Notes: MS 204 Chapter 1/5

Overview

- Multivariate thinking

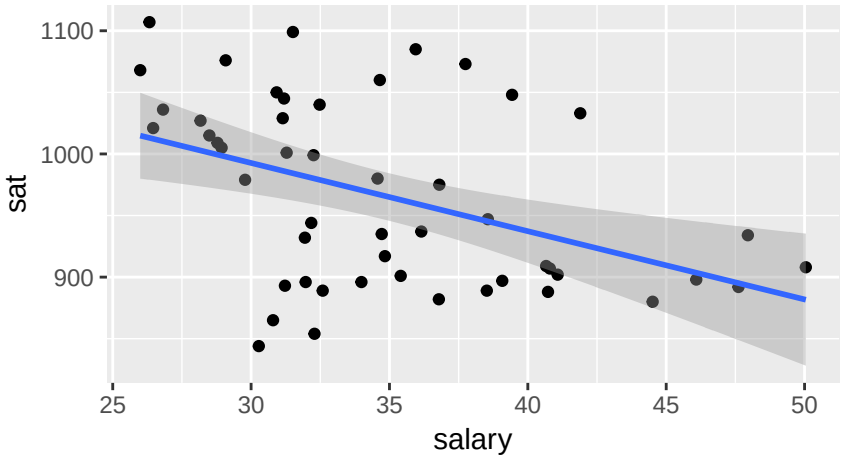
Simple linear regression

Ex: X = salary, Y = SAT

```
library(tidyverse); library(mosaic)
SAT %>% summarise(cor.SAT = cor(sat, salary))
```

```
##      cor.SAT
## 1 -0.4398834
```

```
qplot(x = salary, y = sat, data = SAT) +
  geom_smooth(method = "lm")
```




```
fit <- lm(sat ~ salary, data = SAT)
msummary(fit)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1158.859     57.659  20.098  < 2e-16 ***
## salary        -5.540      1.632  -3.394  0.00139 **
##
## Residual standard error: 67.89 on 48 degrees of freedom
## Multiple R-squared:  0.1935, Adjusted R-squared:  0.1767
## F-statistic: 11.52 on 1 and 48 DF,  p-value: 0.001391
```

- Describe the association between salary and sat on a state-level basis

- Identify possible explanations for this finding

- Write the estimated regression line

- Interpret the itercept and the slope in the context of this example

- Find the residual for Alabama, a state that pays teachers 31.1 thousand dollars per year and boasts an average SAT of 1029.

Confounding variables

Thinking multivariately

Ex: X = salary, Y = SAT

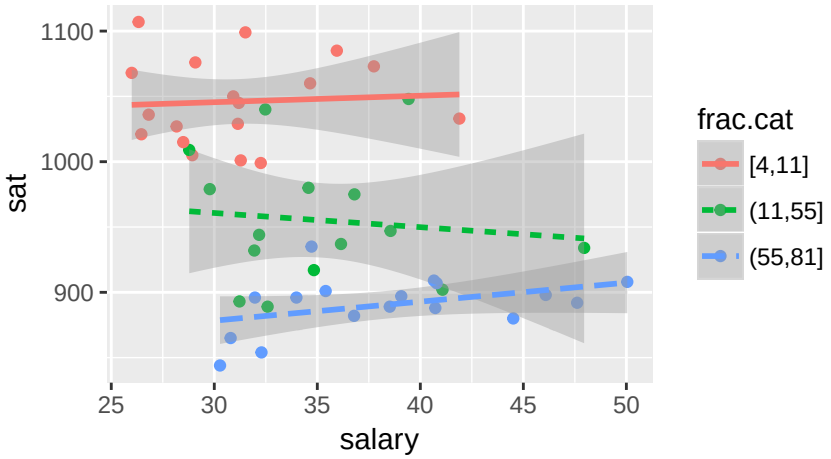
```
library(tidyverse); library(mosaic)
SAT <- SAT %>% mutate(frac.cat = cut_number(frac, 3))
SAT %>% tail(5)
```

```
##            state expend ratio salary frac verbal math  sat frac.cat
## 46      Virginia  5.327  14.6 33.987   65    428  468  896  (55,81]
## 47    Washington  5.906  20.2 36.151   48    443  494  937  (11,55]
## 48 West Virginia  6.107  14.8 31.944   17    448  484  932  (11,55]
## 49     Wisconsin  6.930  15.9 37.746    9    501  572 1073   [4,11]
## 50       Wyoming  6.160  14.9 31.285   10    476  525 1001   [4,11]
```

```
SAT %>%
  group_by(frac.cat) %>%
  summarise(cor.SAT = cor(sat, salary))
```

```
## # A tibble: 3 x 2
##   frac.cat     cor.SAT
##     <fctr>       <dbl>
## 1   [4,11]  0.06430064
## 2  (11,55] -0.10966031
## 3  (55,81]  0.40431822
```

```
qplot(x = salary, y = sat, colour = frac.cat, lty = frac.cat, data = SAT) +
  geom_smooth(method = "lm")
```



- Describe the association between salary and sat on a state-level basis when taking into account the fration of each state that took the SAT

- Possible explanations