PLSC 502 – Fall 2024
Linear Regression II

December 2, 2024

Model Fit

Model Fit

Model fit is:

- The closeness of the mapping between model-based values of Y and actual values of Y...
- Can be *in-sample* or *out-of-sample* ($\rightarrow$ "overfitting")
- Is (in part) a function of *model specification* (choice of predictors, functional form, interactions, etc.)
- Related (but not identical) to prediction / predictive ability

R^2 Introduced

Recall that for

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

We have:

$$\begin{aligned}
\text{“TSS”} &= \sum (Y_i - \bar{Y})^2 \\
\text{“MSS”} &= \sum (\hat{Y}_i - \bar{Y})^2 \\
\text{“RSS”} &= \sum (Y_i - \hat{Y}_i)^2 \equiv \sum \hat{u}_i^2
\end{aligned}$$

Then:

$$\begin{aligned}
R^2 &= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \\
&= \frac{\text{MSS}}{\text{TSS}} \\
&= 1 - \frac{\text{RSS}}{\text{TSS}} \\
&= 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}
\end{aligned}$$

R^2 Is...

R-squared:

- is "the proportion of variance explained"
- $\in [0,1]$
  - $R^2 = 1.0 \equiv$ a "perfect (linear) fit"'
  - $R^2 = 0 \equiv$ no (linear) $X - Y$ association

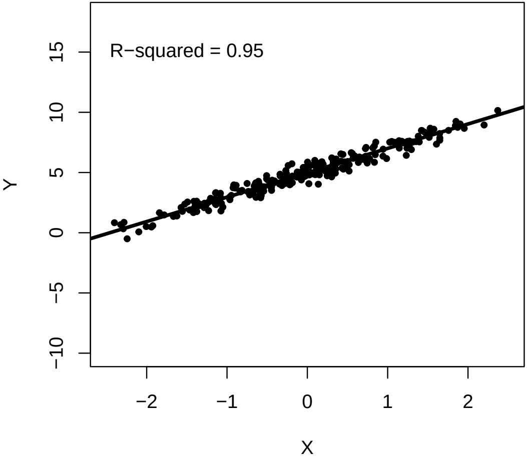
For a single X,

$$\begin{aligned} R^2 &= \hat{\beta}_1^2 \frac{\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} \\ &= (r_{XY})^2 \end{aligned}$$

A (Simulated) Example

```
seed <- 7222009
set.seed(seed)
> X<-rnorm(250)
> Y1<-5+2*X+rnorm(250,mean=0,sd=sqrt(0.2))
> Y2<-5+2*X+rnorm(250,mean=0,sd=sqrt(20))
> fit<-lm(Y1~X)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.97712    0.02846  174.86   <2e-16 ***
X            2.02529    0.02785   72.73   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4491 on 248 degrees of freedom
Multiple R-squared:  0.9552, Adjusted R-squared:  0.955
F-statistic:  5290 on 1 and 248 DF,  p-value: < 2.2e-16
```

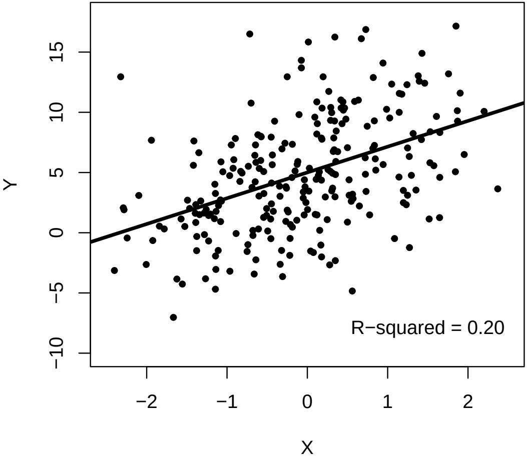
Regression of $Y_i = 5 + 2X_i + u_i$ $(R^2 = 0.95)$



Same Slope/Intercept, Different R^2

```
> fit2<-lm(Y2~X)
> summary(fit2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.0048     0.2757  18.151  < 2e-16 ***
X             2.1402     0.2697   7.934 7.29e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.351 on 248 degrees of freedom
Multiple R-squared:  0.2024, Adjusted R-squared:  0.1992
F-statistic: 62.95 on 1 and 248 DF,  p-value: 7.288e-14
```

Regression of $Y_i = 5 + 2X_i + u_i$ ($R^2 = 0.20$)



R^2 is Also an *Estimate*...

Luskin: Population analogue "P^2":

$$P^2 = 1 - \frac{\sigma^2}{\sigma_Y^2}$$

Then $\hat{P}^2 = R^2$ has variance:

$$\widehat{\text{Var}(R^2)} = \frac{4R^2(1-R^2)^2(N-k)^2}{(N^2-1)(N+3)}$$

and standard error:

$$\widehat{\text{s.e.}(R^2)} = \sqrt{\frac{4R^2(1-R^2)^2(N-k)^2}{(N^2-1)(N+3)}}.$$

Adjusted R^2

"Adjusted" R^2 is:

$$R^2_{adj.} = 1 - \frac{(1 - R^2)(N - c)}{(N - k)}$$

where $c = 1$ if there is a constant in the model and $c = 0$ otherwise.

$R^2_{adj.}$ characteristics:

- $R^2_{adj.} \rightarrow R^2$ as $N \rightarrow \infty$
- $R^2_{adj.}$ can be > 1, or < 0...
- $R^2_{adj.}$ increases with model "fit," but
- The extent of that increase is discounted by a factor proportional to the number of covariates.

Other R^2 / Goodness-Of-Fit Alternatives

- Standard Error of the Estimate:

$$\text{SEE} = \sqrt{\frac{\text{RSS}}{N-k}}$$

- F-statistic (bivariate regression, for $\beta_1 = 0$):

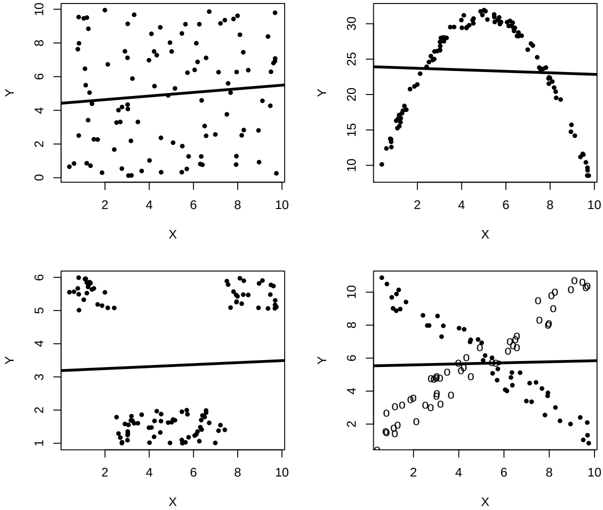
$$\begin{aligned} F &= \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{(N-1)-(N-2)} \div \frac{\sum(Y_i - \hat{Y}_i)^2}{(N-2)} \\ &= \frac{\text{“explained” variance}}{\text{“unexplained” variance}} \end{aligned}$$

  which is $\sim F(1, N-2)$.

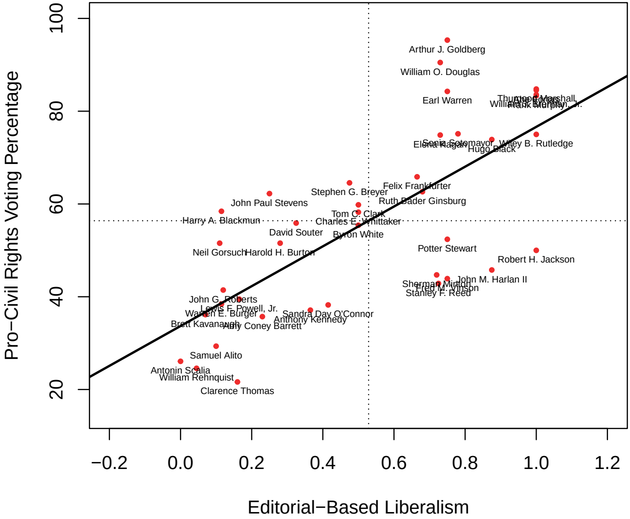
- ROC / AUC (later...)

- Graphical methods

Caution: Different Ways to get $R^2 \approx 0$



Remember This Regression?



Remember This Regression?

```
> fit<-lm(CivLibs˜IdeologyScore,data=SCOTUS)
> summary(fit)

Call:
lm(formula = CivLibs ˜ IdeologyScore, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-26.62  -9.84   2.61   8.05  29.44

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)
(Intercept)      33.69       4.26    7.91 0.0000000018 ***
IdeologyScore    42.94       6.85    6.27 0.0000002699 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

ANOVA → R^2

```
> anova(fit)
Analysis of Variance Table

Response: CivLibs
              Df Sum Sq Mean Sq F value     Pr(>F)
IdeologyScore  1   7753    7753    39.3 0.00000027 ***
Residuals     37   7294     197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # R-squared:
>
> anova(fit)$‘Sum Sq‘[1] / (anova(fit)$‘Sum Sq‘[1] + anova(fit)$‘Sum Sq‘[2])
[1] 0.515

> # F-statistic:
>
> anova(fit)$‘Mean Sq‘[1] / anova(fit)$‘Mean Sq‘[2]
[1] 39.3
```

How Much Improvement?

Consider:

$$Y_i = \beta_0 + u_i$$

...which gives:

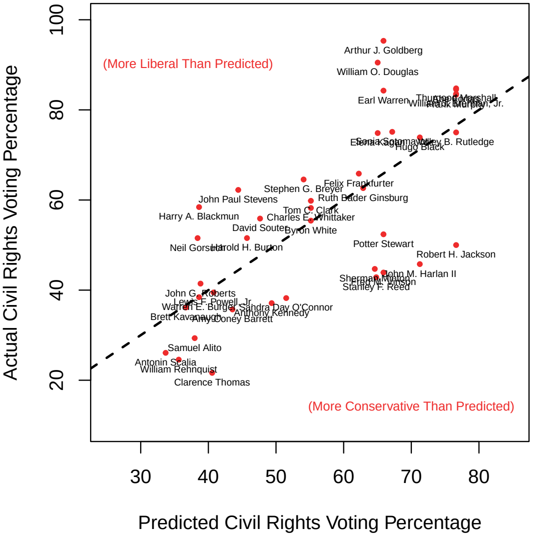
```
> fit0<-lm(CivLibs~1,data=SCOTUS)
> summary(fit0)

Call:
lm(formula = CivLibs ~ 1, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max 
-34.76 -15.93  -0.97  17.98  38.94 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    56.39       3.19    17.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 19.9 on 38 degrees of freedom
```

Model Fit / Improvement

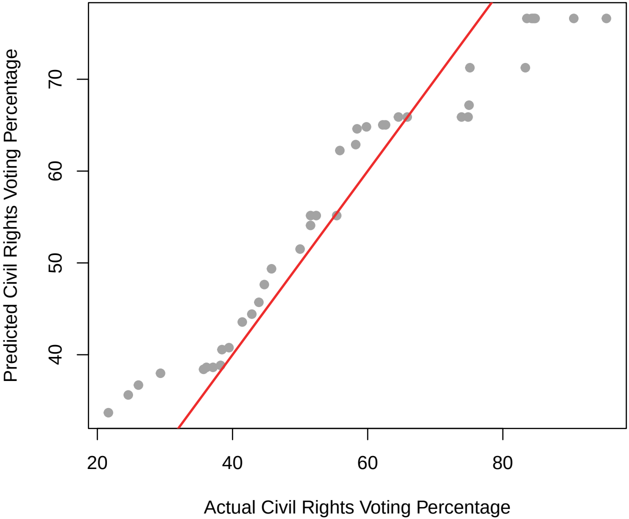
Comparison:

		Model without	Model with
Statistic	Concept	Segal-Cover	Segal-Cover
RSE	"Typical" residual	19.9	14
R-squared	"Variance explained"	0 (N/A)	0.515
F-statistic	$P($"better than chance"$)$	0 (N/A)	39.3

Model Fit: Plotting $\hat{Y}$ vs. Y



Model Fit: Q-Q Plot (Actual vs. Predicted)

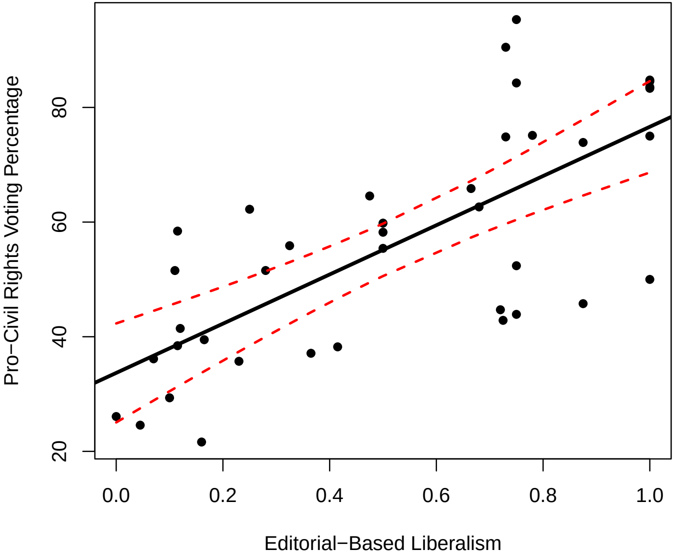


Stupid Regression Tricks

SCOTUS Regression Redux

```
> fit<-lm(CivLibs~IdeologyScore,data=SCOTUS)
> summary(fit)

Call:
lm(formula = CivLibs ~ IdeologyScore, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-26.62  -9.84   2.61   8.05  29.44

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)
(Intercept)      33.69       4.26    7.91 0.0000000018 ***
IdeologyScore    42.94       6.85    6.27 0.0000002699 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

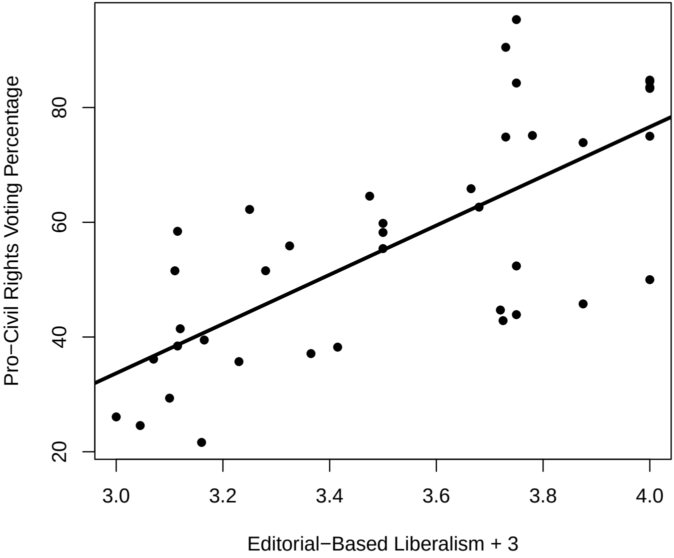
SCOTUS Regression Plot



Add Three to IdeologyScore

```
> SCOTUS$IdeoPlus3 <- SCOTUS$IdeologyScore + 3
>
> fit2<-lm(CivLibs~IdeoPlus3,data=SCOTUS)
> summary(fit2)

Call:
lm(formula = CivLibs ~ IdeoPlus3, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-26.62  -9.84   2.61   8.05  29.44

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)   -95.12      24.26   -3.92    0.00037 ***
IdeoPlus3      42.94       6.85    6.27 0.00000027 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

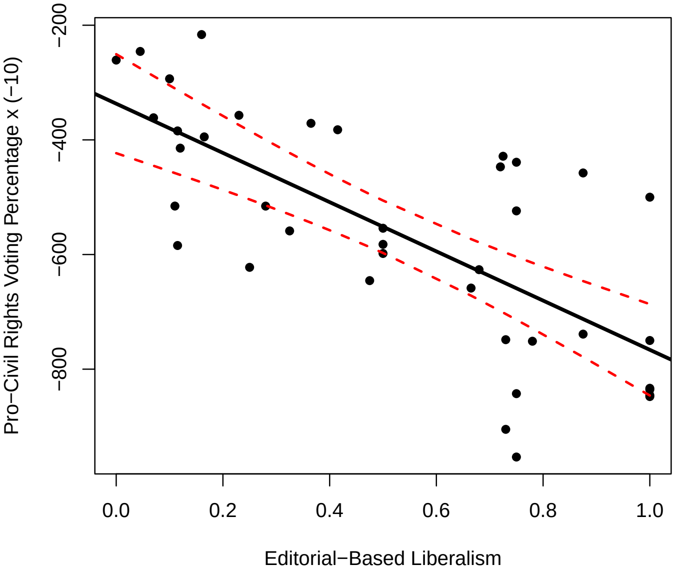
SCOTUS Plot With Rescaled X



Multiply CivLibs Times -10

```
> SCOTUS$CivLibNeg10 <- -10 * SCOTUS$CivLibs
>
> fit3<-lm(CivLibNeg10~IdeologyScore,data=SCOTUS)
> summary(fit3)

Call:
lm(formula = CivLibNeg10 ~ IdeologyScore, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-294.4  -80.5  -26.1   98.4  266.2

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)
(Intercept)     -336.9       42.6   -7.91 0.0000000018 ***
IdeologyScore   -429.4       68.5   -6.27 0.0000002699 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

SCOTUSplot With Rescaled Y



Linear Transformations

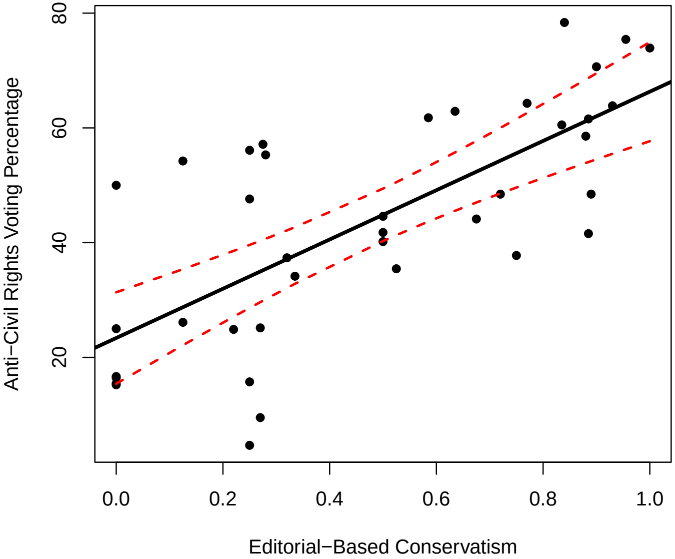
- Adding (subtracting) a positive constant to X shifts the X-axis to the left (right).
- Adding (subtracting) a positive constant to Y shifts the Y-axis downwards (upwards).
- Multiplying X (Y) times a positive constant greater than 1.0 stretches the X (Y) axis.
- Multiplying X (Y) times a positive constant less than 1.0 shrinks the X (Y) axis.
- Multiplying X (Y) times a negative constant inverts the X (Y) axis, and stretches / shrinks it as above.

Linear transformations do not alter the model in a statistically / substantively important way.

Application: Reversing The Scales

```
> SCOTUS$CivLibCons <- 100 - SCOTUS$CivLibs
> SCOTUS$IdeolCons <- 1 - SCOTUS$IdeologyScore
>
> fit4<-lm(CivLibCons~IdeolCons,data=SCOTUS)
> summary(fit4)

Call:
lm(formula = CivLibCons ~ IdeolCons, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-29.44  -8.05  -2.61   9.84  26.62

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    23.38       3.93    5.94 0.00000075 ***
IdeolCons      42.94       6.85    6.27 0.00000027 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

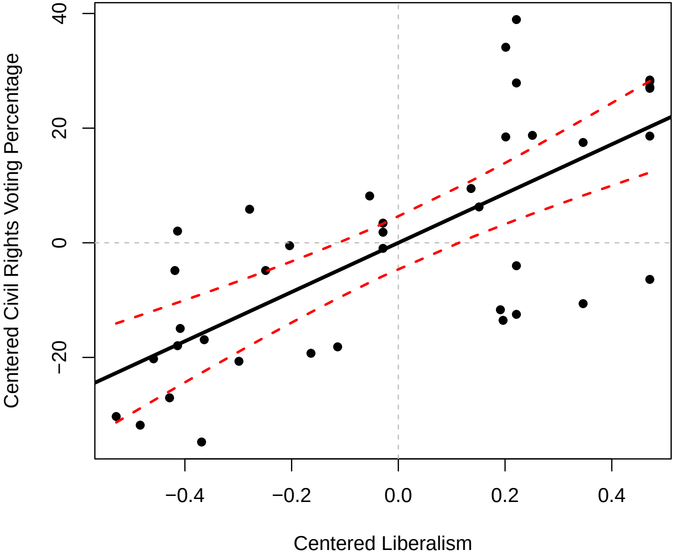
Plot of Civil Liberties Conservatism vs. Ideological Conservatism



Application: “Centering” Variables

```
> SCOTUS$CivLibCentered <- SCOTUS$CivLibs - mean(SCOTUS$CivLibs)
> SCOTUS$IdeolCentered <- SCOTUS$IdeologyScore - mean(SCOTUS$IdeologyScore)
>
> fit5<-lm(CivLibCentered~IdeolCentered,data=SCOTUS)
> summary(fit5)

Call:
lm(formula = CivLibCentered ~ IdeolCentered, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-26.62  -9.84   2.61   8.05  29.44

Coefficients:
              Estimate Std. Error t value   Pr(>|t|)
(Intercept)  -2.15e-15   2.25e+00    0.00          1
IdeolCentered  4.29e+01   6.85e+00    6.27 0.00000027 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

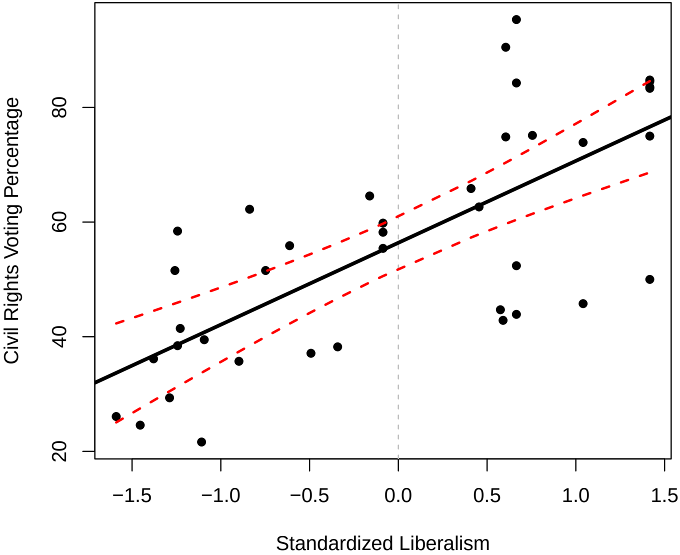
“Regression Through The Origin”



Application: “Standardizing” a Variable

```
> SCOTUS$IdeolStd <- scale(SCOTUS$IdeologyScore)
>
> fit6<-lm(CivLibs~IdeolStd,data=SCOTUS)
> summary(fit6)

Call:
lm(formula = CivLibs ~ IdeolStd, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-26.62  -9.84   2.61   8.05  29.44

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    56.39       2.25   25.08    < 2e-16 ***
IdeolStd       14.28       2.28    6.27 0.00000027 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

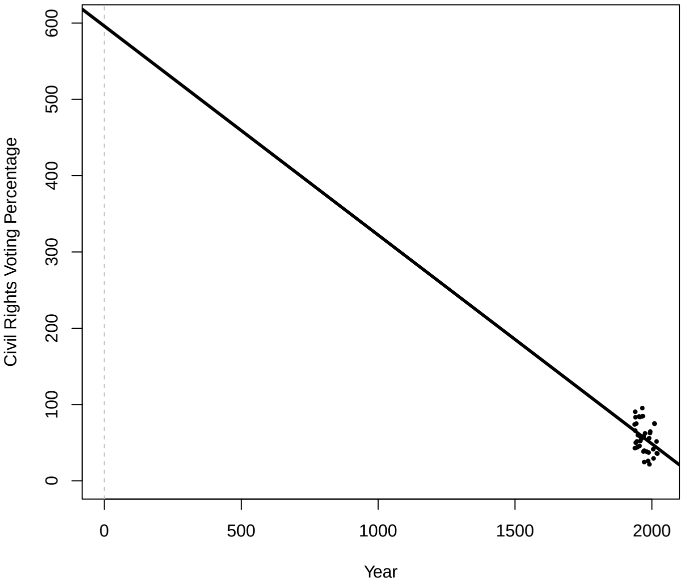
OLS with Standardized X



Rescaling for Interpretability

```
> fit7<-lm(CivLibs~Year,data=SCOTUS)
> summary(fit7)

Call:
lm(formula = CivLibs ~ Year, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max 
 -31.6  -15.2   -2.6   13.4   37.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  596.059    235.072    2.54    0.016 *
Year          -0.274      0.119   -2.30    0.027 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.9 on 37 degrees of freedom
Multiple R-squared:  0.125, Adjusted R-squared:  0.101 
F-statistic: 5.27 on 1 and 37 DF,  p-value: 0.0274
```

What Does That Look Like?



Rescaling for Interpretability (continued)

```
> SCOTUS$Year1950<-SCOTUS$Year-1950
> fit8<-lm(CivLibs~Year1950,data=SCOTUS)
> summary(fit8)

Call:
lm(formula = CivLibs ~ Year1950, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max 
 -31.6  -15.2   -2.6   13.4   37.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   62.145      3.926    15.8   <2e-16 ***
Year1950      -0.274      0.119    -2.3    0.027 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.9 on 37 degrees of freedom
Multiple R-squared:  0.125, Adjusted R-squared:  0.101 
F-statistic: 5.27 on 1 and 37 DF,  p-value: 0.0274
```

Reporting

The results:

```
> summary(fit)

Call:
lm(formula = CivLibs ~ IdeologyScore, data = SCOTUS)

Residuals:
   Min     1Q Median     3Q    Max
-26.62  -9.84   2.61   8.05  29.44

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)
(Intercept)      33.69       4.26    7.91 0.0000000018 ***
IdeologyScore    42.94       6.85    6.27 0.0000002699 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 37 degrees of freedom
Multiple R-squared:  0.515, Adjusted R-squared:  0.502
F-statistic: 39.3 on 1 and 37 DF,  p-value: 0.00000027
```

Reporting

The table:

Table: OLS Regression Model of SCOTUS Voting

Variables	Model I
(Constant)	33.69
	(4.26)
Ideological Liberalism	42.94*
	(6.85)
Adjusted R^2	0.50

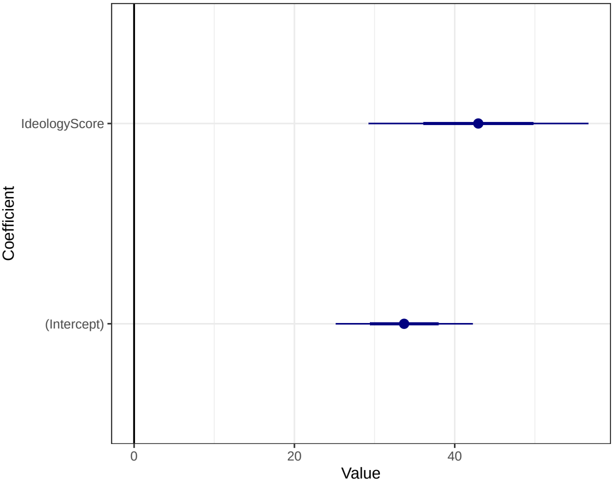
Note: $N = 39$. *Cell entries are coefficient estimates; numbers in parentheses are estimated standard errors. Asterisks indicate* $p < .05$ *(one-tailed). See text for details.*

Another Table (using defaults in `stargazer`)

Table: OLS Regression Model of SCOTUS Voting

	Model I
(Constant)	33.70^{***}
	(4.26)
Ideological Liberalism	42.90^{***}
	(6.85)
Observations	39
R^2	0.52
Adjusted R^2	0.50
Residual Std. Error	14.00 (df = 37)
F Statistic	39.30^{***} (df = 1; 37)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Default-y Ladderplot, using `-fitplot-`



Tools for Tables (→ Figures)

Table tools (in no particular order):

- `stargazer`
- `tinytable`
- `texreg`
- `gt`
- `reactable` (interactive tables)

Figures from regression results:

- `coefplot`
- `jtools`
- `modelsummary`
- `dotwhisker`

See more resources at the Reproducible Research task view.

Some General Guidelines ("Rules"?)

Tables:

- *Use column headings descriptively.*
- *Use multiple rows / columns rather than multiple tables.*
- *Learn about significant digits, and don't report more than 3-4 of them.*
- *Use a figure to replace a table when you can.*
- *Be aware of norms about *s.*

Figures:

- *Report the scale of axes, and label them.*
- *Use as much "space" as you need, but no more.*
- *Use color sparingly.*