Lecture 12 Notes

2024-02-27

Regression examples

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

```
debt <- read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/sc_debt.Rds')

glimpse(debt)
```

```
## Rows: 2,546
## Columns: 16
## $ unitid         <int> 100654, 100663, 100690, 100706, 100724, 100751, 100760,…
## $ instnm         <chr> "Alabama A & M University", "University of Alabama at B…
## $ stabbr         <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "…
## $ grad_debt_mdn  <int> 33375, 22500, 27334, 21607, 32000, 23250, 12500, 19500,…
## $ control        <chr> "Public", "Public", "Private", "Public", "Public", "Pub…
## $ region         <chr> "Southeast", "Southeast", "Southeast", "Southeast", "So…
## $ preddeg        <chr> "Bachelor's", "Bachelor's", "Associate", "Bachelor's", …
## $ openadmp       <int> 2, 2, 1, 2, 2, 2, 1, NA, 2, 2, 2, 1, 1, 2, 1, 1, 2, 2, …
## $ adm_rate       <dbl> 0.9175, 0.7366, NA, 0.8257, 0.9690, 0.8268, NA, NA, 0.9…
## $ ccbasic        <int> 18, 15, 20, 16, 19, 15, 2, 22, 18, 15, 21, 1, 5, 19, 7,…
## $ sat_avg        <int> 939, 1234, NA, 1319, 946, 1261, NA, NA, 1082, 1300, 123…
## $ md_earn_wne_p6 <int> 25200, 35100, 30700, 36200, 22600, 37400, 23100, 33400,…
## $ ugds           <int> 5271, 13328, 365, 7785, 3750, 31900, 1201, 2677, 4407, …
## $ costt4_a       <int> 23053, 24495, 14800, 23917, 21866, 29872, 10493, NA, 19…
## $ selective      <dbl> 0, 0, NA, 0, 0, 0, NA, NA, 0, 0, 0, NA, NA, 0, NA, NA, …
## $ research_u     <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

Running first regression

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

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
```

```
# Predicting SAT scores of 400
-12053.9 + 42.6*400
```

```
## [1] 4986.1
```

```
# Predicting Vandy's future earnings
debt %>%
  filter(grepl('Vanderbil',instnm)) %>%
  select(instnm,sat_avg,md_earn_wne_p6)
```

```
## # A tibble: 1 × 3
##   instnm                sat_avg md_earn_wne_p6
##   <chr>                   <int>          <int>
## 1 Vanderbilt University    1515          53400
```

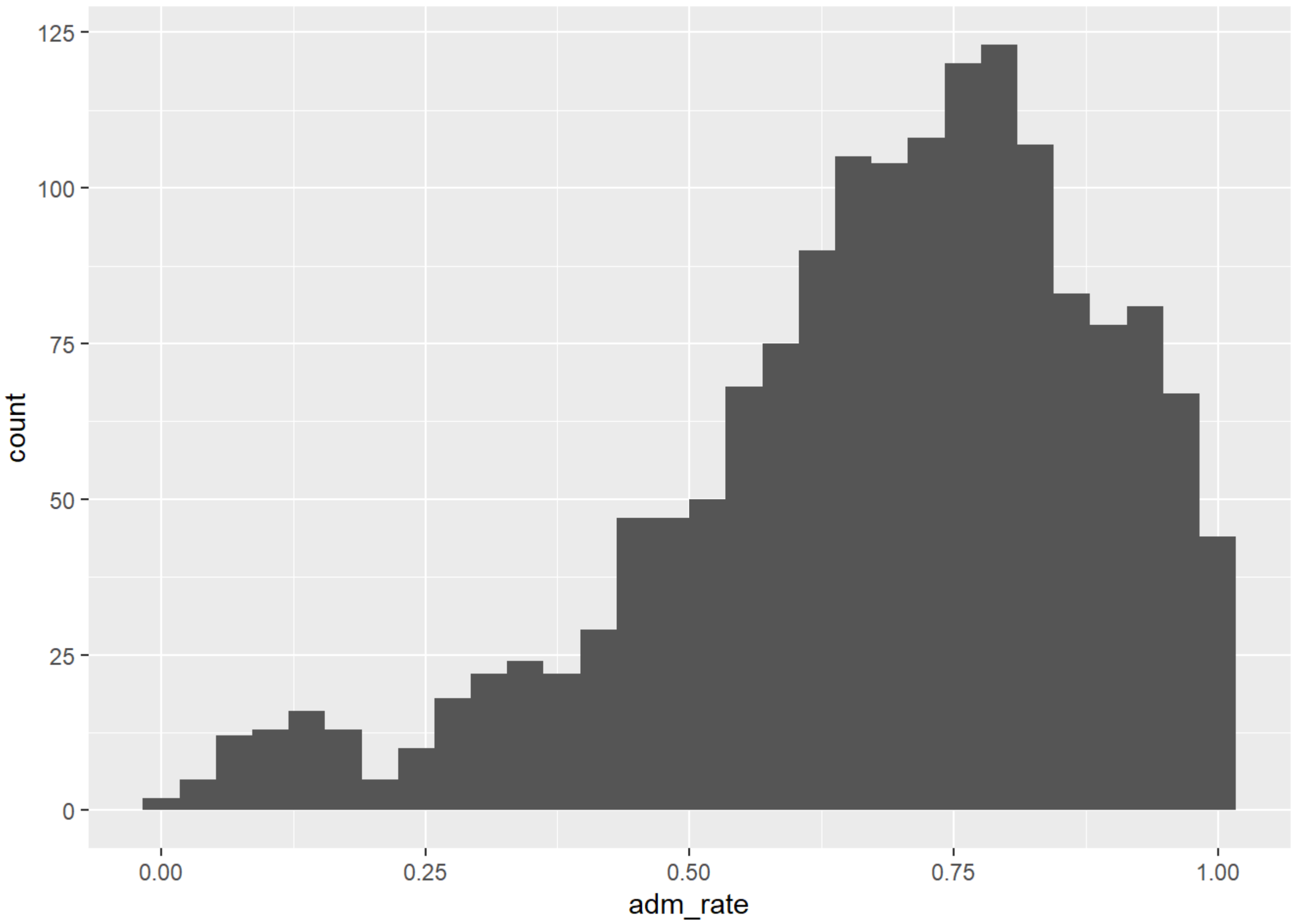
```
-12053.9 + 42.6*1515
```

```
## [1] 52485.1
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 958 rows containing non-finite values (`stat_bin()`).
```



```
lm(formula = md_earn_wne_p6 ~ adm_rate,
   data = debt)
```

```
##
## Call:
## lm(formula = md_earn_wne_p6 ~ adm_rate, data = debt)
##
## Coefficients:
## (Intercept)     adm_rate
##       44274       -12232
```