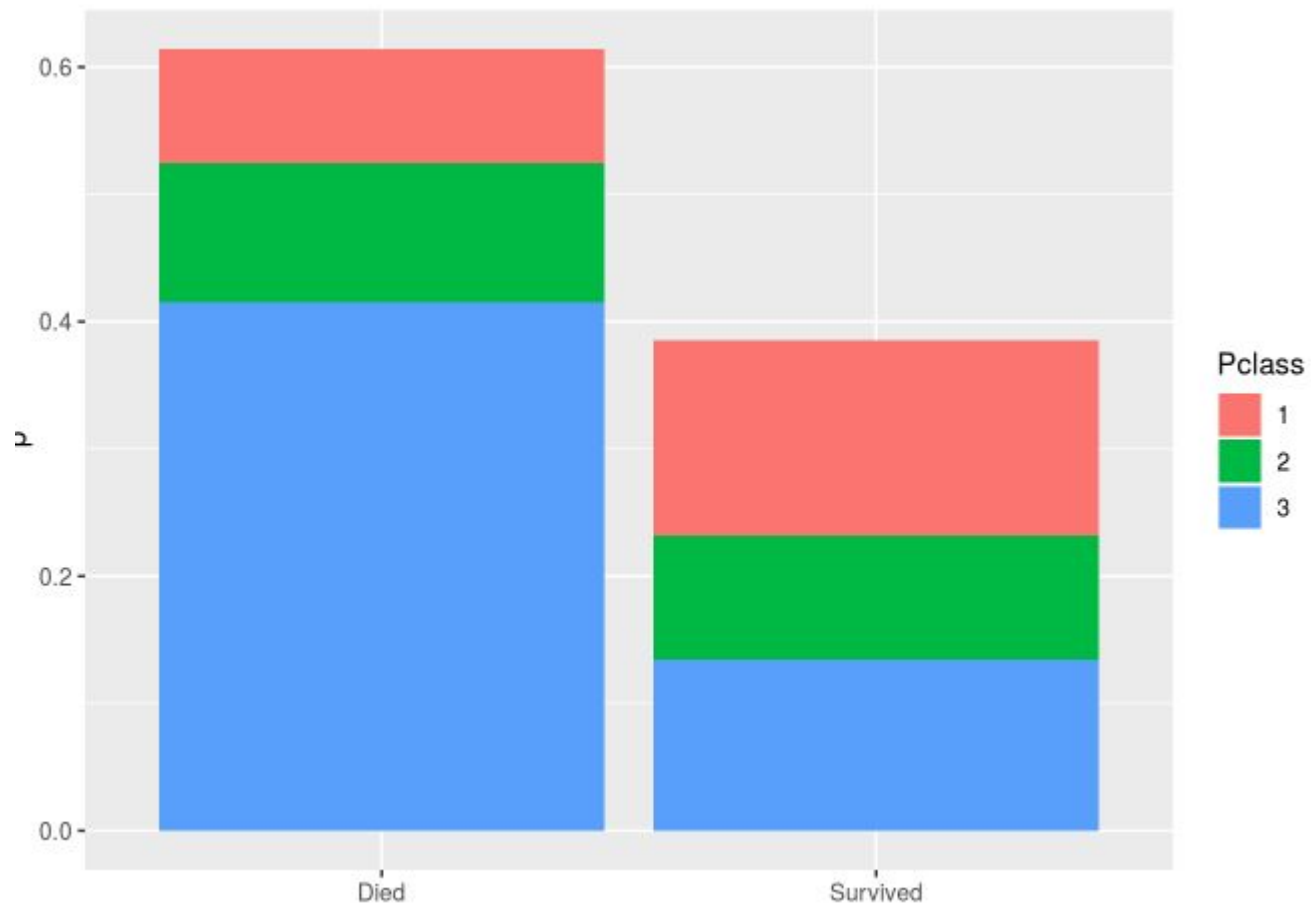


SML480 Week 5, Meeting 3



Aside: ggplot example



Association vs. Causation

```
gapminder$y <- gapminder$year - 1952  
summary(lm(lifeExp ~ log(gdpPercap) + y, data = gapminder))
```

Call:

```
lm(formula = lifeExp ~ log(gdpPercap) + y, data = gapminder)
```

Residuals:

Min	1Q	Median	3Q	Max
-27.2291	-3.8454	0.6065	4.7737	17.8644

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-9.300131	1.108141	-8.393	<2e-16	***
log(gdpPercap)	7.770320	0.138084	56.273	<2e-16	***
y	0.195569	0.009927	19.702	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

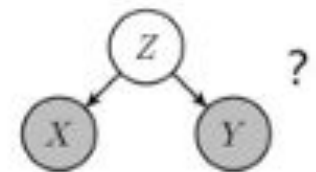
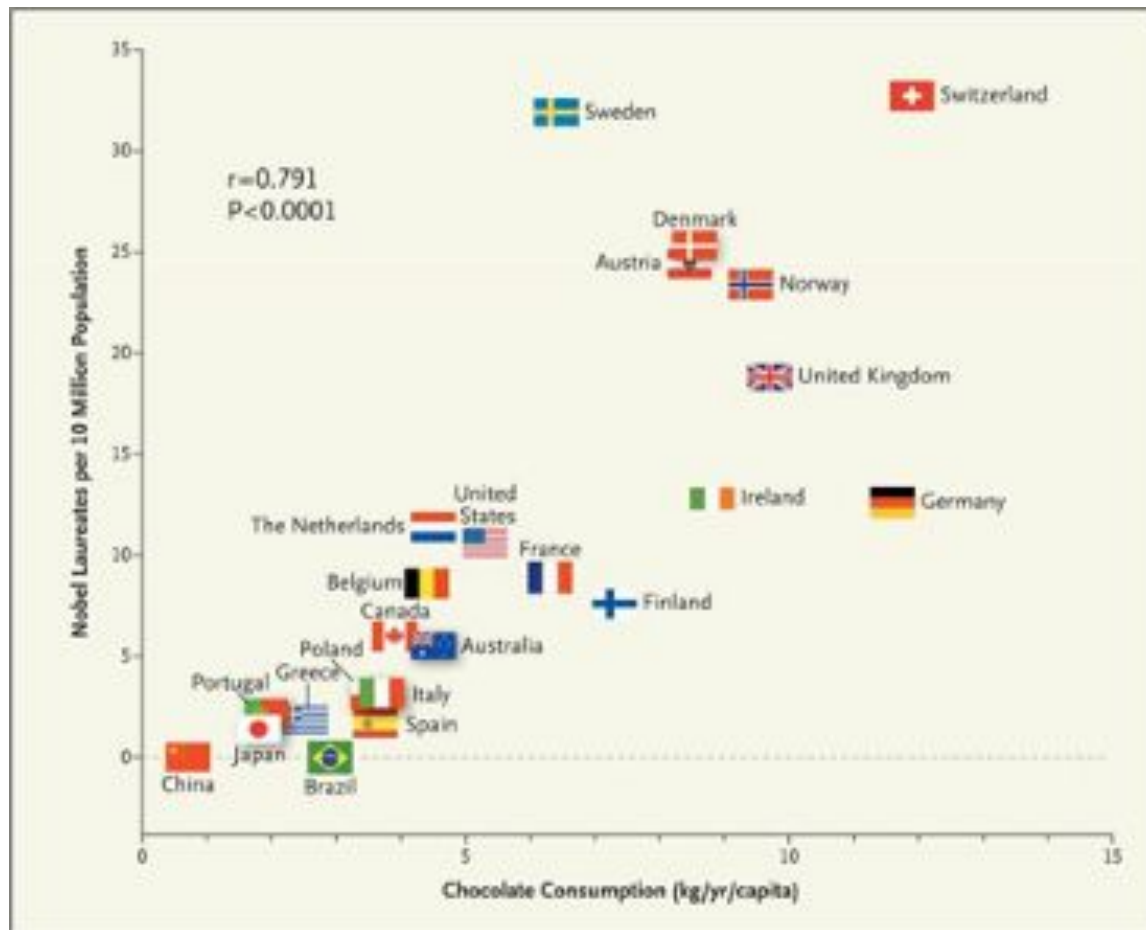
Residual standard error: 6.877 on 1701 degrees of freedom

Multiple R-squared: 0.7169, Adjusted R-squared: 0.7165

F-statistic: 2153 on 2 and 1701 DF, p-value: < 2.2e-16

What does it mean to say that X caused Y ?

- Low barometer reading predicts rain
- Experiment:
 - ?



Back to gapminder

- Approach: tell stories
 - Possible stories about the relationship of gdpPercap to lifeExp

Modeling outliers

- Model:
 - (on the whiteboard)
- Generate data from the model

Three approaches to regression

- Minimizing a cost function
- Maximum likelihood
- Maximum A-Posteriori

Teaching discussion

- Why we are starting with cost functions
- Going from cost functions to inference