Logistic regression assumptions and diagnostics

Class activity, Part I

https://sta214-s23.github.io/class_activities/ca_lecture_12.html

> - Simulate data with a potential outlier
> - Assess the impact on estimated coefficients

Class activity

> How does an outlier influence the fitted regression model?

- extreme outliers in the data can make our $\hat{\beta}$s quite different from βs (bias!)
- outliers have more potential influence in small samples

Cook's distance

how much influence does each observation have on fitted model?

$k+1 = \# \beta s$

$$D_i = \frac{(Y_i - \hat{\pi}_i)^2}{(k+1)\hat{\pi}_i(1-\hat{\pi}_i)} \cdot \frac{h_i}{(1-h_i)^2}$$

(cook's distance ith observation)

$h_i =$ leverage (how unusual is an observation in X direction)

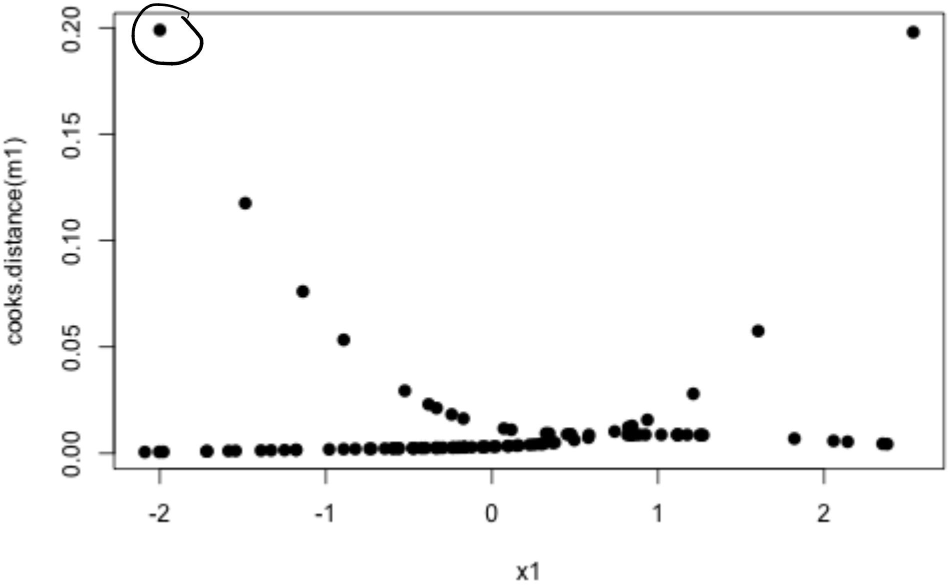
Intuition: a point is influential if both Y_i is far from $\hat{\pi}_i$, and the values of the explanatory variables are unusual

Thresholds: concerned if $D_i > 0.5$ or 1

Cook's distance in R

```r
x1 <- c(x, -2)
y1 <- c(y, 1)
m1 <- glm(y1 ~ x1, family = binomial)

plot(x1, cooks.distance(m1))
```

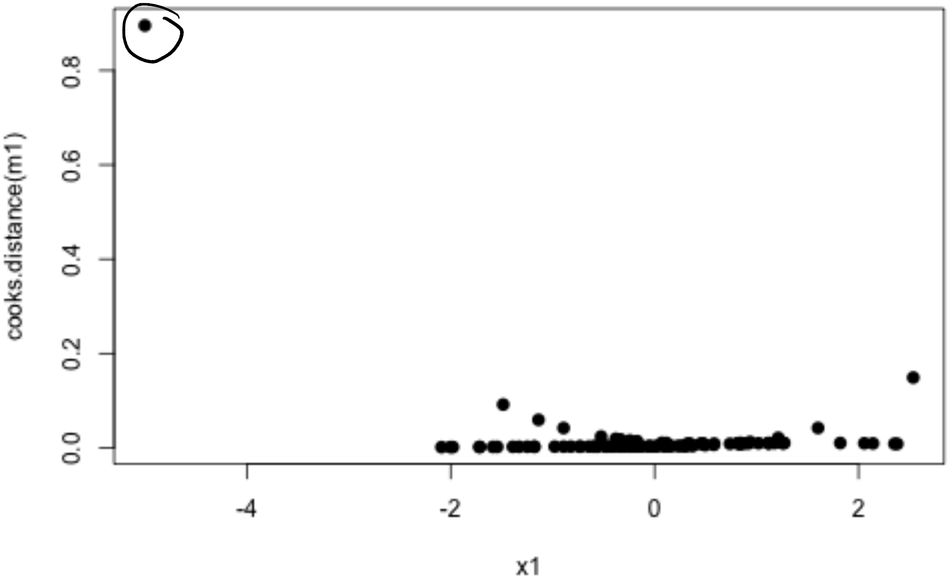


Cook's distance in R

```r
x1 <- c(x, -5)
y1 <- c(y, 1)
m1 <- glm(y1 ~ x1, family = binomial)

plot(x1, cooks.distance(m1))
```

$D_i \approx 0.9$



Addressing model issues

> How should we handle outliers and influential points? Discuss with a neighbor for a few minutes, then we will discuss as a group.

- remove outliers if is a clear error
- remove outliers, report results (p-values, CIs) with and without outliers
- try transformations for skewed explanatory variables

Summary

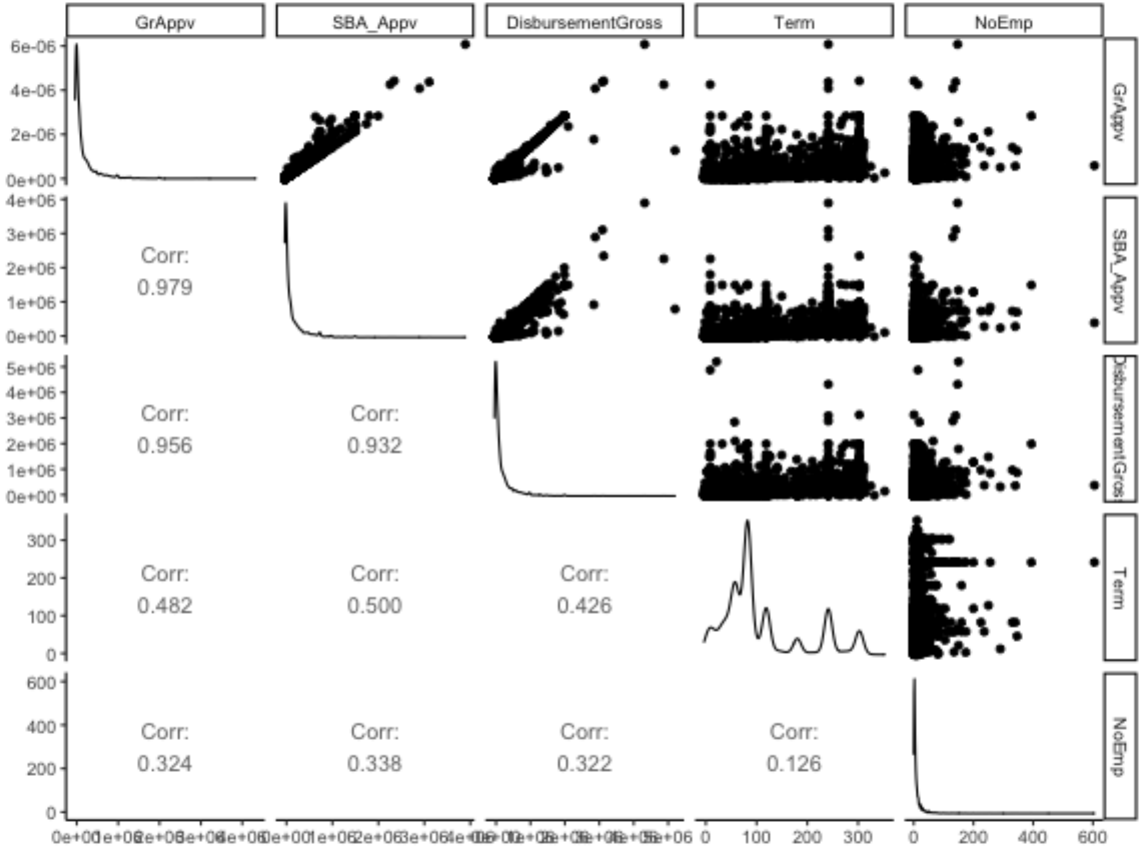
- Shape assumption
  - Diagnostics: empirical logit plots, quantile residual plots
  - Addressing violations: transformations
- Multicollinearity
  - Diagnostics: correlation matrix, scatterplot matrix, VIFs
  - Addressing violations: remove or combine some variables
- Outliers and influential points
  - Diagnostics: Cook's distance
  - Addressing violations: remove clear errors; otherwise report conclusions (p-values, confidence intervals, etc.) with and without potential outliers

Class activity, Part II

https://sta214-s23.github.io/class_activities/ca_lecture_12.html

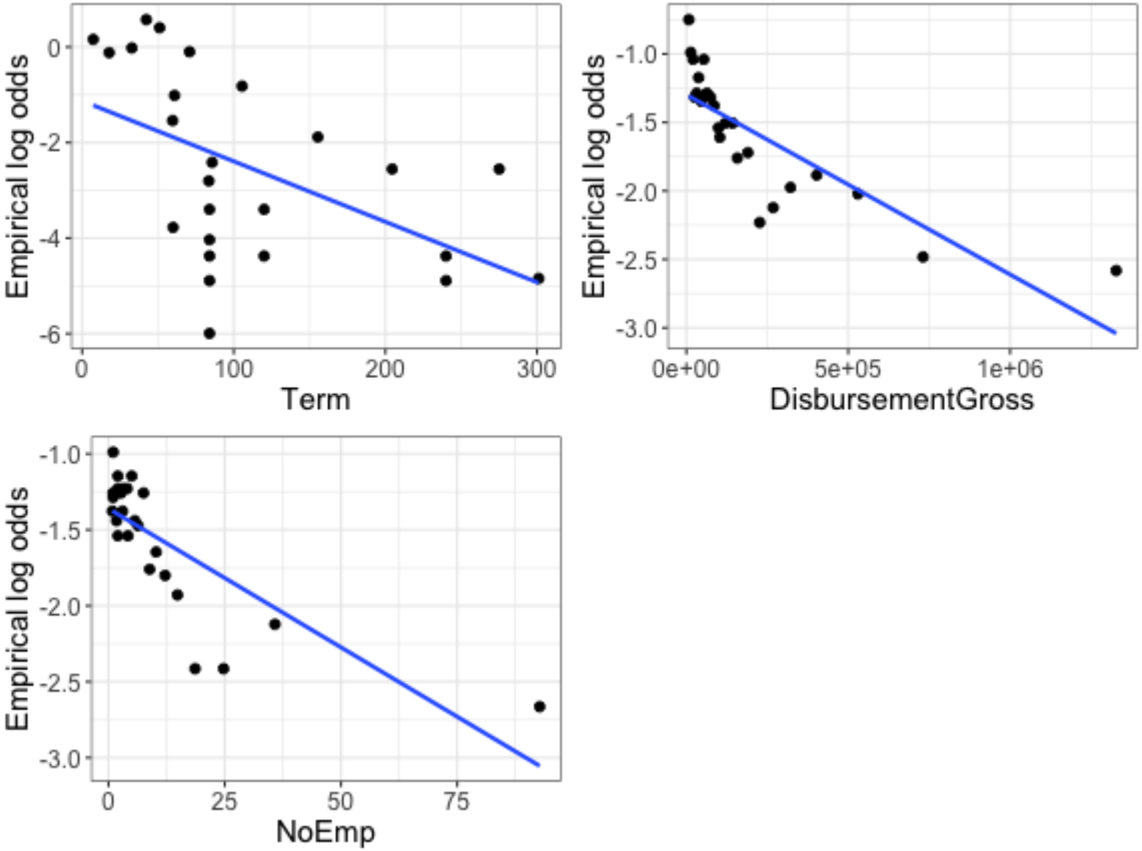
> - Explore a dataset on small business loans
> - Perform diagnostics and build a model
>
> Work with a neighbor on the class activity questions. We will discuss as a group towards the end of the class period. Note: some of the questions are open-ended, with multiple reasonable answers

Correlation



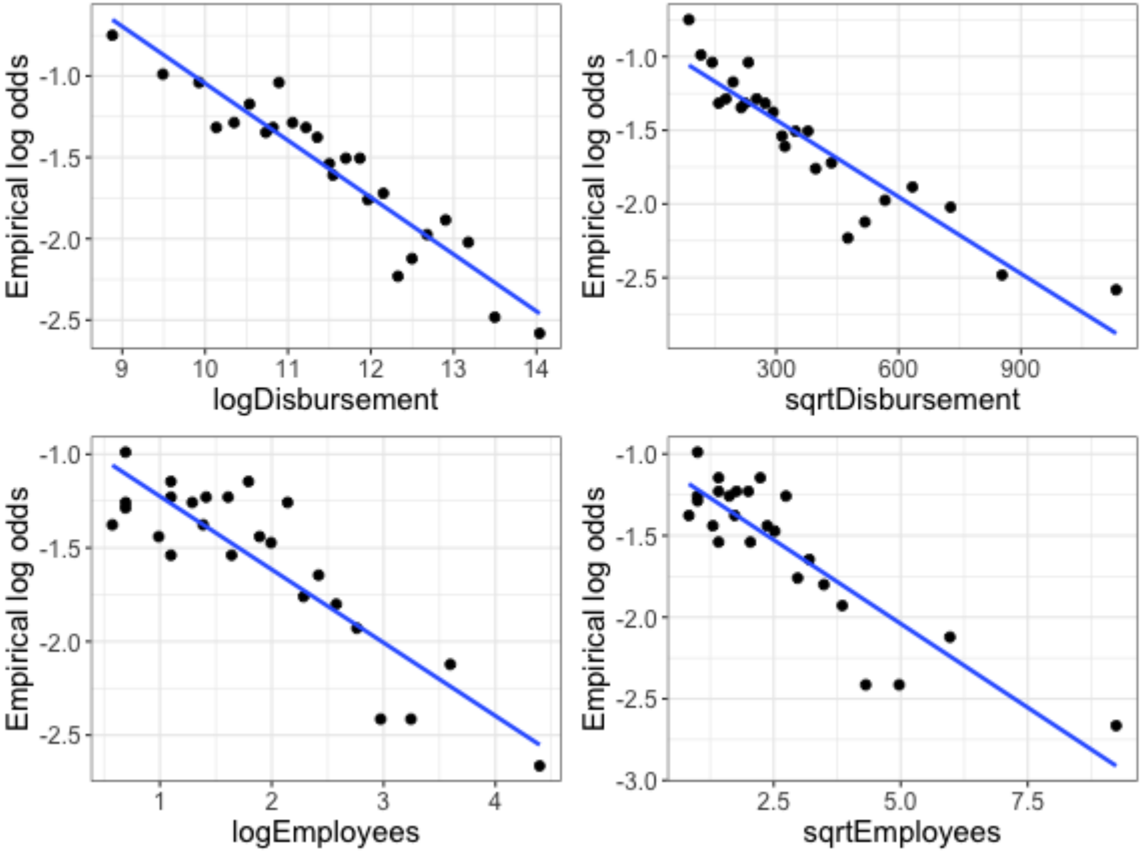
How should we handle correlation in these variables?

Empirical logit plots



> How does the shape assumption look?

Trying some transformations



Model output

New Exist

0 1 2

7

NewExist 1 $\begin{bmatrix}1\\0\\1\\1\end{bmatrix}$ NewExist2 $\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$

```
m1 <- glm(Default ~ log(DisbursementGross) + Term +
        sqrt(NoEmp) + as.factor(NewExist) + as.factor(UrbanRural),
          data = sba, family = binomial)
summary(m1)
```

```
...
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -13.165174 287.140564  -0.046  0.96343
## log(DisbursementGross)  0.100402   0.038974   2.576  0.00999 **
## Term                   -0.021929   0.001192 -18.396  < 2e-16 ***
## sqrt(NoEmp)            -0.101943   0.029505  -3.455  0.00055 ***
## as.factor(NewExist)1   11.656026 287.140216   0.041  0.96762
## as.factor(NewExist)2   11.770036 287.140224   0.041  0.96730
## as.factor(UrbanRural)1  1.145921   0.109647  10.451  < 2e-16 ***
## as.factor(UrbanRural)2  0.870859   0.145871   5.970 2.37e-09 ***
...
```

> Why are the standard errors for NewExist so large?