STA 235H - Multiple Regression: Outliers

Fall 2023

McCombs School of Business, UT Austin

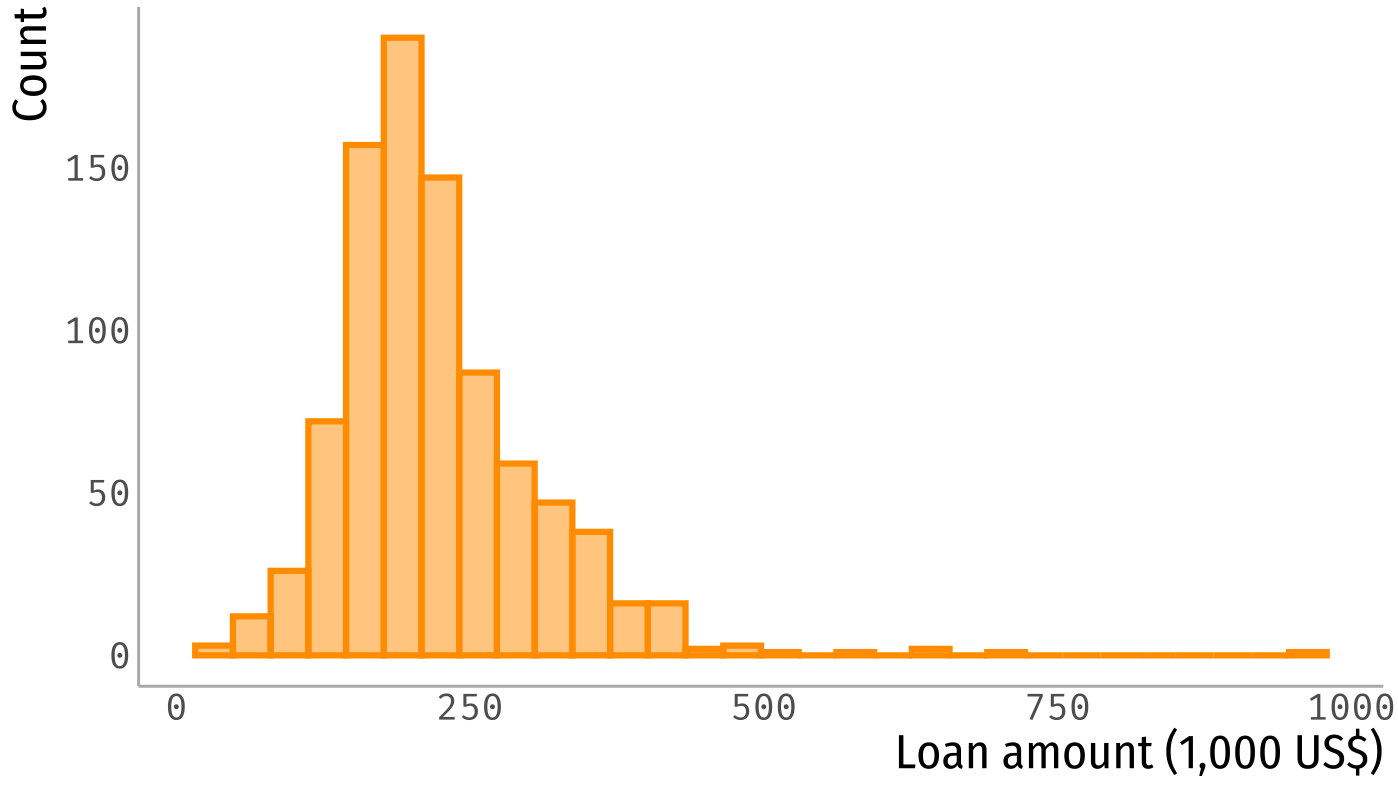
Why should we inspect our data before doing anything else?

Identifying outliers

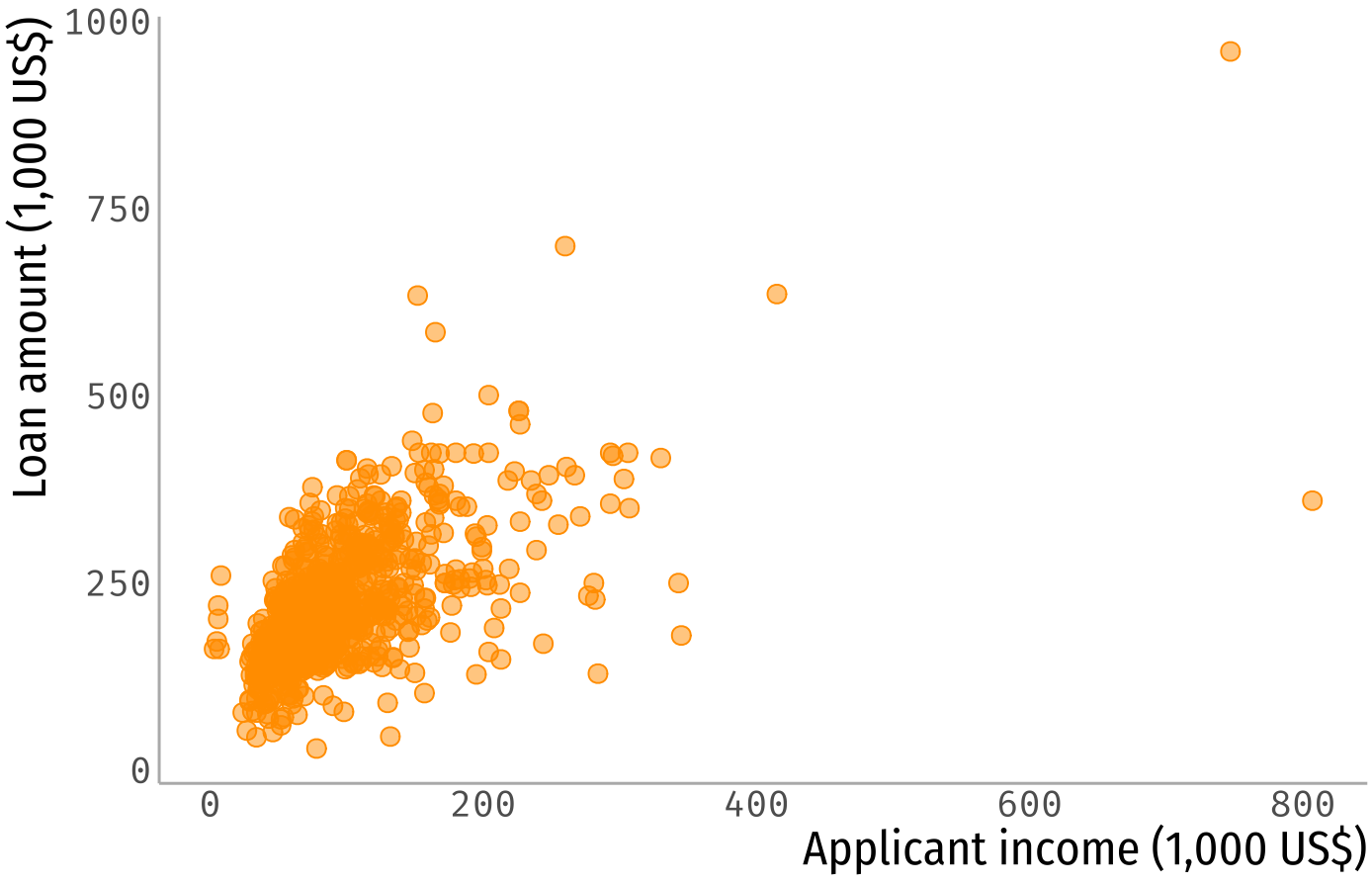
- How do we identify outliers?
  - Visual inspection (e.g. plots, tables)
  - Creating thresholds (e.g. z-scores, IQ)
- There is no definite way to identify outliers
  - Like the characterization of pornography, "I know it when I see it" (P. Stewart, 1964)

HMDA Data for Bastrop County

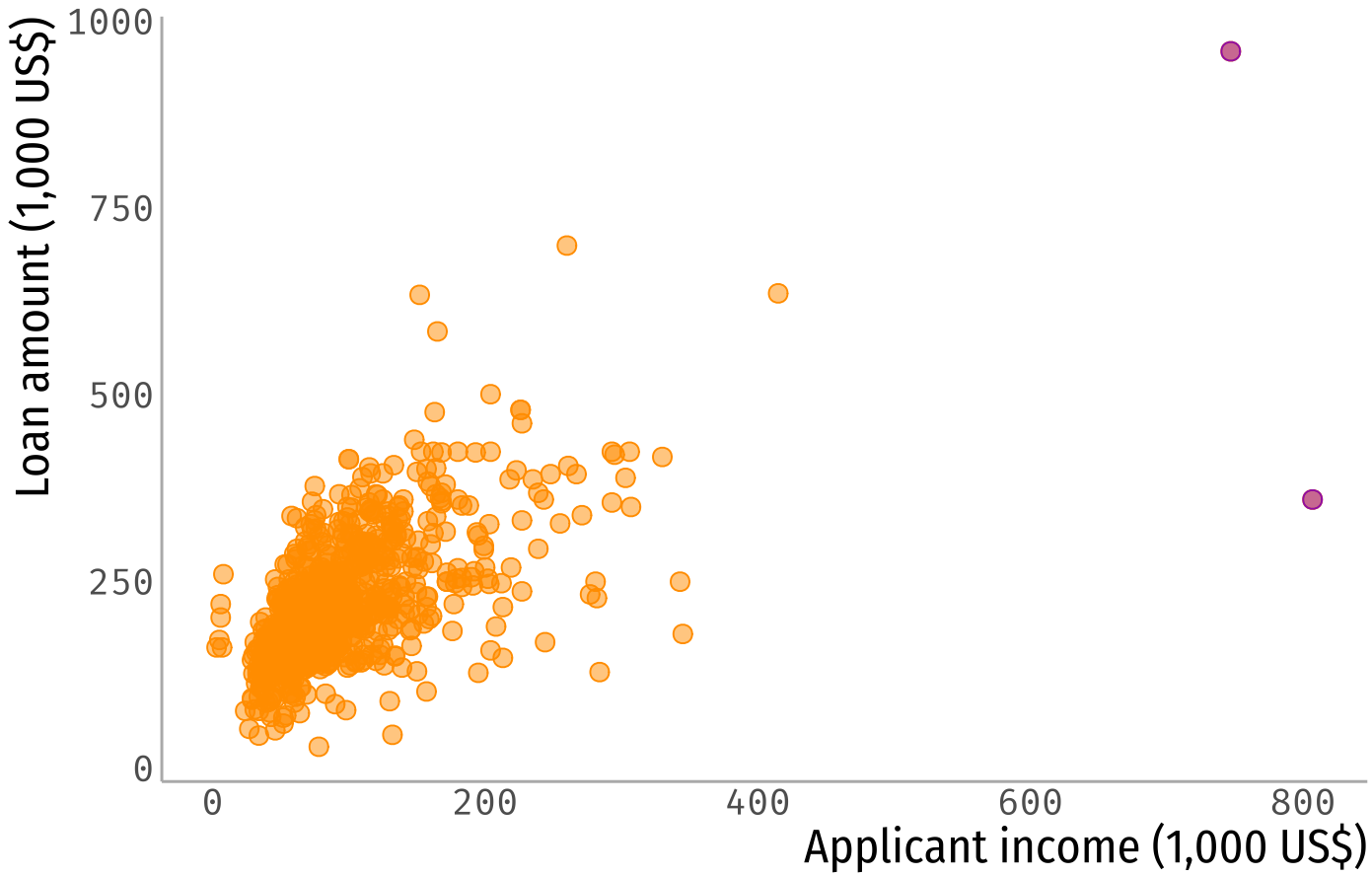
- Data from the Home Mortgage Disclosure Act (HMDA) from 2017 in Bastrop County (near Austin)



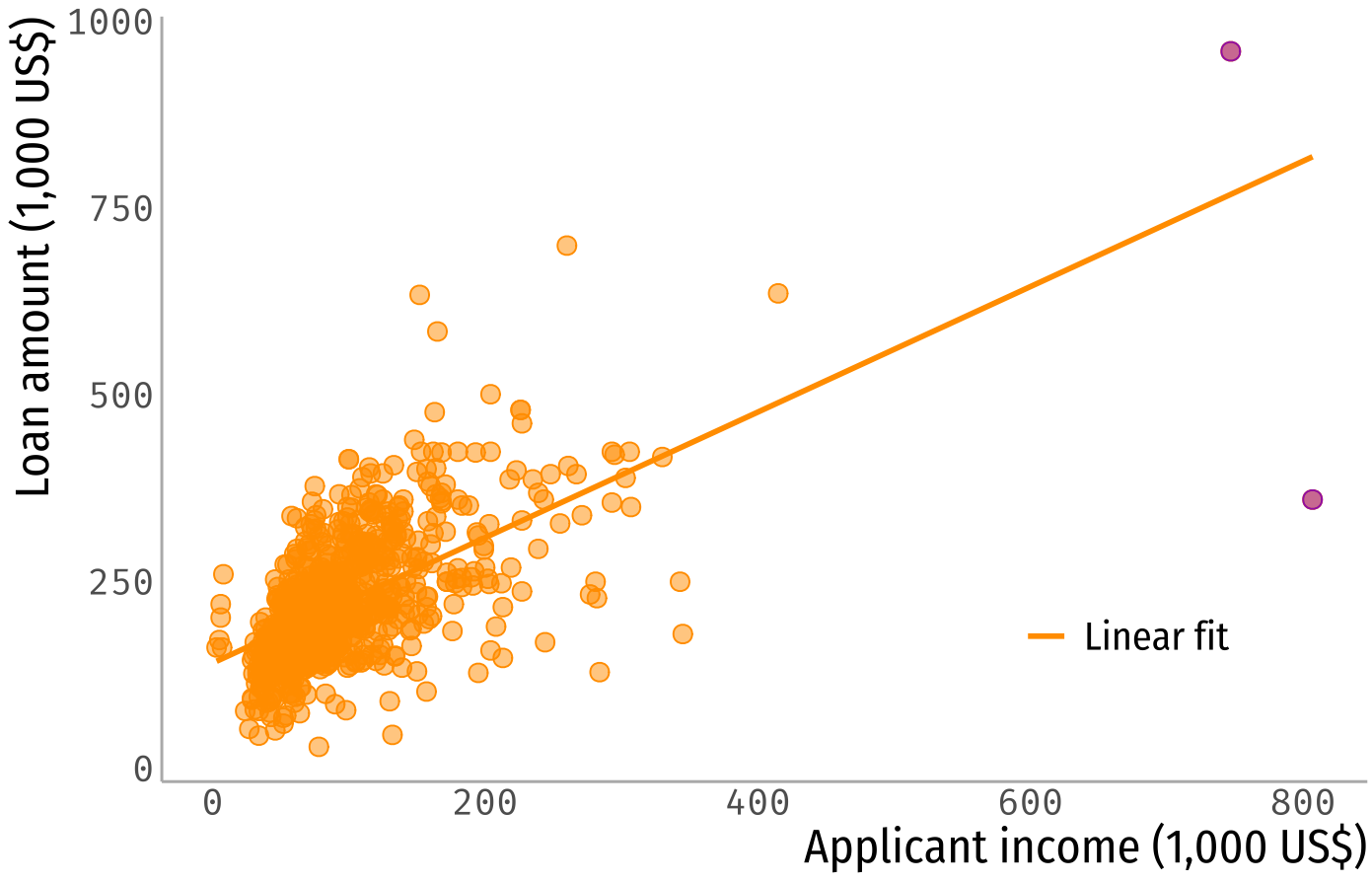
Association between loan amount and income



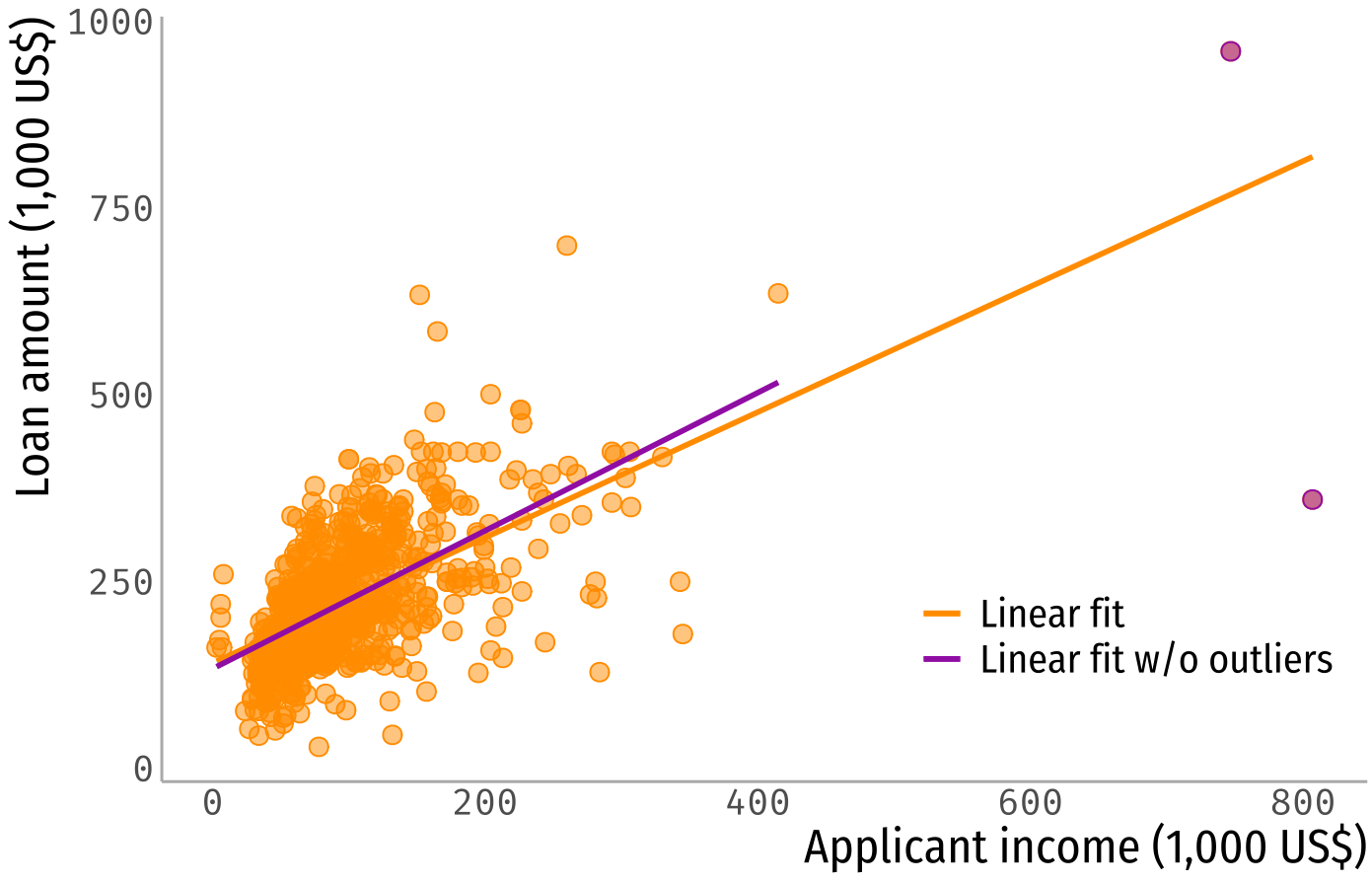
Identifying outliers



Association with complete data



Association after removing outliers



Compare both coefficients: Complete data

```
summary(lm(loan_amount_000s ~ applicant_income_000s, data = hmda))
```

```
##
## Call:
## lm(formula = loan_amount_000s ~ applicant_income_000s, data = hmda)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -458.93  -36.97   -8.77   35.47  365.27
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           141.05028    4.15313   33.96   <2e-16 ***
## applicant_income_000s   0.84000    0.03663   22.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.04 on 875 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3754,    Adjusted R-squared:  0.3747
## F-statistic: 525.8 on 1 and 875 DF,  p-value: < 2.2e-16
```

Compare both coefficients: Data without outliers

```
summary(lm(loan_amount_000s ~ applicant_income_000s, data = hmda_without_outliers))
```

```
##
## Call:
## lm(formula = loan_amount_000s ~ applicant_income_000s, data = hmda_without_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -272.22  -36.09   -6.82   34.12  360.06
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           133.52408    4.47317   29.85   <2e-16 ***
## applicant_income_000s   0.92376    0.04171   22.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.82 on 873 degrees of freedom
## Multiple R-squared:  0.3597,    Adjusted R-squared:  0.359
## F-statistic: 490.5 on 1 and 873 DF,  p-value: < 2.2e-16
```

What to do with outliers?

1. Check them!
    - Make sure there's no coding error; try to understand what's happening there.

2a. If they are wrongly coded:

- You can remove them, always adding a note of why you did so
- Be aware of sample selection!

2b. If they are correctly coded:

- Run analysis both with and without outliers (don't just drop them!).
- Robust results: Do not depend exclusively on a few observations.

Let's do some exercises!