Scaling by Two Standard Deviations

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Make the most of regression by making coefficients directly interpretable.

Introduction

You all should be familiar with regression by now.

Introduction

Regression coefficients communicate:

- Estimated change in y for one-unit change in x.
    - This is in linear regression.
- Estimated change in *logged odds* of y for one-unit change in x.
    - This is the interpretation for logistic regression.

These communicate some quantities of interest.

- After all, you want to know the effect of x on y!

Introduction

However, it's easy (and tempting) to provide misleading quantities of interest.

- Our variables are seldom (if ever) on the same scale.
    - e.g. age can be anywhere from 18 to 100+, but years of education are typically bound between 0 and 25 (or so).
- Worse yet, zero may not occur in any variable.
    - We would have an uninterpretable *y*-intercept.
- From my experience, this can lead to false convergence of the model itself.

Your goal: regression results should be as easily interpretable as possible.

Interpreting by Standardizing the Input

Gelman (2008) offers a technique for interpreting regression results: scale the non-binary input data by two standard deviations.

- This makes continuous inputs (roughly) on same scale as binary inputs.
- It allows a preliminary evaluation of relative effect of predictors otherwise on different scales.

Standardization

Standardization follows z transformations, which you should know.

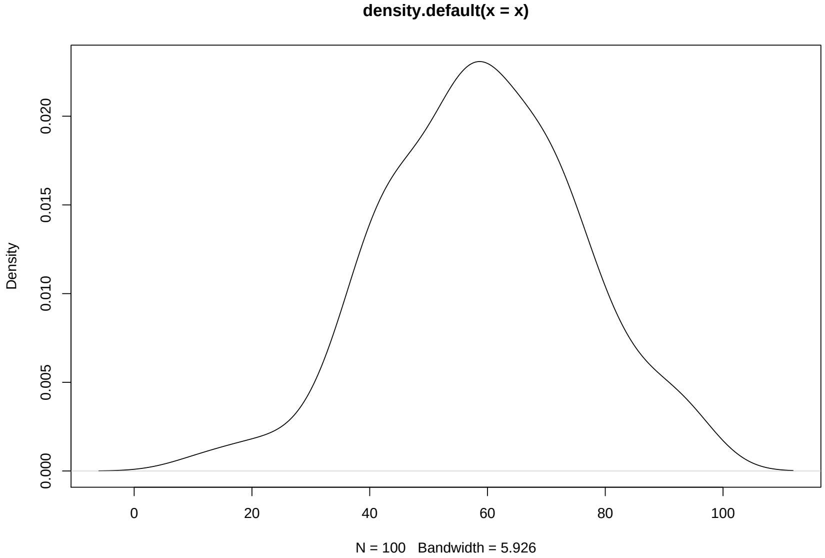
- $z = (\bar{x} - \mu)/\sigma$

This transforms any variable to have a mean of zero and a standard deviation of one.

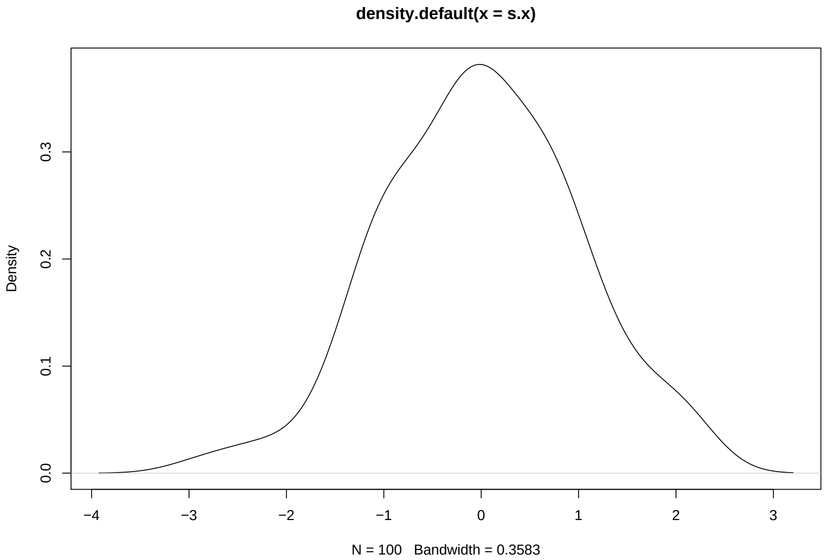
Observe for Normally Distributed Data

```
> set.seed(8675309) ### for reproducibility
> x <- rnorm(100, 58, 17.8)
> mean(x)
[1] 58.93099
> sd(x)
[1] 16.53876
> s.x <- (x - mean(x))/(sd(x))
> mean(s.x)
[1] 1.998618e-16
> sd(s.x)
[1] 1
```

Density Plot



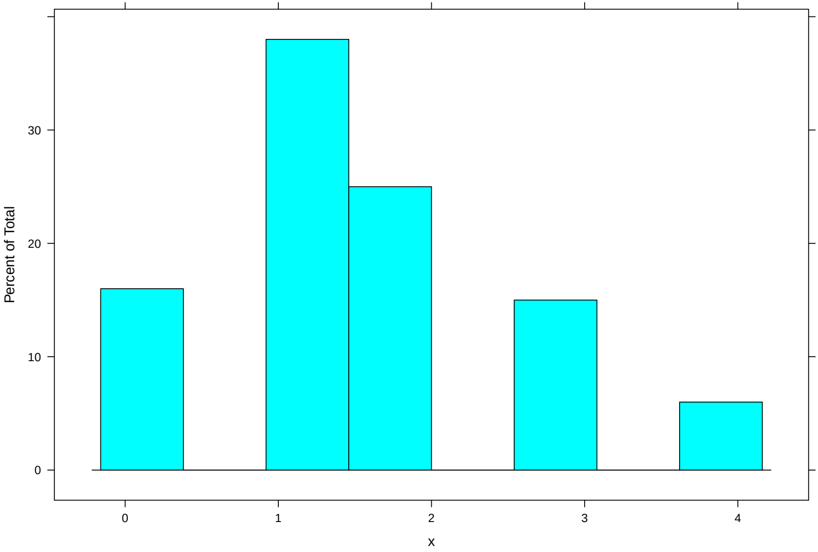
Density Plot (Standardized)



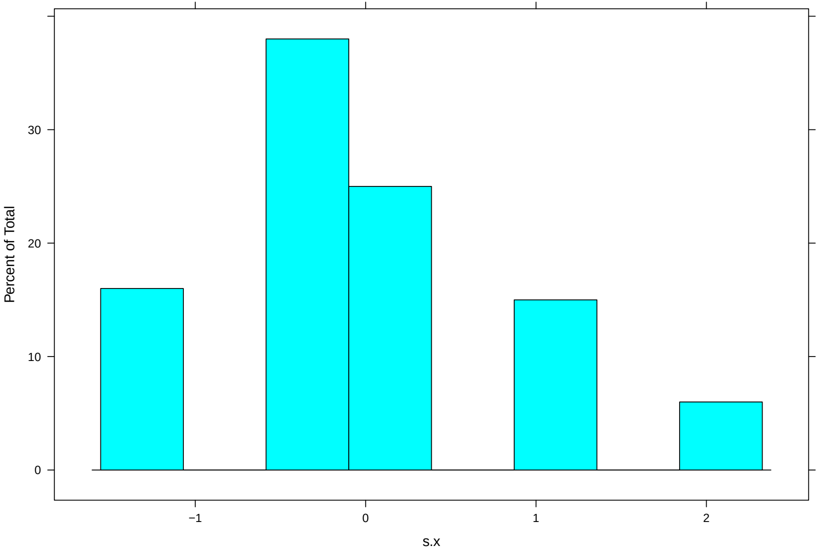
Works with Non-Normally Distributed Data Too

```
> set.seed(8675309)
> x <- rpois(100, 1.5)
> mean(x)
[1] 1.57
> sd(x)
[1] 1.112418
> s.x <- (x - mean(x))/(sd(x))
> mean(s.x)
[1] -7.327933e-17
> sd(s.x)
[1] 1
```

Density Plot



Density Plot (Standardized)



Don't Let the Scale Fool You

```r
unique(x)
```

```
## [1] 0 1 2 4 3
```

```r
unique(s.x)
```

```
## [1] -1.4113394 -0.5123971  0.3865452  2.1844298  1.2854875
```

What Standardization Does to Regression Coefficients

Recall what standardization does when overlaying standardized x-axis with normal x-axis.

- Distance between 0 and $1 = 34\%$ of data.

Thus, a regression coefficient for standardized variable estimates the effect of x on y for a one-standard deviation change from the mean.

- i.e. effect of a change across 34% of the data of x.

Benefits/Limitations of Standardization

Standardizing by one standard deviation is helpful for a couple reasons.

- It creates a meaningful zero (i.e. the mean) for the y-intercept.
- Regression coefficient captures a magnitude change.

However, it won't help us make preliminary comparisons with dummy variables.

The Problem of Dummy Variables

Dummy variables are special class of nominal variables.

- An indicator is either “there” or “not there”.

In regression, this has an important effect.

- Coefficient goes up, all else equal.
    - So does standard error.

It may be misleading to think that binary variables have the largest effect on an outcome, but a regression coefficient may suggest this.

Scaling by Two Standard Deviations

Take a continuous (non-binary) input variable and divide it by two standard deviations instead of one.

- This will transform the data to have a mean of zero and standard deviation of .5.
- Regression coefficient would communicate estimated change in y for change across 47.7% of data in x.

Observe with Normally Distributed Data

```
> set.seed(8675309)
> x <- rnorm(100, 58, 17.8)
> mean(x)
[1] 58.93099
> sd(x)
[1] 16.53876
> s.x <- (x - mean(x))/(2*sd(x)) ### two SDs instead of one
> mean(s.x)
[1] 9.993091e-17
> sd(s.x)
[1] 0.5
```

Same Thing with Non-Normally Distributed Data

```
> set.seed(8675309)
> x <- rpois(100, 1.5)
> mean(x)
[1] 1.57
> sd(x)
[1] 1.112418
> s.x <- (x - mean(x))/(2*sd(x)) ### two SDs instead of one
> mean(s.x)
[1] -3.663966e-17
> sd(s.x)
[1] 0.5
```

Comparison with Binary Independent Variables

Why do this? Consider a dummy IV with 50/50 split between 0s and 1s.

- $p(dummy = 1) = .5$
- Then, standard deviation equals .5 ($\sqrt{.5 * .5} = \sqrt{.25} = .5$)
- We can directly compare this dummy variable with our new standardized input variable!

This works well in most cases, except when $p(dummy = 1)$ is really small.

- e.g. $p(dummy = 1) = .25$, then $\sqrt{.25 * .75} = .4330127$

An Application with State Education-Turnout

I revisit the state education-turnout example from the Pollock book.

I do use newer data.

- **state education**: % of state having HS diploma (2009).
- **turnout**: state-level turnout of VEP in 2012 general election.
- **region**: factor/"fixed effects"
  - 0 = West, 1 = Northeast, 2 = Midwest, 3 = South.

A Simple Regression

```
> M1 <- lm(turnout ~ perhsdiploma +
           factor(regioncondensed), data=Data)
> summary(M1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -33.7839    27.3483  -1.235  0.22312
perhsdiploma   1.0464     0.3105   3.370  0.00155 **
Northeast      3.4049     2.4388   1.396  0.16952
Midwest        4.1645     2.2680   1.836  0.07294 .
South          3.7147     2.4732   1.502  0.14009
```

Some Confusion (with the Results)

- We see that more educated states have higher turnout.
- Midwestern states have higher turnout in comparison to states in West.

However, are we to believe that the Midwest is the largest predictor?

- and what about that uninterpretable y-intercept?

Let's standardize the education variable by two standard deviations.

A More Readable Regression

```
> M2 <- lm(turnout ~ z.perhsdiploma +
    factor(regioncondensed), data=Data)
    > summary(M2)

    Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      57.133      1.592  35.881  < 2e-16 ***
z.perhsdiploma    7.128      2.115   3.370  0.00155 **
Northeast         3.405      2.439   1.396  0.16952
Midwest           4.164      2.268   1.836  0.07294 .
South             3.715      2.473   1.502  0.14009
```

Interpretation

Notice that the effects ultimately didn't change for the region fixed effects.

- t value for education variable is unchanged too.

However, this regression table is much more readable.

- y-intercept is much more meaningful.
- We also see that education does appear to have the largest effect.

Table of Contents