

# Regression

## Part 1

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/10/11

Slides Updated: 2023-09-30

# Agenda

1. Modeling Conditional Variation
2. Adding Regression to the **Process**
3. Introducing the Data
4. Demonstrating Regressions

# Regression & Conditional Analysis

- Recall our discussion of **conditional** analysis
  - Conditional → **depends on**
  - Analyze with **conditional means**

# Reminder of the **Process**

## 1. Determine variable **type**

- I.e., categorical (ordered, unordered, binary) or continuous
- In R terms: `chr`, `fct` for categorical, `dbl` for continuous

## 2. Type informs **univariate analysis**

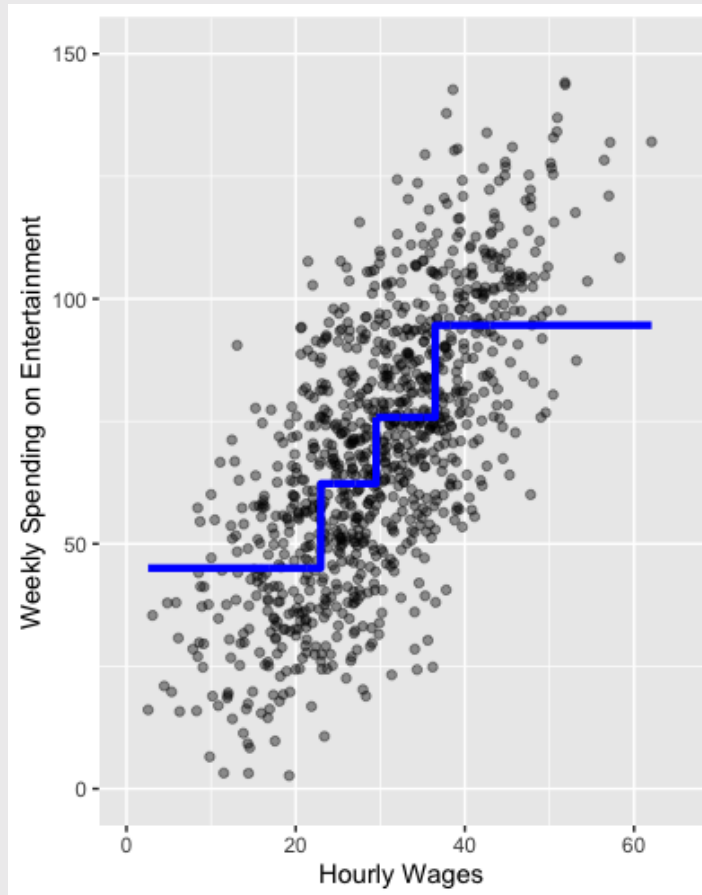
- I.e., histograms for continuous, barplots for categorical

## 3. Combination of types informs **conditional analysis**

- Categorical X Categorical: proportions by categories (`geom_bar`)
- Binary X Continuous: histograms by categories (`geom_histogram` / `geom_density`)
- Categorical X Continuous: distributions by categories (`geom_boxplot` / `geom_violin`)
- Continuous X Continuous: scatter plots (`geom_point`)

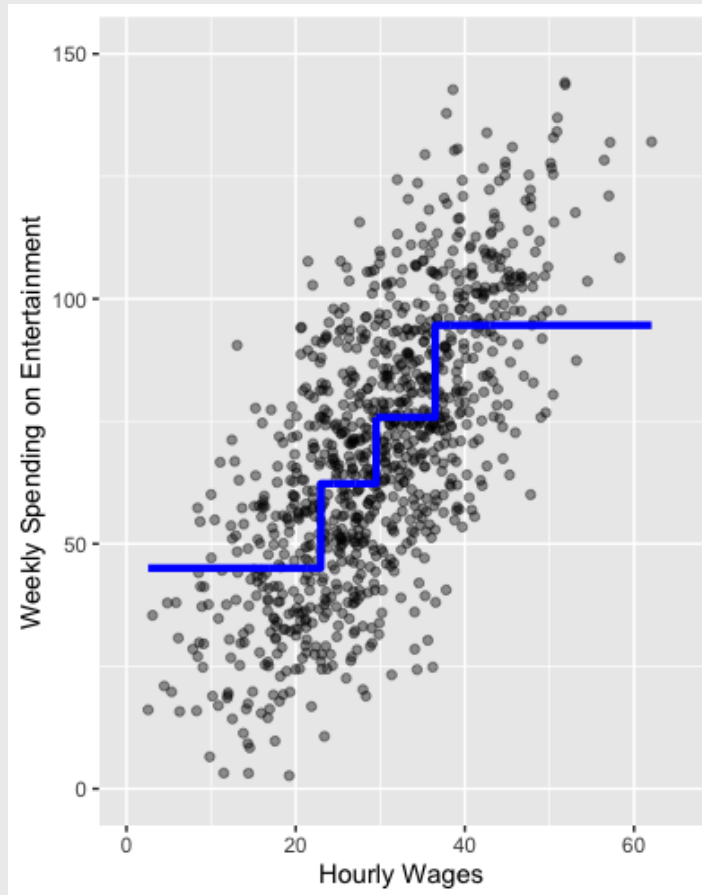
# What is regression?

- Conditional means for continuous data



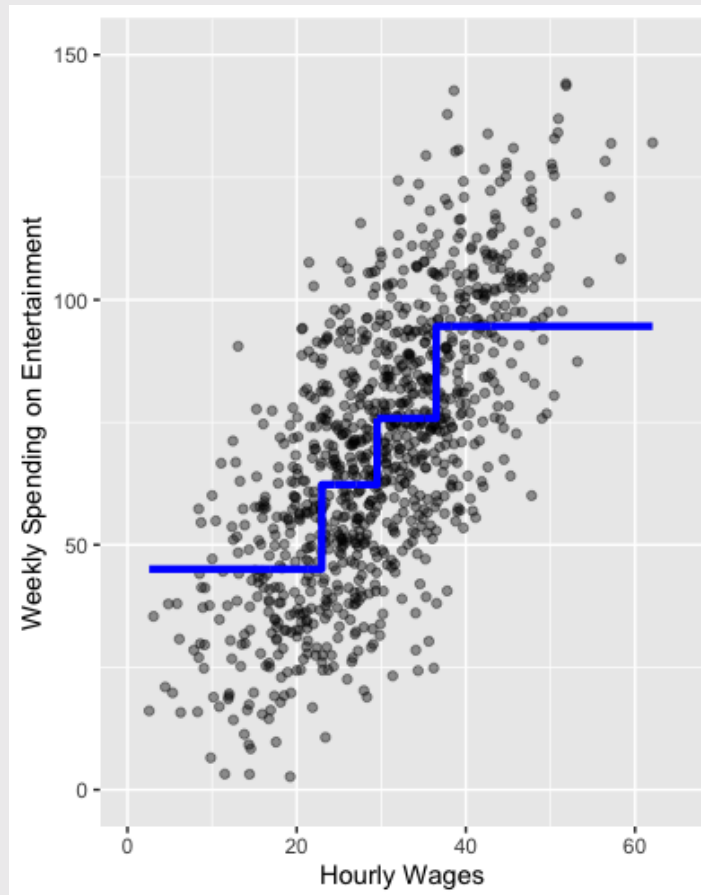
# What is regression?

- People with hourly wages < \$20 spend ~\$50 on entertainment per week



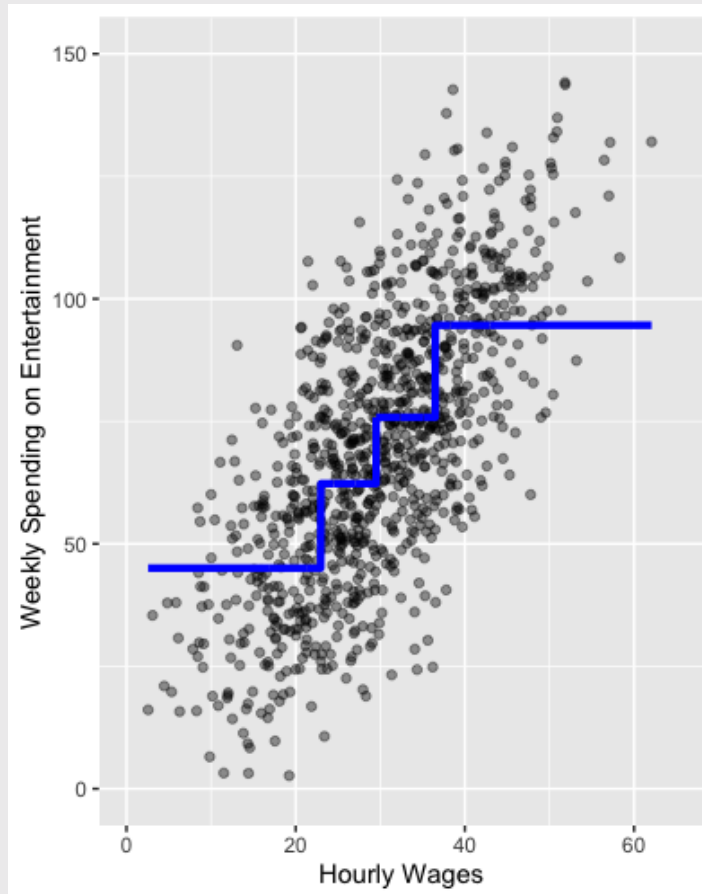
# What is regression?

- People with hourly wages  $> \$40$  spend  $\sim \$95$  on entertainment per week



# What is regression?

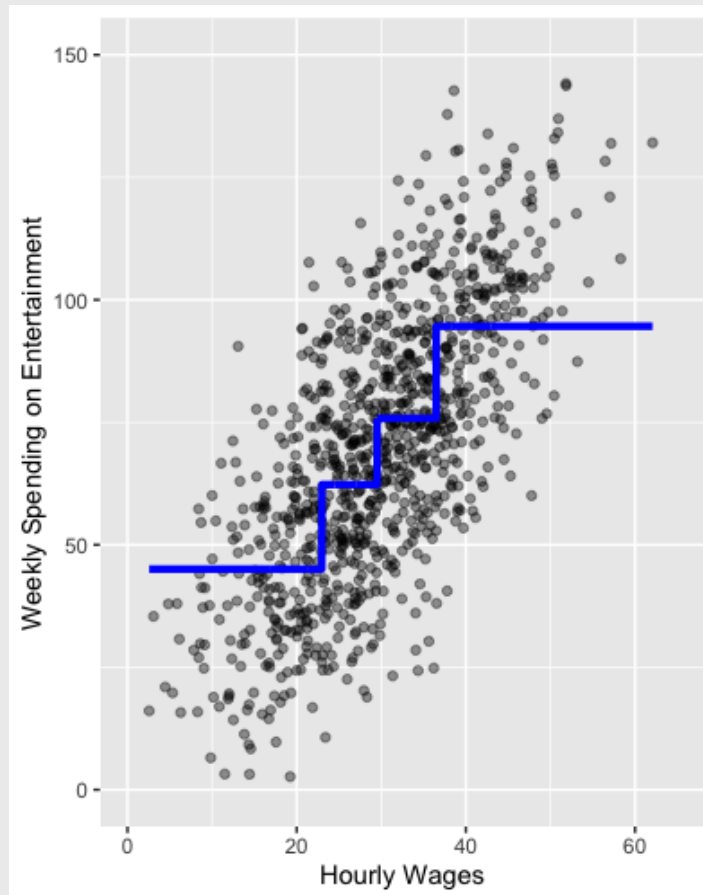
- **Theory**: the more you earn, the more you spend





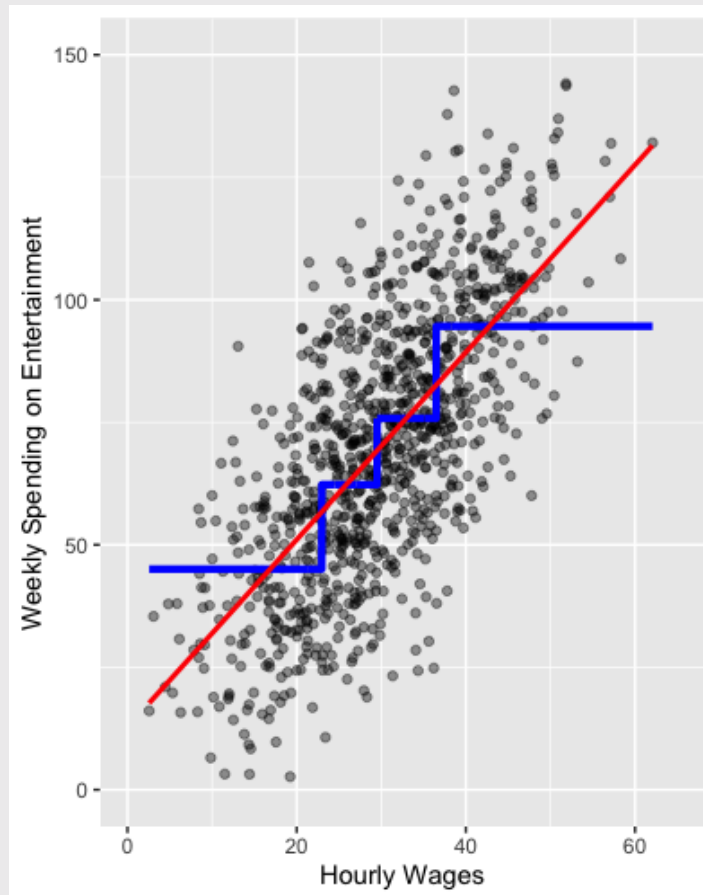
# What is regression?

- But **conditional means** make a lot of mistakes. Can we do better?



# What is regression?

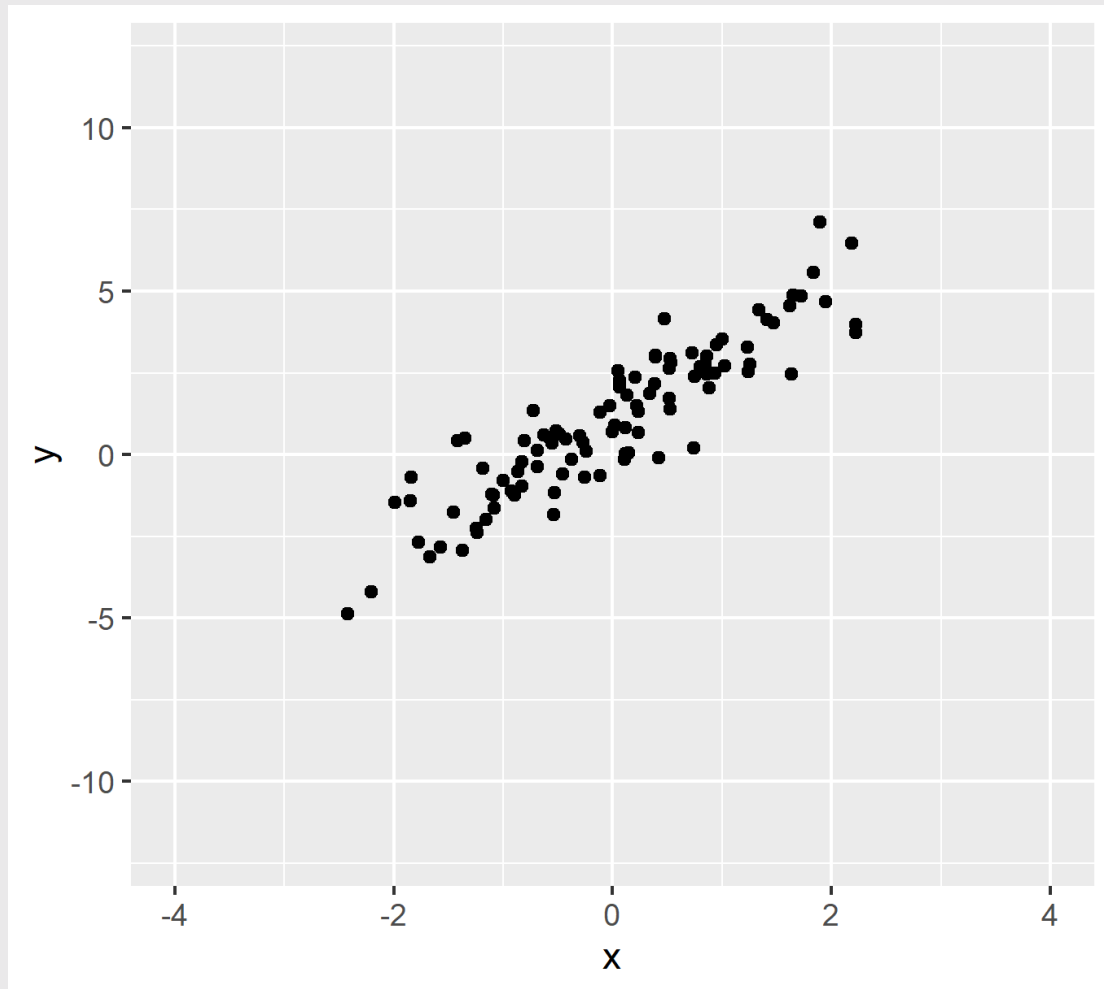
- But **conditional means** make a lot of mistakes. Can we do better?



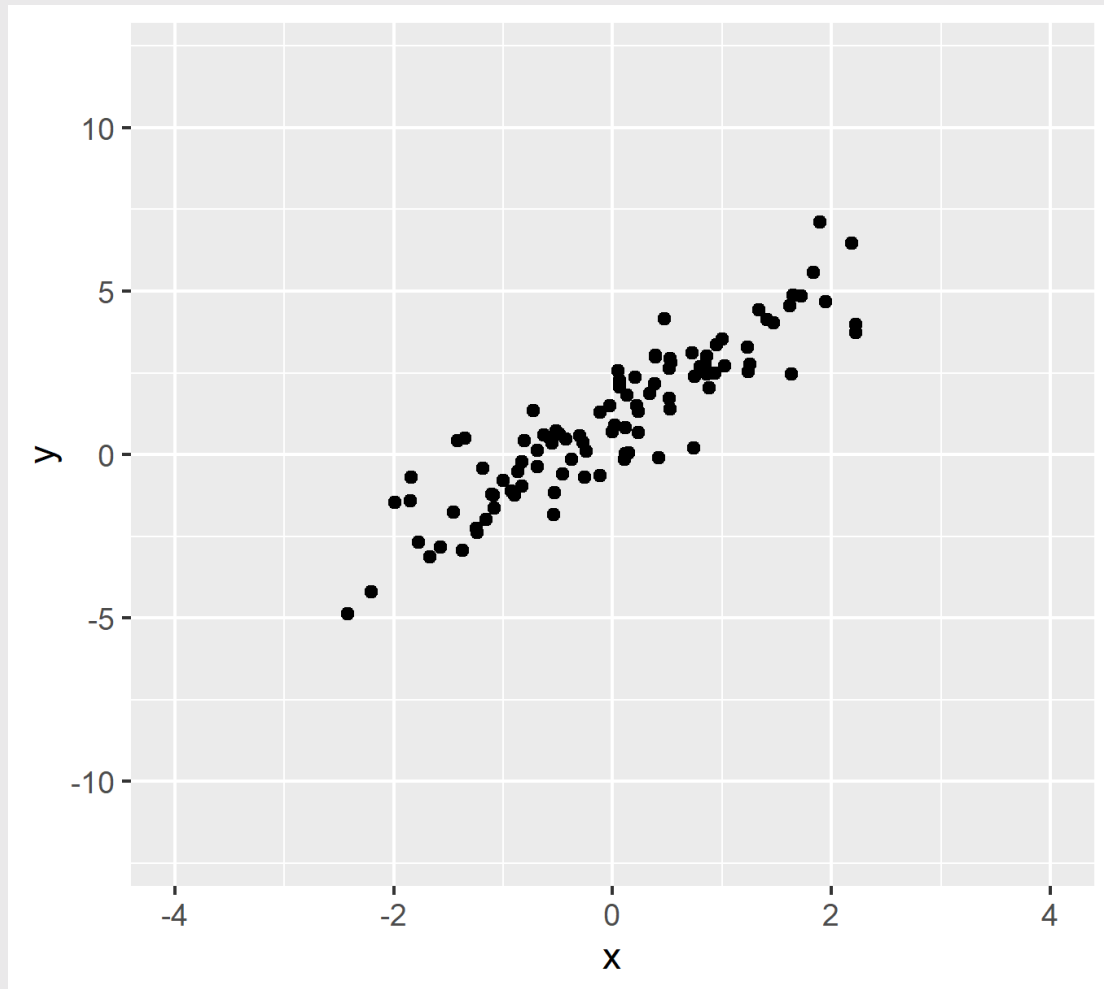
# Regression

- Calculating a **line** that minimizes mistakes *for every observation*
  - NB: could be a curvey line! For now, just assume straight
- Recall from geometry how to graph a straight line
- $Y = a + bX$ 
  - $a$ : the "intercept" (where the line intercepts the y-axis)
  - $b$ : the "slope" (how much  $Y$  changes for each increase in  $X$ )
- (Data scientists use  $\alpha$  and  $\beta$  instead of  $a$  and  $b$  b/c nerds)
- Regression analysis simply chooses the best line
  - "Best"?
  - The line that minimizes the mistakes (the **line of best fit**)

# Visual Intuition



# Visual Intuition



# Regression

- The line is **substantively meaningful**
- Red line on scatter plot of spending and wages:  $Y = 12 + 2 * X$
- $\alpha$  tells us the value of  $Y$  when  $X$  is zero
  - People who don't make any money spend \$12 per week on entertainment
- $\beta$  tells us how much  $Y$  increases for each additional  $X$ 
  - People spend an additional \$2 per week for each additional \$1 in hourly wages

# Regression

- These are called "**linear models**"
  - **Not** because the line is straight (it might not be)
  - but because the components are additive (  $\alpha + \beta X$  )
- Can extend to multiple predictors (  $X$  's)
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$
  - $X_1$  might be wages and  $X_2$  might be age (for example)
  - The final term  $\varepsilon$  measures how bad our mistakes are

# Regression

- Let's demonstrate with the `debt` data

```
require(tidyverse)

debt <- readRDS('../data/sc_debt.Rds')

glimpse(debt)
```

```
## Rows: 2,546
## Columns: 16
## $ unitid      <int> 100654, 100663, 100690, 100706, 100...
## $ instnm     <chr> "Alabama A & M University", "Univer...
## $ stabbr     <chr> "AL", "AL", "AL", "AL", "AL", "AL",...
## $ grad_debt_mdn <int> 33375, 22500, 27334, 21607, 32000, ...
## $ control    <chr> "Public", "Public", "Private", "Pub...
## $ region     <chr> "Southeast", "Southeast", "Southeas...
## $ preddeg    <chr> "Bachelor's", "Bachelor's", "Associ...
## $ openadmp   <int> 2, 2, 1, 2, 2, 2, 1, NA, 2, 2, 2, 1...
## $ adm_rate   <dbl> 0.9175, 0.7366, NA, 0.8257, 0.9690,...
## $ ccbasic    <int> 18, 15, 20, 16, 19, 15, 2, 22, 18, ...
## $ sat_avg    <int> 939, 1234, NA, 1319, 946, 1261, NA,...
## $ md_earn_wne_p6 <int> 25200, 35100, 30700, 36200, 22600, ...
```



# Research Camp

- **Research Question:** What is the relationship between SAT scores and median future earnings?
- **Theory:** Students with higher SAT scores work harder and have learned more. Employers reward these attributes with higher wages in the private market.
- **Hypothesis:** The relationship between SAT scores and future earnings should be positive.
  - **NB:** Important caveats to this simplistic theory!
  - Socioeconomic status: predicts both higher SAT scores and higher wages
  - **Correlation  $\neq$  Causation**

# Set Up

- Linking Theory to Data
- Our SAT scores are theorized to explain future earnings
  - Thus the SAT scores are the independent / explanatory / predictor variable  $X$
  - And earnings are the dependent / outcome variable  $Y$

# Regression

- There is a simple recipe to follow
- And it is exactly how the syllabus for the class is designed!
  1. Look at your data to identify **missingness**
  2. **Univariate** visualization of your variables
  3. **Multivariate** visualization of your variables
  4. **Regression** (today)
  5. Evaluation of **errors** (next lecture)

# Step 1: Look

- Why worry about **missingness**?

1. **Substantive:** external validity

2. **Technical:** cross validation won't work! (Wednesday's lecture)

```
summary(debt %>% select(sat_avg,md_earn_wne_p6))
```

```
##      sat_avg      md_earn_wne_p6
##  Min.      : 737    Min.      : 10600
## 1st Qu.:1053    1st Qu.: 26100
## Median :1119    Median : 31500
## Mean   :1141    Mean   : 33028
## 3rd Qu.:1205    3rd Qu.: 37400
## Max.   :1557    Max.   :120400
## NA's   :1317    NA's   :240
```

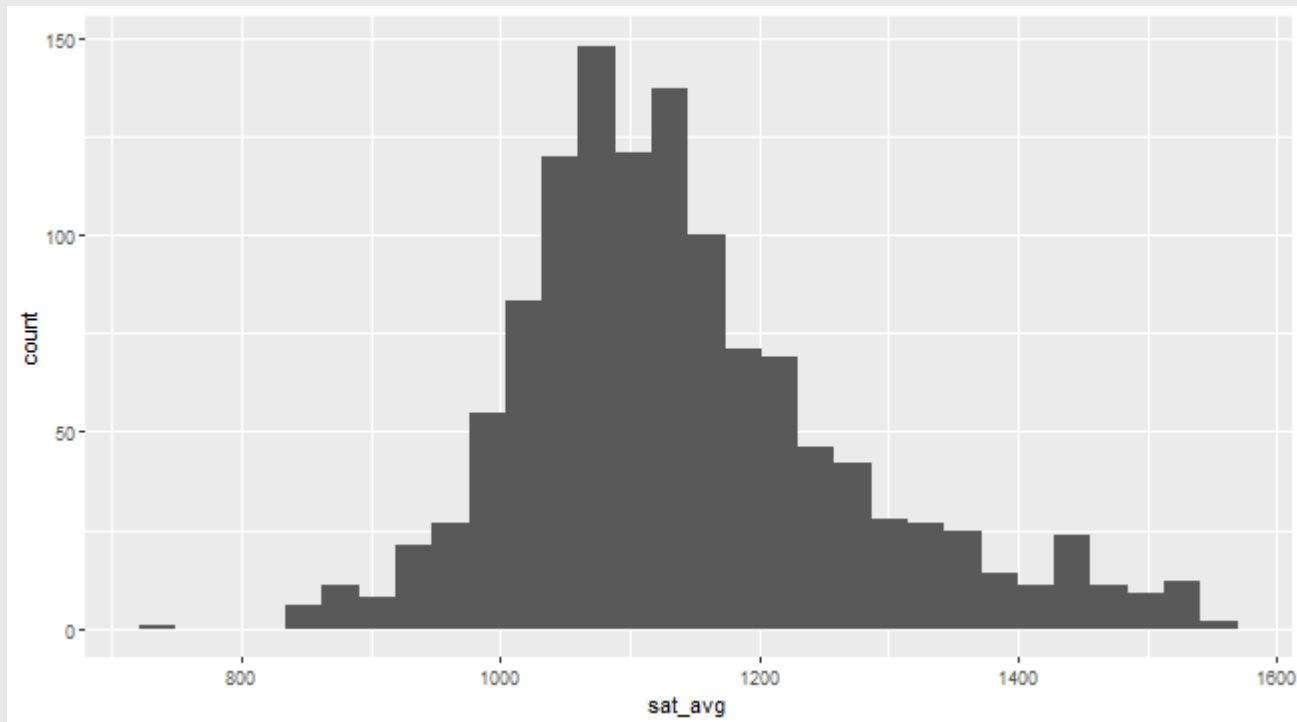
# Step 2: Univariate Viz

- Why visualize both  $Y$  and  $X$ ?
  1. **Substantive:** See which units you are talking about
  2. **Technical:** Adjust for *skew*

# Step 2: Univariate Viz

- Why visualize both  $Y$  and  $X$ ?

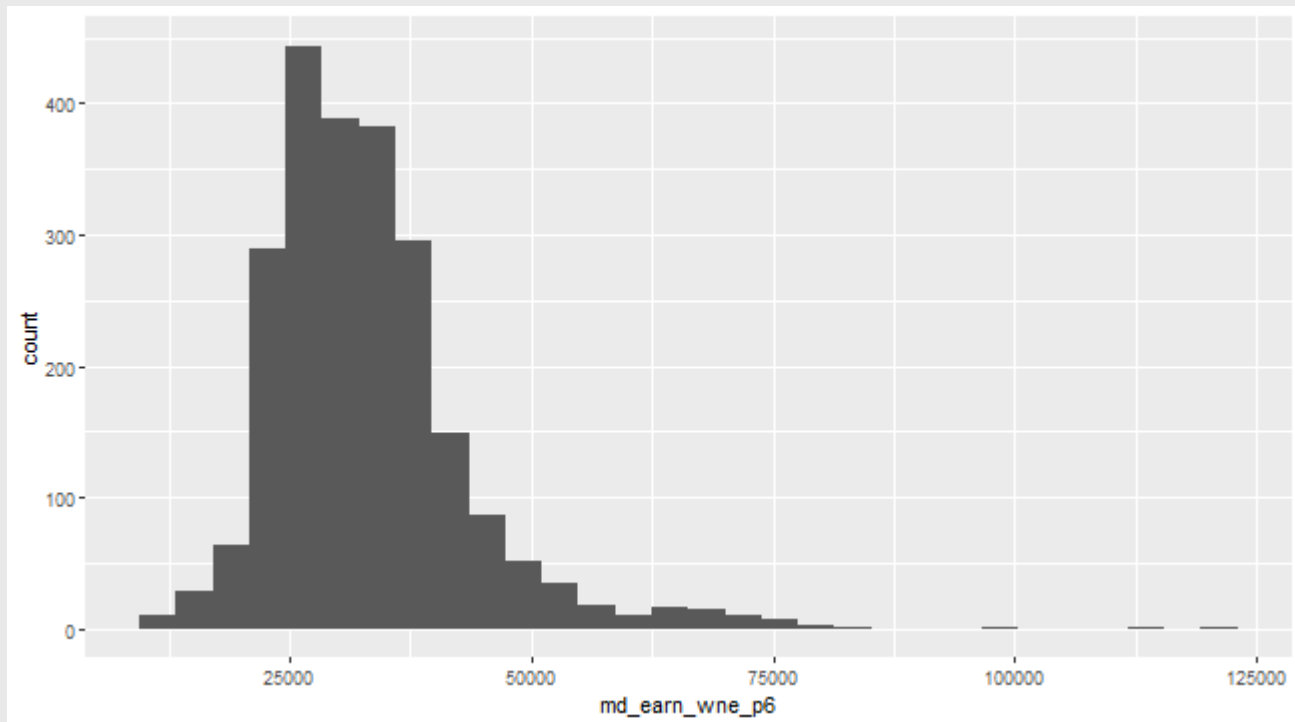
```
debt %>%  
  ggplot(aes(x = sat_avg)) +  
  geom_histogram()
```



# Step 2: Univariate Viz

- Why visualize both  $Y$  and  $X$ ?

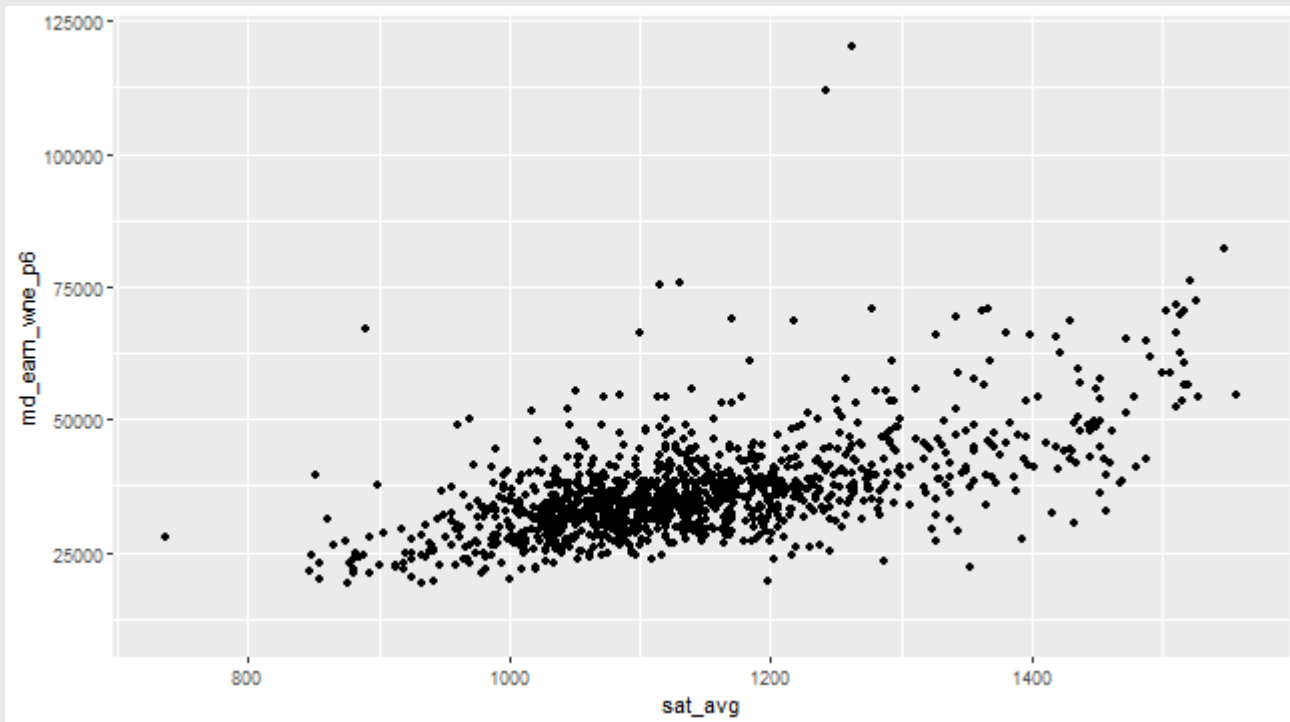
```
debt %>%  
  ggplot(aes(x = md_earn_wne_p6)) +  
  geom_histogram()
```



# Step 3: Multivariate

- Eyeball the relationship first!

```
debt %>%  
  ggplot(aes(x = sat_avg, y = md_earn_wne_p6)) +  
  geom_point()
```

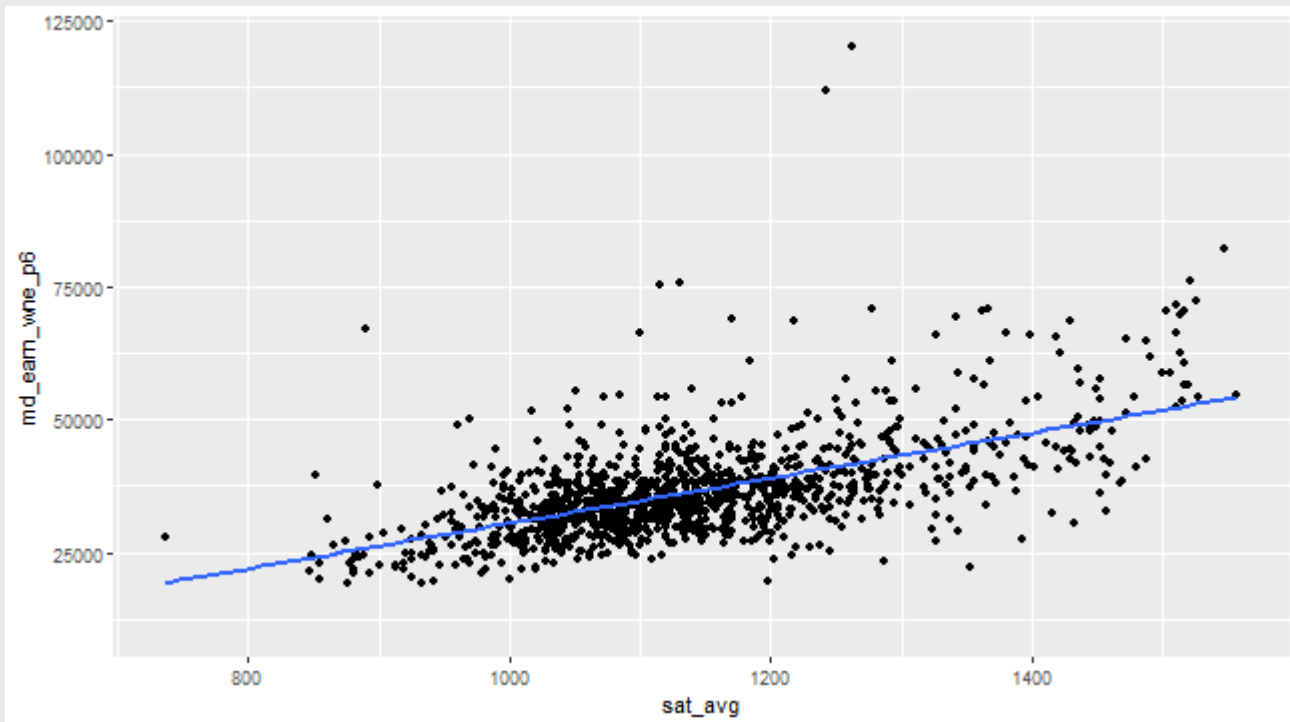




# Step 3: Multivariate Viz

- Adding regression line

```
debt %>%  
  ggplot(aes(x = sat_avg, y = md_earn_wne_p6)) +  
  geom_point() + geom_smooth(method = 'lm', se = F)
```



# Step 3: Multivariate Viz

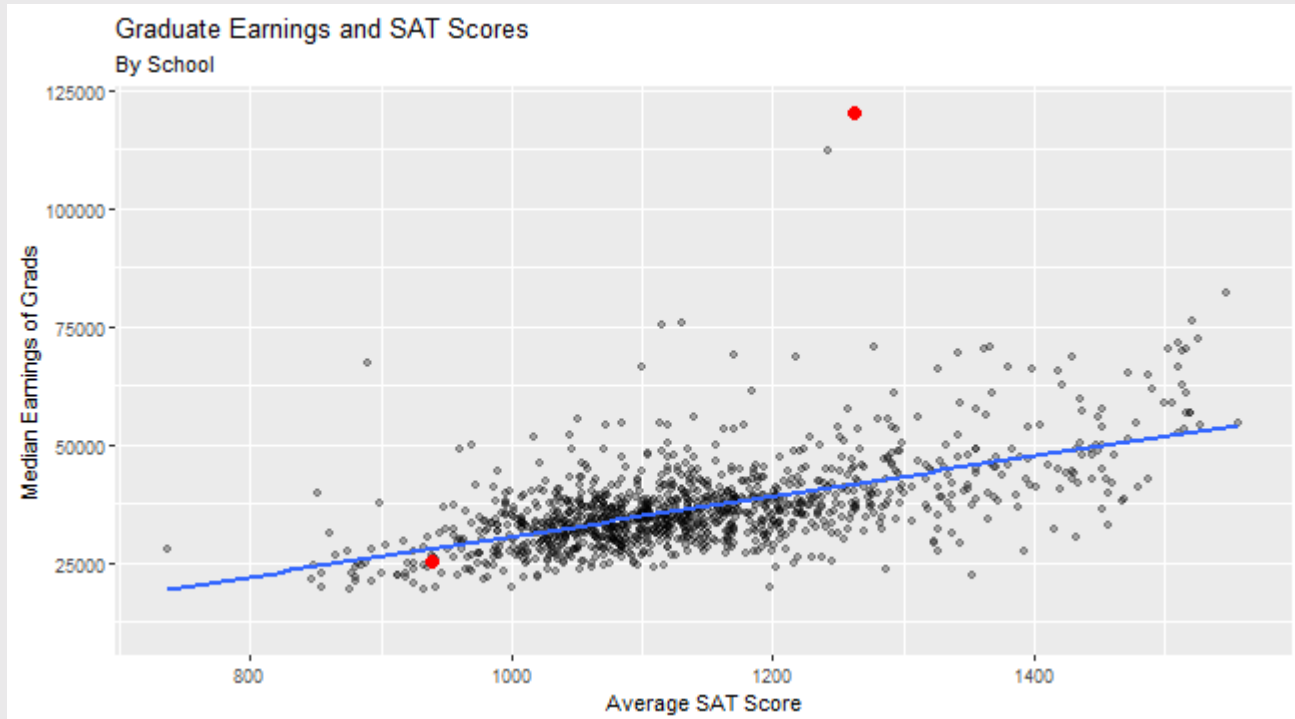
- Let's focus on two schools

```
toplot <- debt %>%
  mutate(h1 = ifelse(unitid %in% c(100654,179265), 'h1', 'none')) #
Choosing two examples
p2 <- toplot %>%
  ggplot(aes(x = sat_avg, y = md_earn_wne_p6,color = h1,group =
1,alpha = h1)) +
  geom_point(data = toplot %>% filter(h1 == 'none')) +
  geom_point(data = toplot %>% filter(h1 == 'h1'),size =3) +
  scale_alpha_manual(values = c(1,.3)) +
  scale_color_manual(values = c('red','black')) +
  geom_smooth(method = 'lm',se = F) +
  theme(legend.position = 'none') +
  labs(title = "Graduate Earnings and SAT Scores",
        subtitle = "By School",
        x = "Average SAT Score",
        y = "Median Earnings of Grads")
```

# Step 3: Multivariate Viz

- Adding regression line

p2



# Step 3: Multivariate Viz

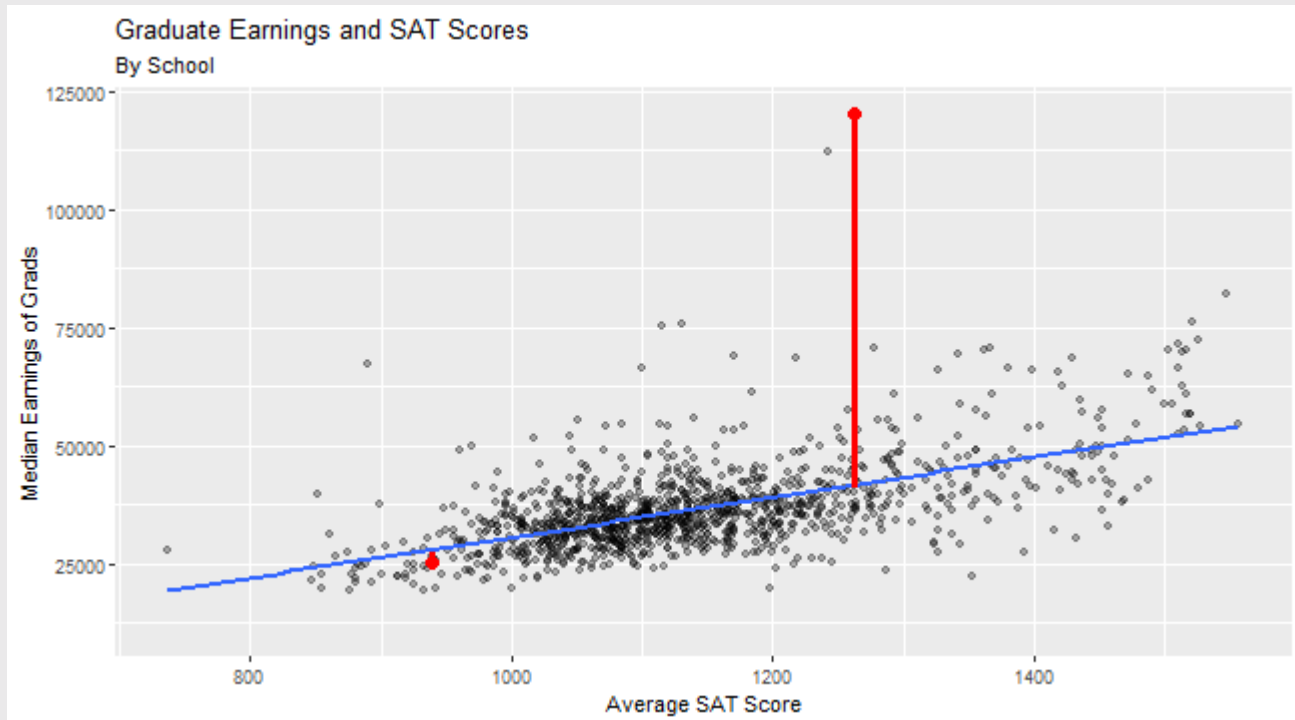
- Defining  $\varepsilon$

```
p3 <- toplot %>%
  ggplot(aes(x = sat_avg, y = md_earn_wne_p6,color = h1,group =
1,alpha = h1)) +
  geom_point(data = toplot %>% filter(h1 == 'none')) +
  geom_point(data = toplot %>% filter(h1 == 'h1'),size =3) +
  scale_alpha_manual(values = c(1,.3)) +
  scale_color_manual(values = c('red','black')) +
  geom_smooth(method = 'lm',se = F) +
  annotate(geom = 'segment',
    x = toplot %>% filter(h1 == 'h1') %>% .$sat_avg,
    y = toplot %>% filter(h1 == 'h1') %>% .$md_earn_wne_p6,
    xend = toplot %>% filter(h1 == 'h1') %>% .$sat_avg,
    yend = c(27500,41000),color = 'red',lwd = 1.2) +
  theme(legend.position = 'none') +
  labs(title = "Graduate Earnings and SAT Scores",
    subtitle = "By School",
    x = "Average SAT Score",
    y = "Median Earnings of Grads")
```

# Step 3: Multivariate Viz

- Measuring errors

p3

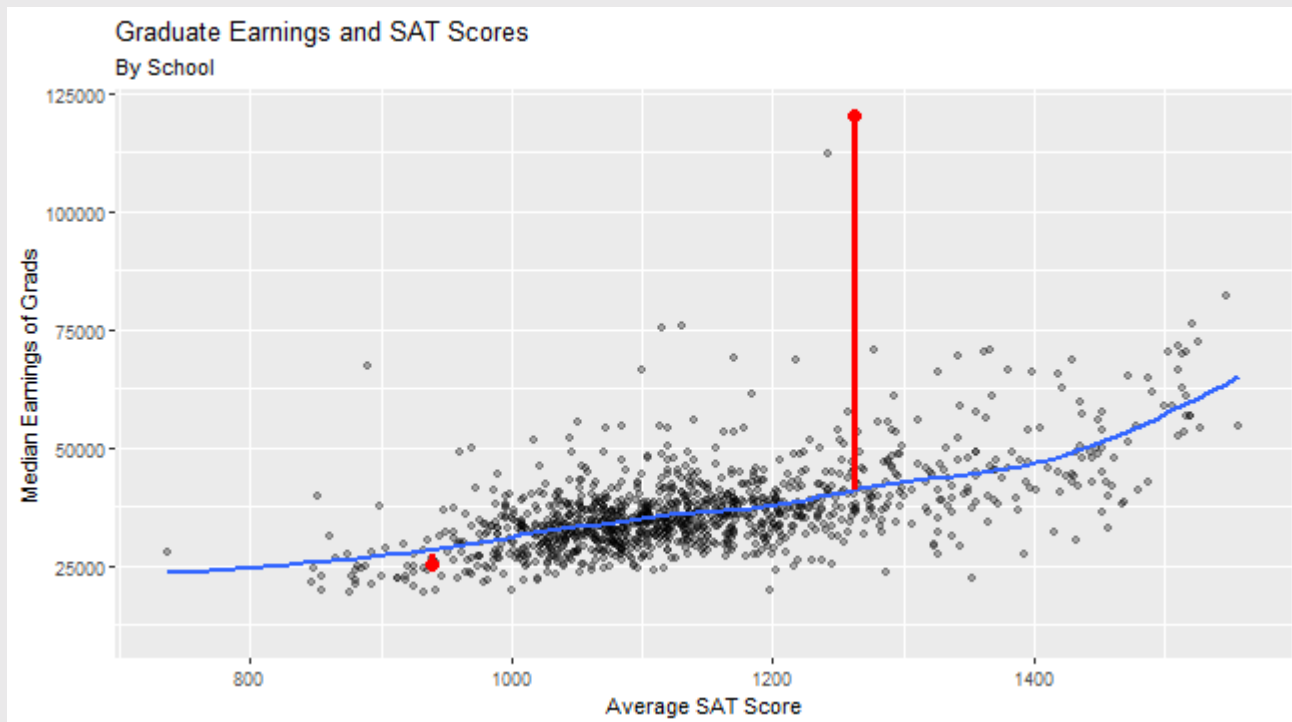


# The Data Scientist's Trade-off

- Those mistakes seem pretty big!
- Why not use a curvier line?

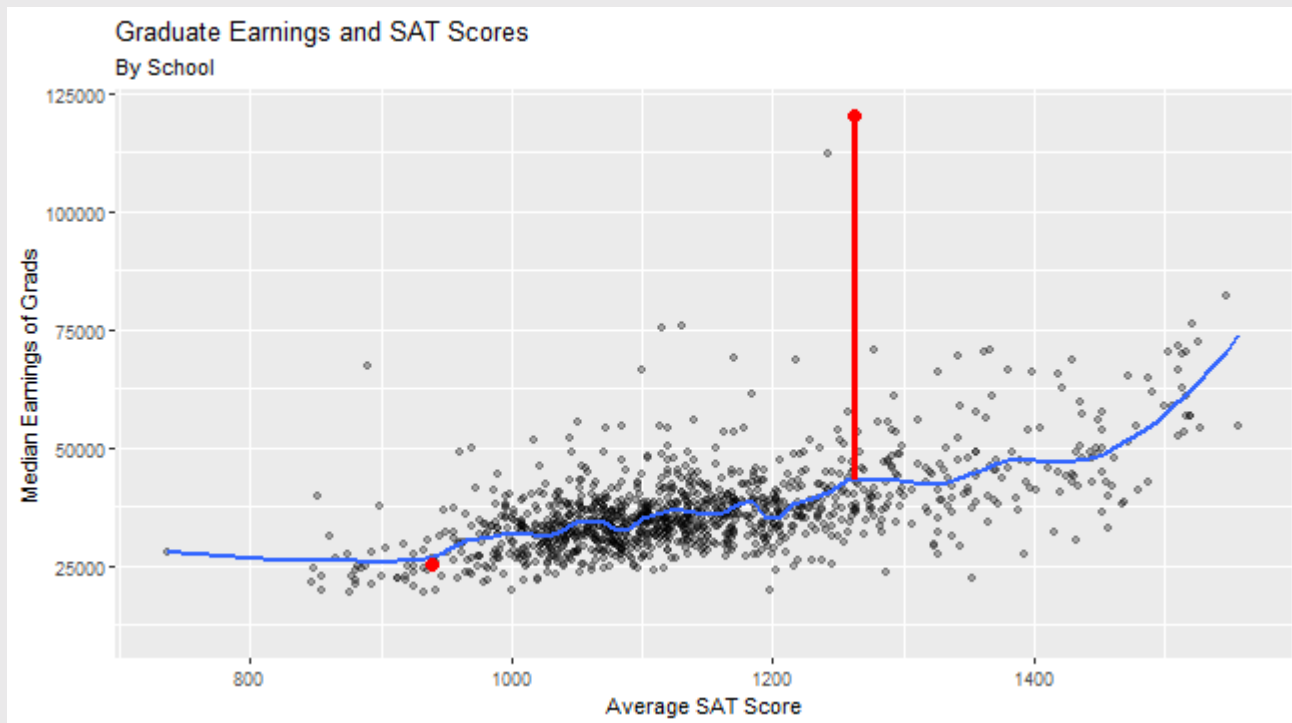
# The Data Scientist's Trade-off

- Those mistakes seem pretty big!
- Why not use a curvier line?



# The Data Scientist's Trade-off

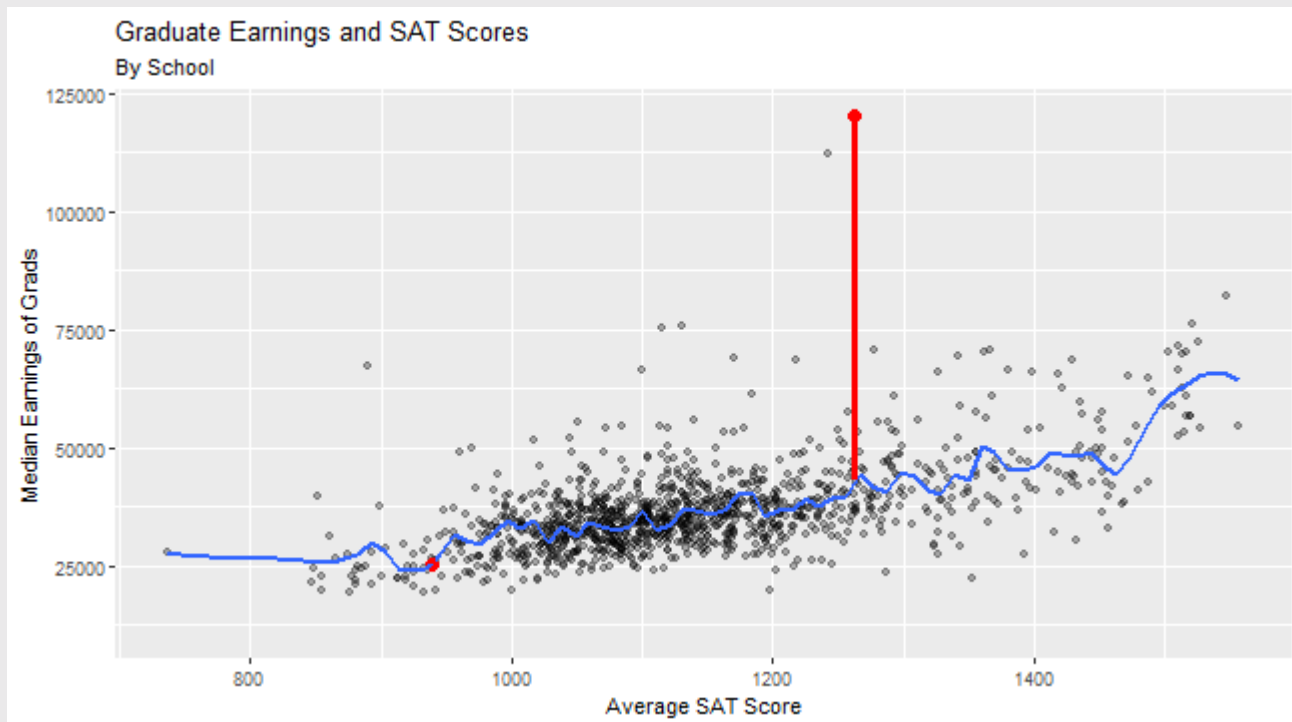
- Those mistakes seem pretty big!
- Why not use a curvier line?





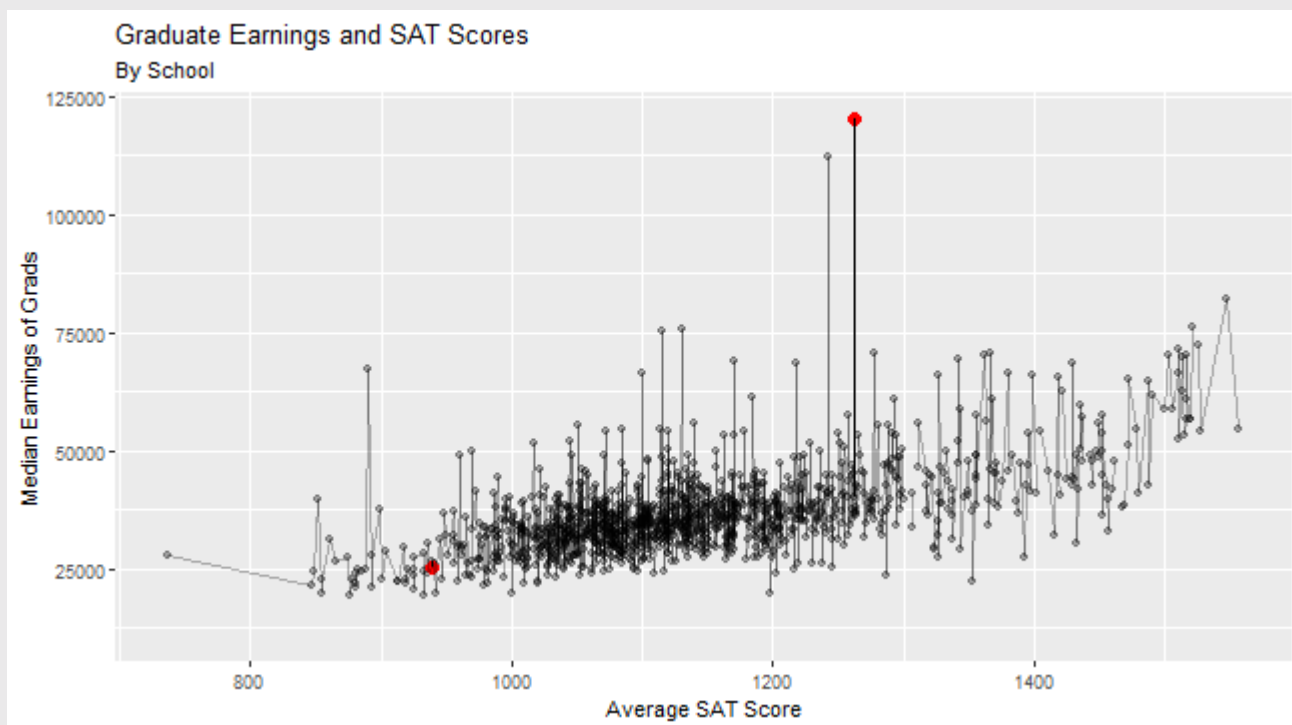
# The Data Scientist's Trade-off

- Those mistakes seem pretty big!
- Why not use a curvier line?



# The Data Scientist's Trade-off

- Those mistakes seem pretty big!
- Why not use a curvier line?



# The Data Scientist's Trade-off

- Want to **reduce complexity**
- But also want to be **accurate**
- What is the right answer?
  - It depends on your **theory** and the **data**
  - It is context-dependent
- And this is still only using *linear regression models*!
  - This is a deep area of study, for those interested

# Step 4: Regression

- Introducing the `lm(formula,data)` function
- Two inputs to care about:
  - `formula`: Code for  $Y = \alpha + \beta X$
  - `data`: What is the data we are using?
- `formula` is written as  $Y \sim X$ 
  - `R` will calculate  $\alpha$  and  $\beta$  for us
  - Just need to tell it what is  $Y$  (`md_earn_wne_p6`) and  $X$  (`sat_avg`)
  - The tilde (`~`) is `R`'s version of the equals sign
- Save the model to an object

```
model_earn_sat <- lm(formula = md_earn_wne_p6 ~ sat_avg, data = debt)
```

# Step 4: Interpretation

- What is in this object?
- The regression results! Look at them with `summary()`

```
summary(model_earn_sat)
```

```
##
## Call:
## lm(formula = md_earn_wne_p6 ~ sat_avg, data = debt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23239  -4311   -852    2893   78695
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12053.87    1939.80  -6.214 7.12e-10 ***
## sat_avg      42.60       1.69   25.203 < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

# Step 4: Interpretation

- Starting with the first column called **Estimate**
- 1st Row (**Intercept**) is  $\alpha$ : the predicted value of  $Y$  when  $X$  is zero
  - Schools with average SAT scores of 0 produce graduates who earn -\$12,053.87
  - Sensible?
- 2nd Row **sat\_avg** is the  $\beta$ : the increase in  $Y$  when  $X$  increases by one
  - For each unit increase in the average SAT score, recent graduates earn \$42.60 more
  - Sensible?

# Step 4: Interpretation

- Other 3 columns?
  - `Std. Error` is the "standard error"
  - `t value` is the "t-statistic"
  - `Pr(>|t|)` is the "p-value"
- $t\text{-statistic} = \text{Estimate} / \text{standard error}$
- $p\text{-value} = \text{function}(t\text{-statistic})$ 
  - Only really need to remember the p-value for this course
  - This is 1 minus confidence
  - The lower the p-value, the more confident we are that the `Estimate` is not zero

# Step 4: Interpretation

```
summary(model_earn_sat)
```

```
##
## Call:
## lm(formula = md_earn_wne_p6 ~ sat_avg, data = debt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23239  -4311   -852    2893   78695
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12053.87    1939.80  -6.214 7.12e-10 ***
## sat_avg       42.60       1.69   25.203 < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7594 on 1196 degrees of freedom
## (1348 observations deleted due to missingness)
## Multiple R-squared:  0.3469,    Adjusted R-squared:  0.3463
## F-statistic: 635.2 on 1 and 1196 DF,  p-value: < 2.2e-16
```



# Another Example

- We will come back to the RMSE next class
- For now, let's try with a different research question!
- What is the relationship between admissions and future earnings?
  - Theory: More selective schools are more prestigious
  - Hypothesis: There should be a negative relationship between the admissions rate and future earnings

# Do It Together!

1. Look at the data and acknowledge missingness
2. Univariate visualization of  $X$  and  $Y$
3. Multivariate visualization of  $X$  and  $Y$
4. Regression

# Quiz & Homework

- Go to Brightspace and take the **10th** quiz
  - The password to take the quiz is ####
- **Homework:**
  1. Work through Regression\_part1\_hw.Rmd