1875838
```

Conditional Distributions: Contrasting Densities

```
ggplot(oster, aes(x = vite, linetype = factor(vigorous_exercise_month))) +
  geom_density(size = 1.5) +
  scale_x_log10() +
  labs(x = 'Vitamin E Taken (Log Scale)',
       y = 'Density') +
  annotate(geom = 'label', x = 3.5, y = .75, label = 'No Vigorous Exercise\nLast Month', hjust = 1, family = 'Garamon
  annotate(geom = 'label', x = 15, y = .75, label = 'Vigorous Exercise\nLast Month', hjust = 0, family = 'Garamond',
  guides(linetype = FALSE) +
  theme_pubr()
```

Conditional Distributions: Contrasting Densities



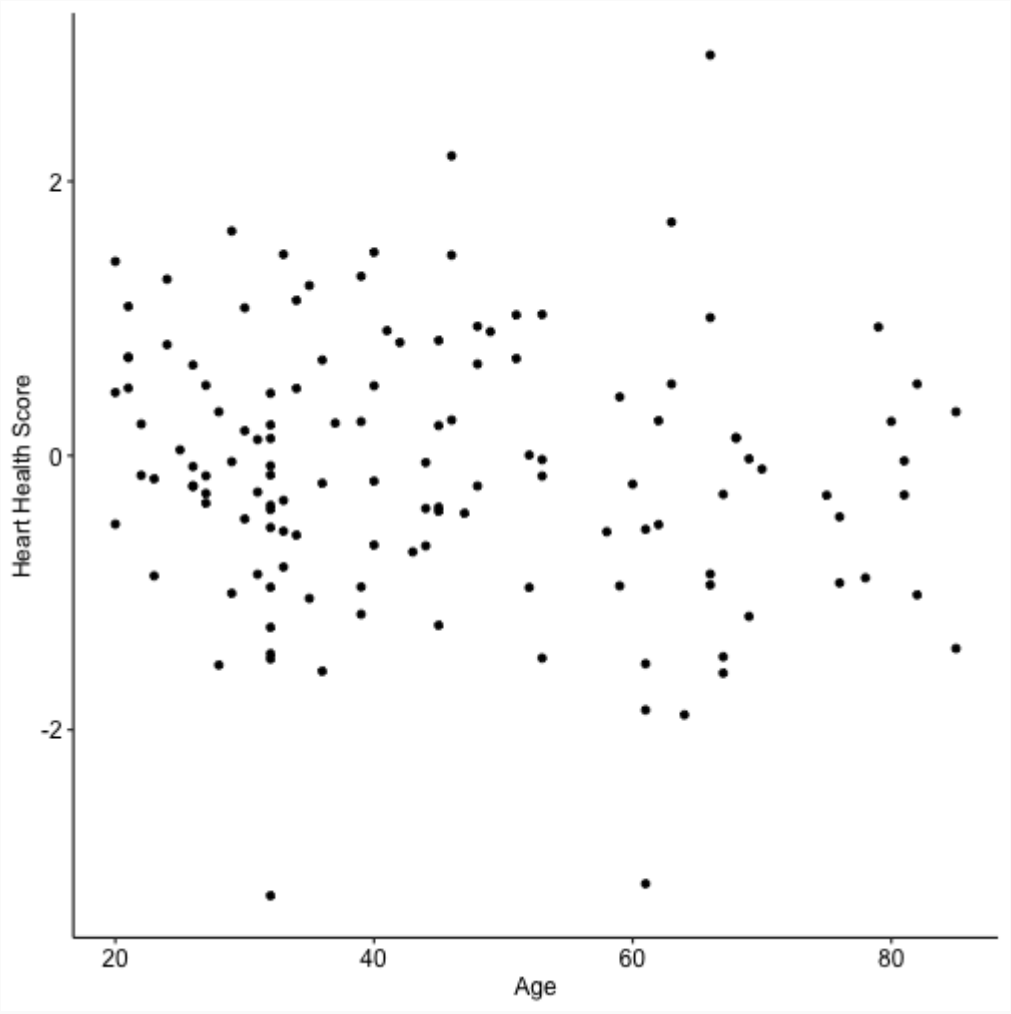
Continuous X Variables

- What if the X variable we want to condition on is continuous?
- We have options!
- A scatter plot will just show *all* the data (which may be too much/busy!)
- Of course, if we really want to *understand* or *describe* the relationship, we'll need some way of summarizing what we see!

Scatterplots

```
ggplot(oster %>% slice(150:300), aes(x= age, y = heart_health)) +
  geom_point() +
  labs(x = 'Age',
       y = 'Heart Health Score') +
  theme_pubr()
```

Scatterplots



Data in the Social Sciences

- Looks like a blob with sort of a suggestion of a relationship, rather than something really clear, doesn't it?
- That's par for the course in the social sciences
- Super-clear scatterplots are for relationships that are truly *bivariate*, i.e. X explains Y almost perfectly
- If there's *a lot of other stuff going on*, i.e. unexplained parts, the data reflects that and looks more blobby
- In the social sciences, there's ALWAYS a lot of other stuff going on

Binning

- How can we calculate a conditional mean when there might only be one observation with a given X value? Can't really estimate $E(Y|X = x)$ well if there's only one observation with $X = x$!
- We'll need to group observations together
- One way is binning - we'll rarely actually use this but it's good for demonstration
- Cut up X into bins, and take the average of Y within each bin

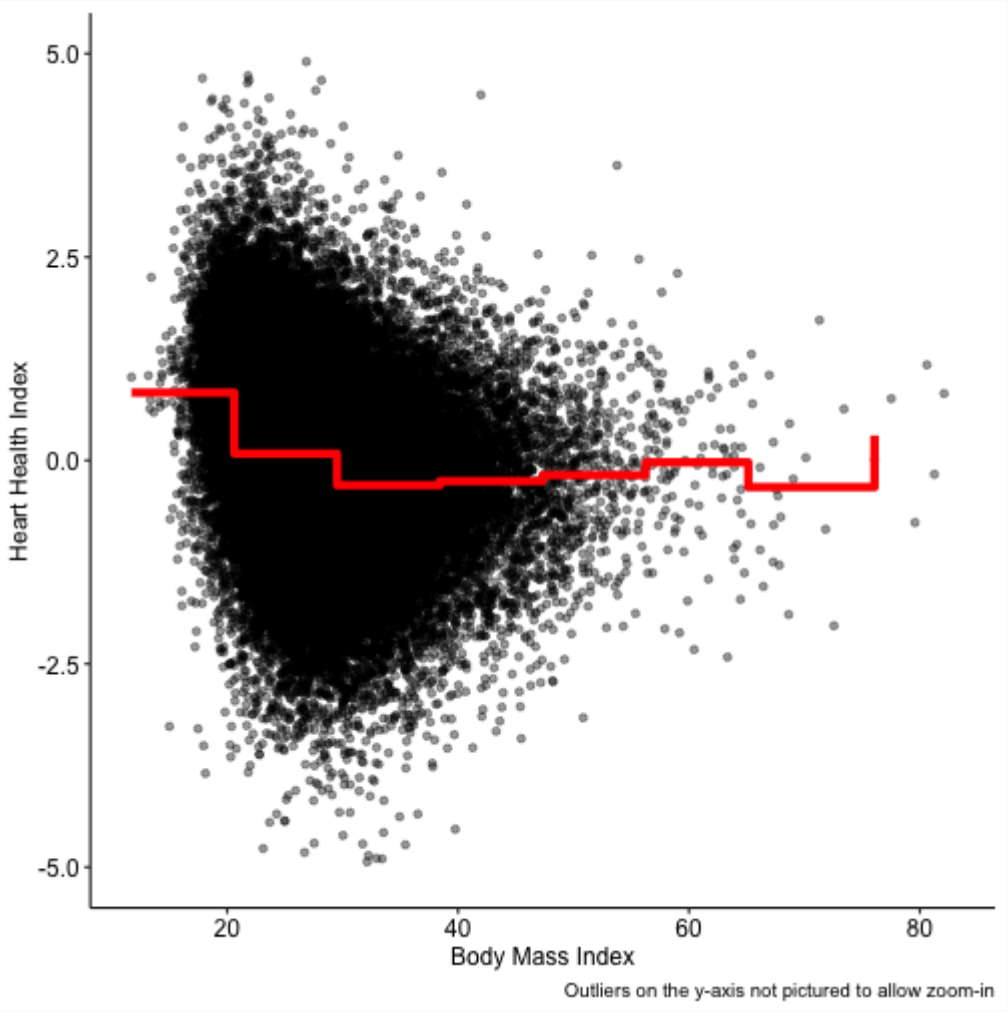
Binning

```
oster %>% filter(bmi < 100) %>%
  mutate(bmi_cut = cut(bmi, 8)) %>% group_by(bmi_cut) %>%
  summarize(vite = mean(heart_health, na.rm = TRUE)) %>%
  mutate(vite = scales::number(vite, accuracy = .001)) %>%
  rename(`BMI Bin` = bmi_cut,
         `Heart Health Index` = vite)
```

```
## # A tibble: 8 × 2
##   `BMI Bin`   `Heart Health Index`
##   <fct>       <chr>
## 1 (11.6,20.6] 0.834
## 2 (20.6,29.5] 0.083
## 3 (29.5,38.4] -0.304
## 4 (38.4,47.3] -0.255
## 5 (47.3,56.2] -0.180
## 6 (56.2,65.1] -0.023
## 7 (65.1,74]   -0.328
## 8 (74,83]     0.306
```

Binning

- Any guesses why that last bin is weird?



Binning

- Doing this binning thing tells us the average of Y within certain defined ranges of X, letting us estimate the population $E(Y|X \in [StartofBin, EndofBin])$
- So I can get $E(Y|X = x)$ by just seeing which bin x is in
- Of course, this is a bit arbitrary and strange - where do the bin definitions come from, how close should we make them, and is it really reasonable to see big jumps going from the edge of one bin to another?

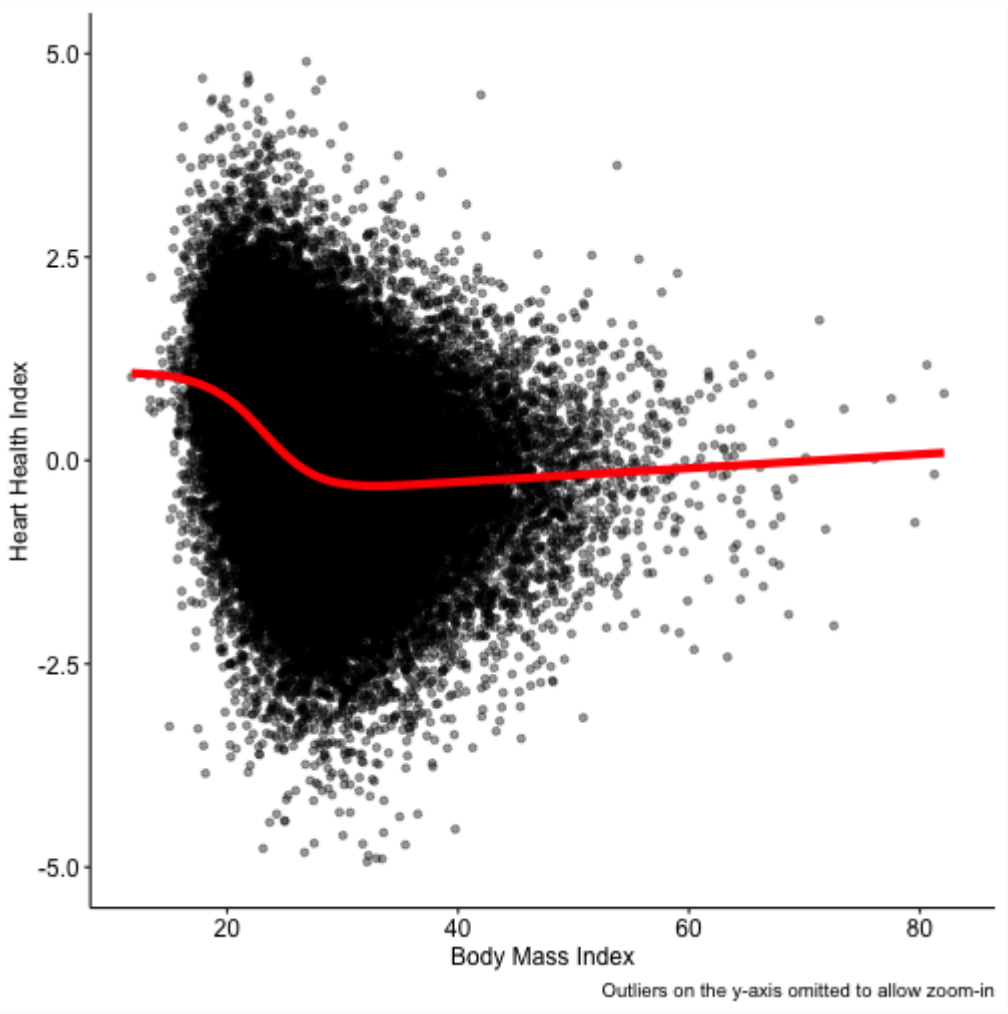
Local Means

- Another approach is to take *local expectations*
- Same binning idea, but we define the bin in a *rolling* manner
- For each value x, use a bin *centered on that* x (perhaps weighting closer values more highly)
- This gives us a smooth estimate (no jumps!) and while we still have decisions to make (width of bin, weights), it's less arbitrary
- If we fit a quadratic shape at each of those points, that's a LOESS (Locally Estimated Scatter Plot Smoothing)

LOESS

```
oster %>% filter(bmi < 100) %>%
  ggplot(aes(x = bmi, y = heart_health)) +
  geom_point(alpha = .4) +
  geom_smooth(size = 2, color = 'red', se = FALSE) +
  scale_y_continuous(limits = c(-5,5)) +
  labs(x = 'Body Mass Index', y = 'Heart Health Index',
       caption = 'Outliers on the y-axis omitted to allow zoom-in') +
  theme_pubr()
```

LOESS


5.0
2.5
0.0
-2.5
-5.0
Heart Health Index
20
40
60
80
Body Mass Index
Outliers on the y-axis omitted to allow zoom-in

LOESS

Benefits of LOESS:

- Nonparametric
- Easy to understand

Downsides:

- Difficult to use to sum up a relationship
- Or try to uncover population relationships

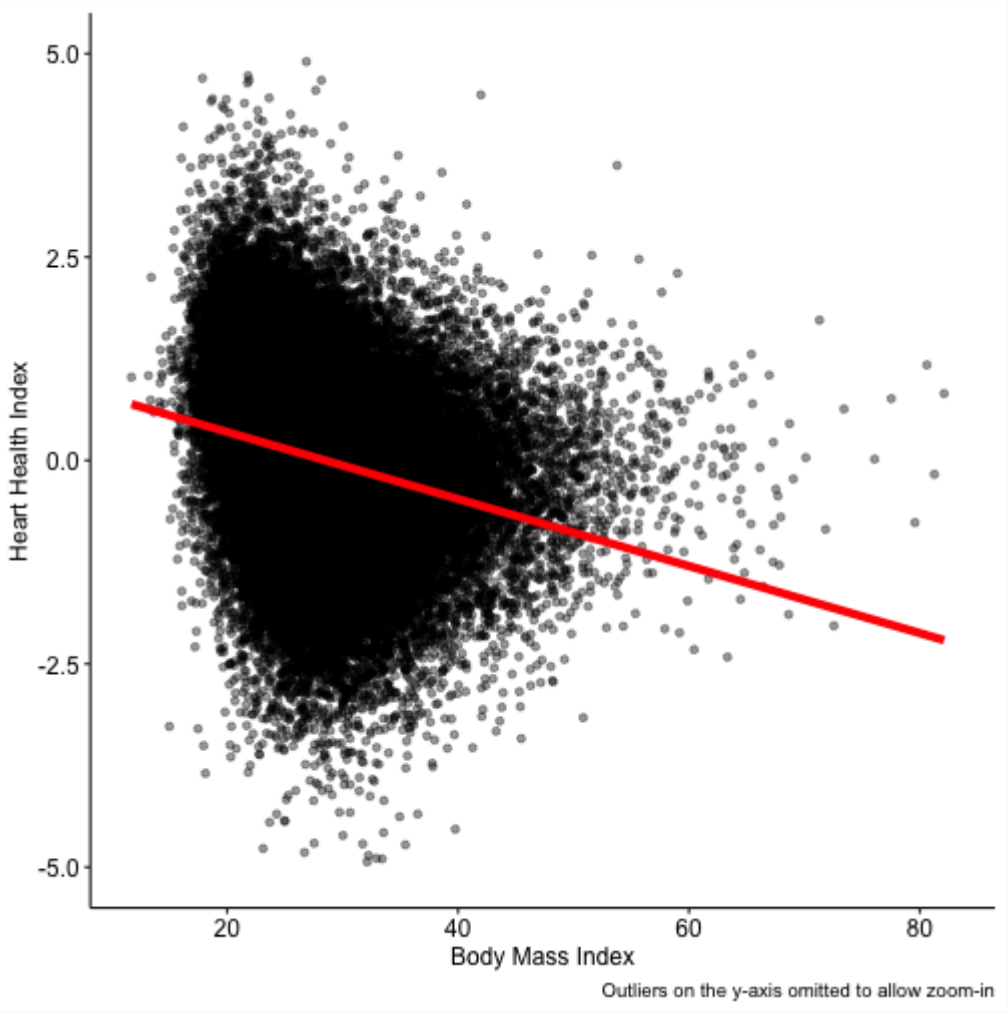
Regression

- This brings us to regression!
- Regression takes the *idea* behind a LOESS curve - use a line to represent the conditional mean at each value of X, smoothing over gaps between the X's - and generalizes it using a *shape*
- Regression is the process of fitting a line to data, *and requiring that that line holds a particular shape*
- In basic forms of regression, that shape is a straight line
- (You can make it a bit curvy by adding polynomial terms)

Linear Regression

```
oster %>% filter(bmi < 100) %>%
  ggplot(aes(x = bmi, y = heart_health)) +
  geom_point(alpha = .4) +
  geom_smooth(size = 2, color = 'red', se = FALSE, method = 'lm') +
  scale_y_continuous(limits = c(-5,5)) +
  labs(x = 'Body Mass Index', y = 'Heart Health Index',
       caption = 'Outliers on the y-axis omitted to allow zoom-in') +
  theme_pubr()
```

Linear Regression



Line-Fitting

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Pros:

- β_1 is a lot easier to interpret than continuously changing local means
- We can extrapolate to areas of X where we have no observations (but don't go beyond edge of data!)
- We understand the sampling distribution of $\hat{\beta}_1$

Cons:

- If we don't use the *right shape* the results will be bad
- We may toss out some interesting information that doesn't fit the shape

Line-Fitting

- Easy interpretation: a one-unit change in X is **associated with** a β_1-unit change in Y
- (If we've *identified a causal effect* then a one-unit change in X **causes** a β_1-unit change in Y)
- $\hat{\beta}_1$ follows a normal distribution (or nearly so, if ε isn't normal), so easy to do hypothesis tests
- Eight million notes and details to go over that we won't today, but they'll come up, and you hopefully covered some in econometrics!

Line-Fitting

Running a regression. This sees whether restaurant chains with more locations get higher/lower health inspection scores:

```
df ← read_csv('restaurant_data.csv')
m1 ← lm(inspection_score ~ NumberofLocations, data = df)
library(modelsummary)
msummary(m1, stars = TRUE,gof_omit = 'AIC|BIC|Lik|F|R2')
```

Interpret!

	Model 1
(Intercept)	94.866***
	(0.046)
NumberofLocations	−0.019***
	(0.000)
Num.Obs.	27178
RMSE	6.05
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Conditional Conditional Means

- The other bonus regression gives us is the ability to add *control variables*

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

- This gives us a "conditional conditional mean", i.e. "the conditional mean of Y given X, conditional on Z "
- "What is the part of the relationship between X and Y that is not explained by differences in Z?"

Conditional Conditional Means

- Adding a control *removes the part of* X *that is explained by* Z and also *removes the part of* Y *that is explained by* Z
- What's left over is the unrelated parts, and we can see how those unrelated parts relate

Conditional Conditional Means

```
m2 ← lm(inspection_score ~ NumberofLocations + Year, data = df)
df ← df %>%
  mutate(inspection_score_res = resid(lm(inspection_score ~ Year)),
         NumberofLocations_res = resid(lm(NumberofLocations ~ Year)))
m3 ← lm(inspection_score_res ~ NumberofLocations_res, data = df)
msummary(list(m2, m3), stars = TRUE, fmt = 5,
         gof_omit = 'AIC|BIC|Lik')
```

Conditional Conditional Means

	Model 1	Model 2
(Intercept)	225.33267***	0.00000
	(12.41118)	(0.03663)
NumberofLocations	-0.01919***	
	(0.00044)	
Year	-0.06489***	
	(0.00617)	
NumberofLocations_res		-0.01919***
		(0.00044)
Num.Obs.	27178	27178
R2	0.068	0.067
R2 Adj.	0.068	0.067

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001