

PLSC 502 – Fall 2022

Linear Regression II

December 1, 2022

- 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

Recall that for

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

We have:

$$\text{"TSS"} = \sum (Y_i - \bar{Y})^2$$

$$\text{"MSS"} = \sum (\hat{Y}_i - \bar{Y})^2$$

$$\text{"RSS"} = \sum (Y_i - \hat{Y}_i)^2 \equiv \sum \hat{u}_i^2$$

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-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 ***

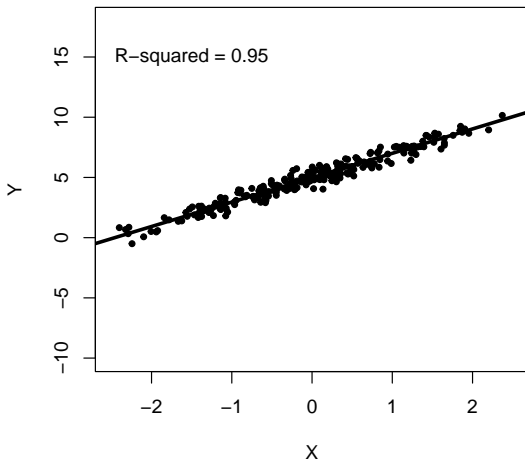
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 ***

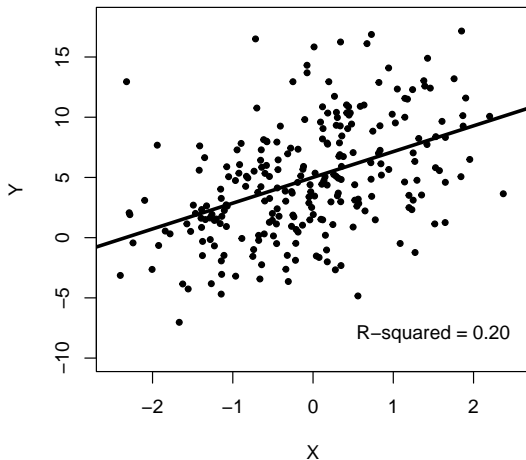
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 is:

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

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

$R_{adj.}^2$ characteristics:

- $R_{adj.}^2 \rightarrow R^2$ as $N \rightarrow \infty$
- $R_{adj.}^2$ can be > 1 , or < 0 ...
- $R_{adj.}^2$ 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:

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

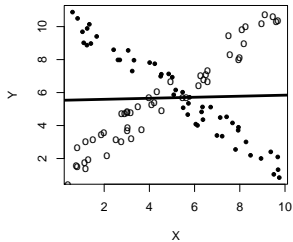
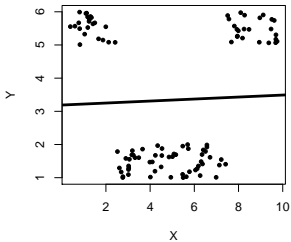
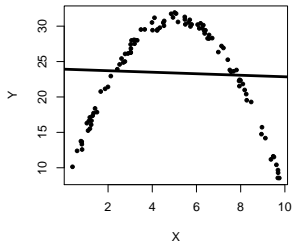
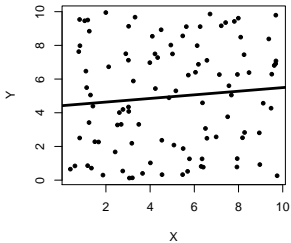
- F -statistic (bivariate regression, for $\beta_1 = 0$):

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

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

- ROC / AUC (later...)
- Graphical methods

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



SCOTUS Redux (OT1946-2021)

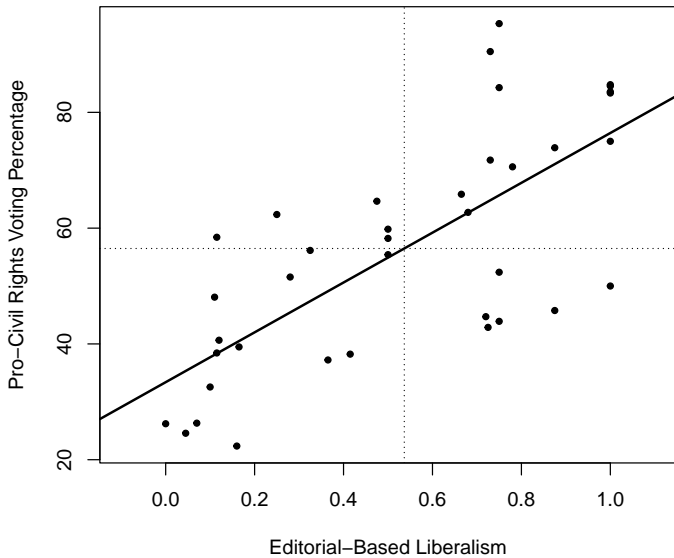
Data from the [Supreme Court Database](#) and the justices' [Segal-Cover](#) scores...

- Y is CivLibs = liberal voting percentage in civil rights & liberties cases
- X is IdeologyScore $\in [0, 1] \rightarrow$ SCOTUS justice liberalism

```
> describe(SCOTUS,skew=FALSE,trim=0)
```

	vars	n	mean	sd	min	max	range	se
justice	1	38	97.37	11.32	78.00	116.00	38.00	1.84
justiceName*	2	38	19.50	11.11	1.00	38.00	37.00	1.80
CivLibs	3	38	56.49	19.94	22.36	95.33	72.97	3.23
Nom.Order*	4	38	19.50	11.11	1.00	38.00	37.00	1.80
Nominee*	5	38	19.50	11.11	1.00	38.00	37.00	1.80
ChiefJustice*	6	4	1.00	0.00	1.00	1.00	0.00	0.00
SenateVote*	7	38	17.05	8.23	1.00	25.00	24.00	1.33
IdeologyScore	8	38	0.54	0.33	0.00	1.00	1.00	0.05
QualificationsScore*	9	38	16.45	7.91	1.00	25.00	24.00	1.28
Nominator (Party)*	10	38	7.03	3.72	1.00	13.00	12.00	0.60
Year	11	38	1969.74	24.70	1937.00	2018.00	81.00	4.01

Scatterplot



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

Call:

```
lm(formula = CivLibs ~ IdeologyScore, data = SCOTUS)
```

Residuals:

Min	1Q	Median	3Q	Max
-26.433	-10.587	2.460	7.858	29.655

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	33.389	4.354	7.669	4.44e-09	***
IdeologyScore	43.044	6.917	6.223	3.51e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.03 on 36 degrees of freedom

Multiple R-squared: 0.5182, Adjusted R-squared: 0.5048

F-statistic: 38.72 on 1 and 36 DF, p-value: 3.505e-07

```

> anova(fit)
Analysis of Variance Table

Response: CivLibs
          Df Sum Sq Mean Sq F value    Pr(>F)
IdeologyScore  1   7621    7621   38.7 0.00000035 ***
Residuals    36   7086     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)$'Sum Sq'[2])
[1] 0.5182
>
> # F-statistic:
>
> anova(fit)$'Mean Sq'[1] / anova(fit)$'Mean Sq'[2]
[1] 38.72

```


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
-26.433	-10.587	2.460	7.858	29.655

Coefficients:

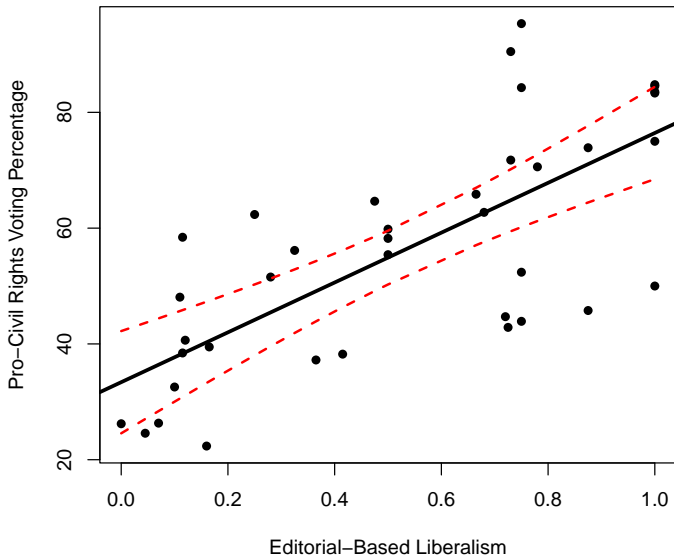
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.389	4.354	7.669	4.44e-09 ***
IdeologyScore	43.044	6.917	6.223	3.51e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.03 on 36 degrees of freedom

Multiple R-squared: 0.5182, Adjusted R-squared: 0.5048

F-statistic: 38.72 on 1 and 36 DF, p-value: 3.505e-07



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
-26.43	-10.59	2.46	7.86	29.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-95.74	24.57	-3.90	0.00041	***
IdeoPlus3	43.04	6.92	6.22	0.00000035	***

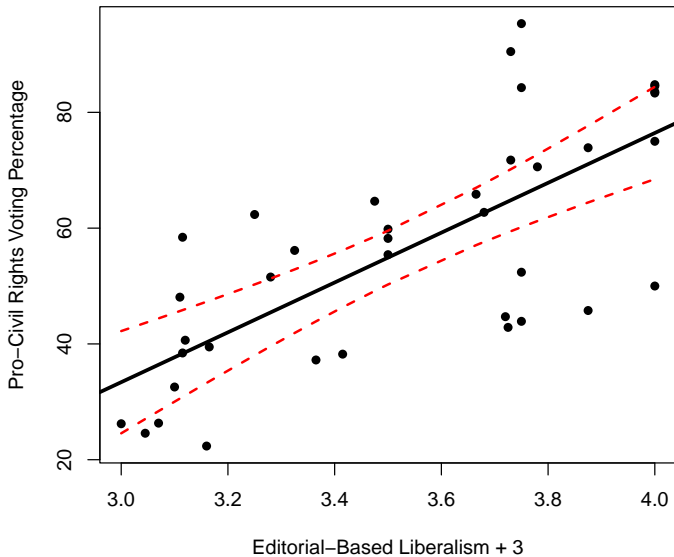
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14 on 36 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.505

F-statistic: 38.7 on 1 and 36 DF, p-value: 0.000000351

SCOTUSplot 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
-296.6	-78.6	-24.6	105.9	264.3

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-333.9	43.5	-7.67	4.4e-09	***
IdeologyScore	-430.4	69.2	-6.22	3.5e-07	***

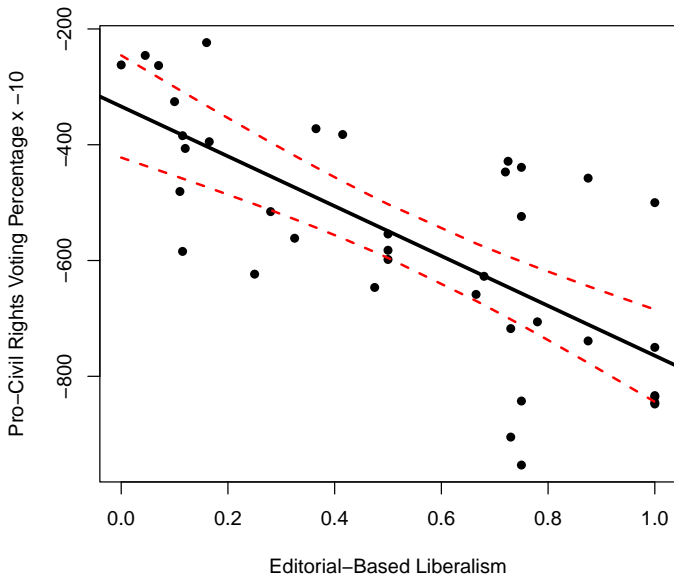
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140 on 36 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.505

F-statistic: 38.7 on 1 and 36 DF, p-value: 3.51e-07

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
-29.66	-7.86	-2.46	10.59	26.43

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	23.57	3.93	5.99	7.1e-07 ***
IdeolCons	43.04	6.92	6.22	3.5e-07 ***

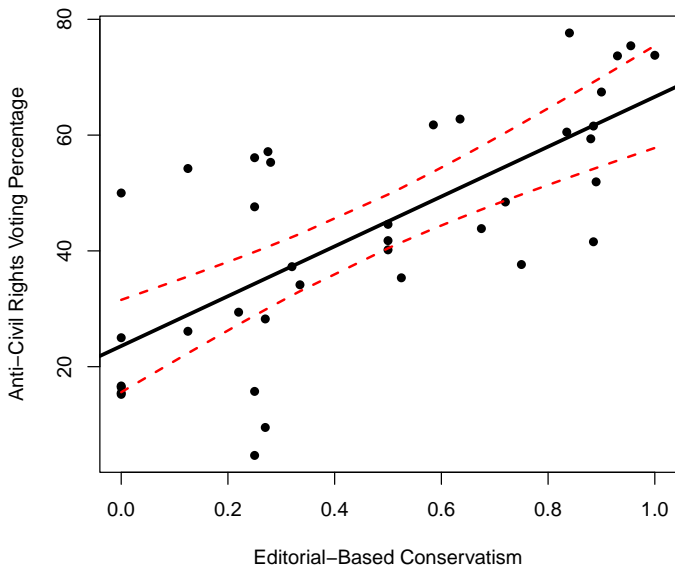
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14 on 36 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.505

F-statistic: 38.7 on 1 and 36 DF, p-value: 3.51e-07

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
-26.43	-10.59	2.46	7.86	29.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.00000000000000025	2.2758451083205631	0.00	1
IdeolCentered	43.0436722235377758	6.9171283031872104	6.22	0.00000035 ***

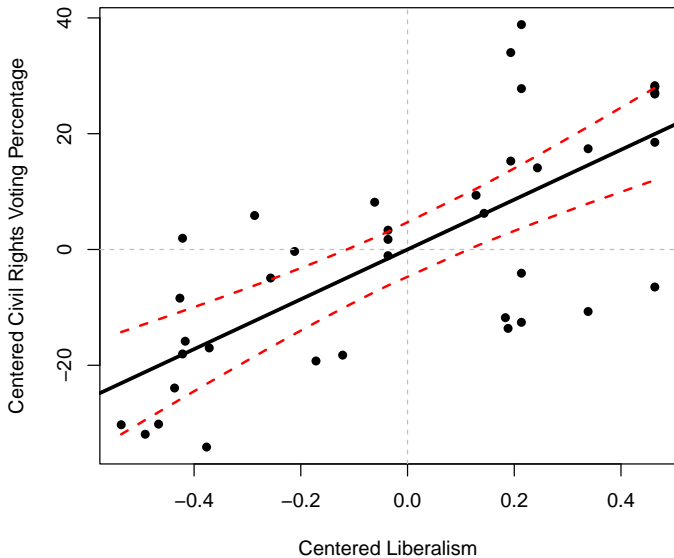
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 36 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.505

F-statistic: 38.7 on 1 and 36 DF, p-value: 0.000000351

“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
-26.43	-10.59	2.46	7.86	29.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	56.49	2.28	24.82	< 0.0000000000000002 ***
IdeolStd	14.35	2.31	6.22	0.00000035 ***

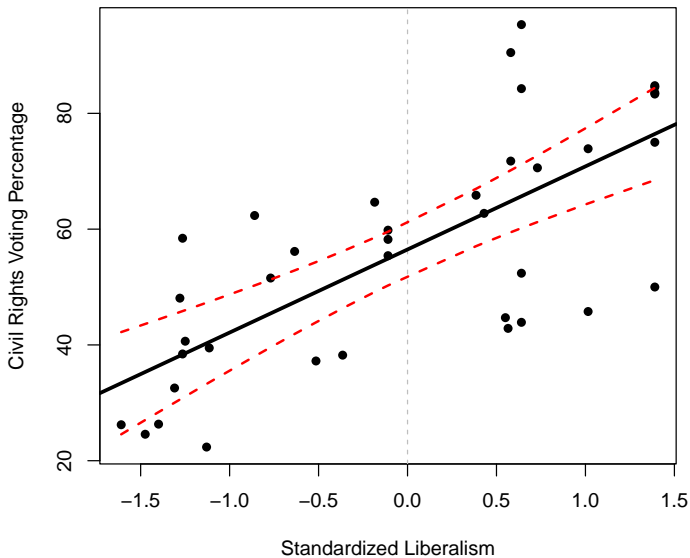
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14 on 36 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.505

F-statistic: 38.7 on 1 and 36 DF, p-value: 0.000000351

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
-31.25	-15.02	-2.38	14.75	37.45

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	635.783	246.825	2.58	0.014 *
Year	-0.294	0.125	-2.35	0.025 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 36 degrees of freedom

Multiple R-squared: 0.133, Adjusted R-squared: 0.109

F-statistic: 5.51 on 1 and 36 DF, p-value: 0.0245

Rescaling for Interpretability (continued)

```
> SCOTUS$Year1900<-SCOTUS$Year-1900
> fit8<-lm(CivLibs~Year1900,data=SCOTUS)
> summary(fit8)
```

Call:

```
lm(formula = CivLibs ~ Year1900, data = SCOTUS)
```

Residuals:

Min	1Q	Median	3Q	Max
-31.25	-15.02	-2.38	14.75	37.45

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	76.995	9.256	8.32	0.00000000067 ***
Year1900	-0.294	0.125	-2.35	0.025 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 36 degrees of freedom

Multiple R-squared: 0.133, Adjusted R-squared: 0.109

F-statistic: 5.51 on 1 and 36 DF, p-value: 0.0245

Binary $X \equiv t$ -test

```
> SCOTUS$Chief<-ifelse(is.na(SCOTUS$ChiefJustice),0,1)
> fit9<-lm(CivLibs~Chief,data=SCOTUS)
> summary(fit9)
```

Call:

```
lm(formula = CivLibs ~ Chief, data = SCOTUS)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	57.04	3.45	16.51	<0.0000000000000002 ***
Chief	-5.22	10.65	-0.49	0.63

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.1 on 36 degrees of freedom

Multiple R-squared: 0.00664, Adjusted R-squared: -0.021

F-statistic: 0.241 on 1 and 36 DF, p-value: 0.627

```
> t.test(CivLibs~Chief,data=SCOTUS,var.equal=TRUE)
```

Two Sample t-test

data: CivLibs by Chief

t = 0.49, df = 36, p-value = 0.6

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

-16.37 26.82

sample estimates:

mean in group 0 mean in group 1

57.04

51.81

The results:

```
> summary(fit)
```

Call:

```
lm(formula = CivLibs ~ IdeologyScore, data = SCOTUS)
```

Residuals:

Min	1Q	Median	3Q	Max
-26.43	-10.59	2.46	7.86	29.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.39	4.35	7.67	0.0000000044 ***
IdeologyScore	43.04	6.92	6.22	0.0000003505 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14 on 36 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.505

F-statistic: 38.7 on 1 and 36 DF, p-value: 0.000000351

The table:

Table: OLS Regression Model of SCOTUS Voting

Variables	Model I
(Constant)	33.39 (4.35)
Ideological Liberalism	43.04* (6.92)
Adjusted R^2	0.50

Note: $N = 38$. Cell entries are coefficient estimates; numbers in parentheses are estimated standard errors. Asterisks indicate $p < .05$ (one-tailed). See text for details.

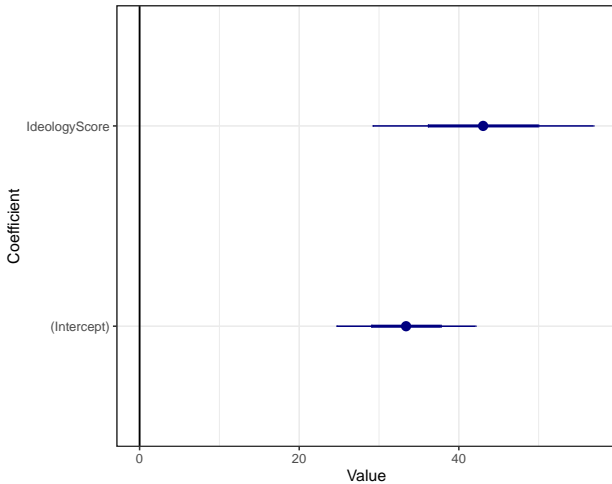
Another Table (using default-y stargazer)

Table: OLS Regression Model of SCOTUS Voting

	Model 1
(Constant)	33.39*** (4.35)
Ideological Liberalism	43.04*** (6.92)
Observations	38
R ²	0.52
Adjusted R ²	0.50
Residual Std. Error	14.03 (df = 36)
F Statistic	38.72*** (df = 1; 36)

Note: *p<0.1; **p<0.05; ***p<0.01

Default-y Ladderplot -fitplot-



Some 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 4-5 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.*