SMI 606: Prediction

Dr. Todd Hartman
Sheffield Methods Institute

Appearance and Voting

Which person is the more competent?

Appearance and Voting

Which person is the more competent?

2004 Senate race in Wisconsin
Russ Feingold (D) 55% vs. Tim Micheles (R) 44%

What about this Pair?

What about this Pair?

2010 Senate race in Wisconsin
Russ Feingold (D) 47% vs. Ron Johnson (R) 52%

Facial Appearance Experiment

Does the "competence measure" predict election outcome?

Name	Description
congress	session of congress
year	year of election
state	state of election
winner	name of winner
loser	name of runner-up
w.party	party of winner
l.party	party of loser
d.votes	number of votes, Democratic candidate
r.votes	number of votes, Republican candidate
d.comp	competence measure, Democratic candidate
r.comp	competence measure, Republican candidate

Let's Plot the Data

Load the data:

```
face <- read.csv("data/face.csv")
```

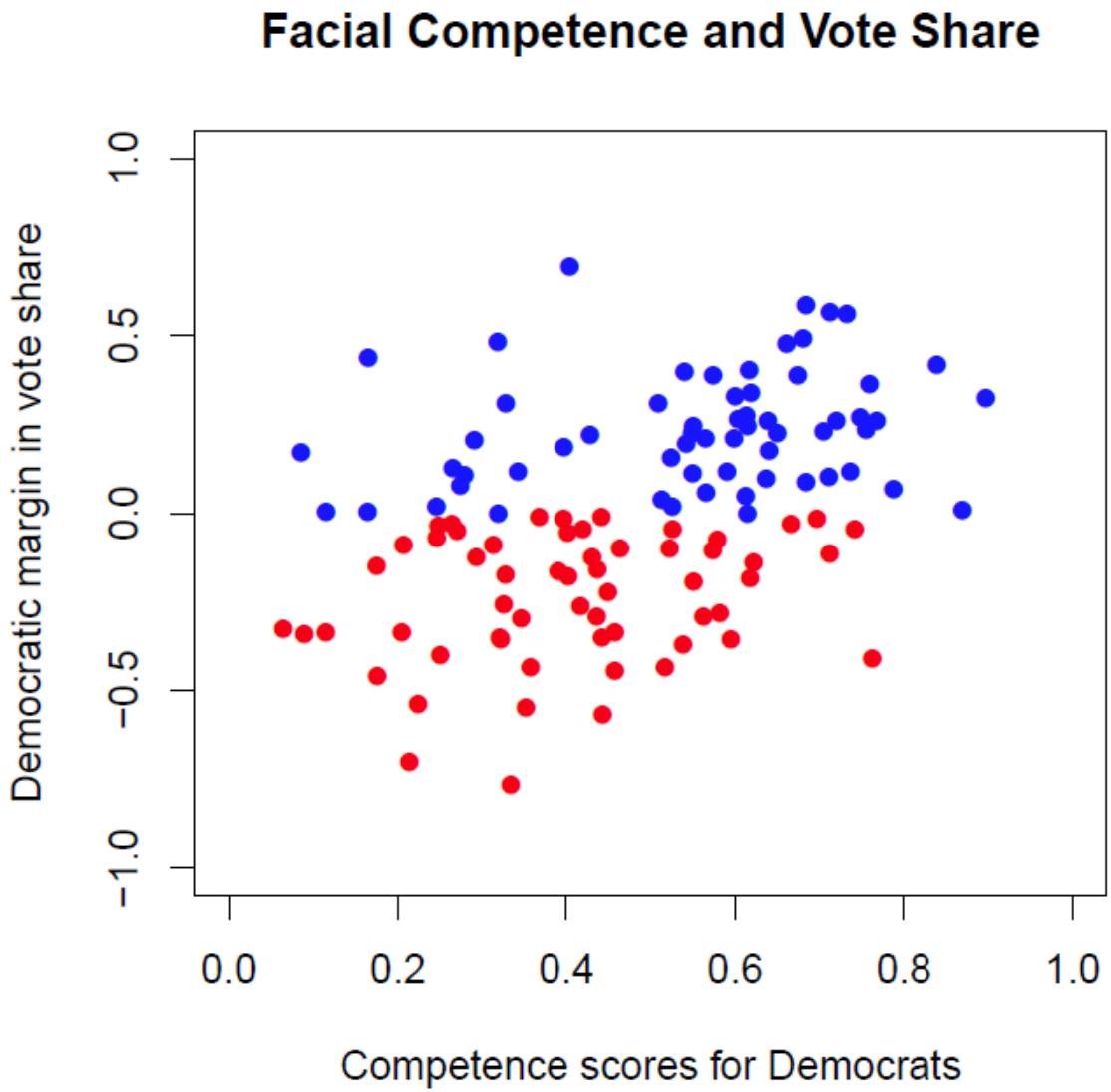
two-party vote share for Democrats and Republicans

```
face$d.share <- face$d.votes /
    (face$d.votes + face$r.votes)
face$r.share <- face$r.votes /
    (face$d.votes + face$r.votes)
face$diff.share <- face$d.share - face$r.share
```

Plot the data

```
## coerce to character
face$w.party <- as.character(face$w.party)
plot(face$d.comp, face$diff.share, pch = 16,
     col = ifelse(face$w.party == "R", "red", "blue"),
     xlim = c(0, 1), ylim = c(-1, 1),
     xlab = "Competence scores for Democrats",
     ylab = "Democratic margin in vote share",
     main = "Facial Competence and Vote Share")
```

Scatterplot of the Data



Correlation between Competence and Vote Share

Recall the definition of correlation

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \text{mean of } x}{\text{standard deviation of } x} \times \frac{y_i - \text{mean of } y}{\text{standard deviation of } y}\right)$$

$=$ mean of products of z-scores

correlation between competence measure and vote margin

```
cor(face$d.comp, face$diff.share)

## [1] 0.433
```

Linear Regression

Model:

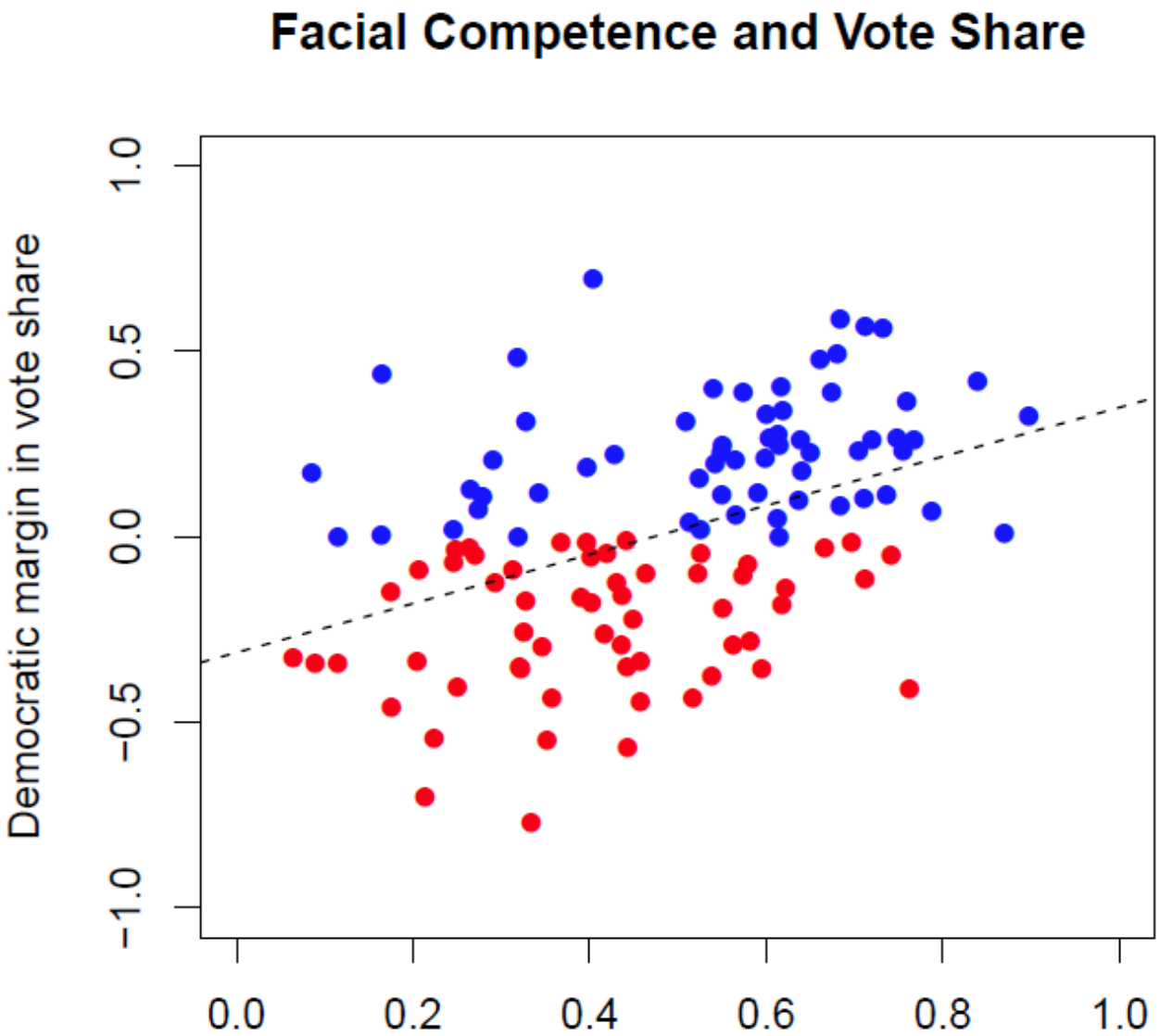
$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

- Y: dependent/outcome/response variable
- X: independent/explanatory variable, predictor
- (α, β): coefficients (parameters of the model)
- ϵ: unobserved error/disturbance term (mean zero)

Interpretation:

- $\alpha + \beta X$: mean of Y given the value of X
- α: the value of Y when X is zero
- β: increase in Y associated with one unit increase in X

Line of Best Fit



Least Squares

- Estimate the model parameters from the data
  - $(\hat{\alpha}, \hat{\beta})$: estimated coefficients
  - $\widehat{Y} = \hat{\alpha} + \hat{\beta}x$: predicted/fitted value
  - $\hat{\epsilon} = Y - \widehat{Y}$: residuals
- We obtain these estimates via the least squares method
- Minimize the sum of squared residuals (SSR):

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

- This also minimizes the root mean squared error: $\text{RMSE} = \sqrt{\frac{1}{n}\text{SSR}}$
- In **R**, use the `lm()` function and the `coef()` to extract the estimated coefficients

Fit the Model in R

Fit the model:

```
fit <- lm(diff.share ~ d.comp, data = face)
fit

##
## Call:
## lm(formula = diff.share ~ d.comp, data = face)
##
## Coefficients:
## (Intercept)       d.comp
##      -0.312        0.660
```

Obtain the estimated coefficients

```
coef(fit)

## (Intercept)      d.comp
##      -0.312       0.660
```

Some Algebra for Least Squares

Estimated coefficients in the mathematical expressions:

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{aligned}$$

Slope coefficient and correlation:

$$\hat{\beta} = \text{correlation of } X \text{ and } Y \times \frac{\text{standard deviation of } Y}{\text{standard deviation of } X}$$

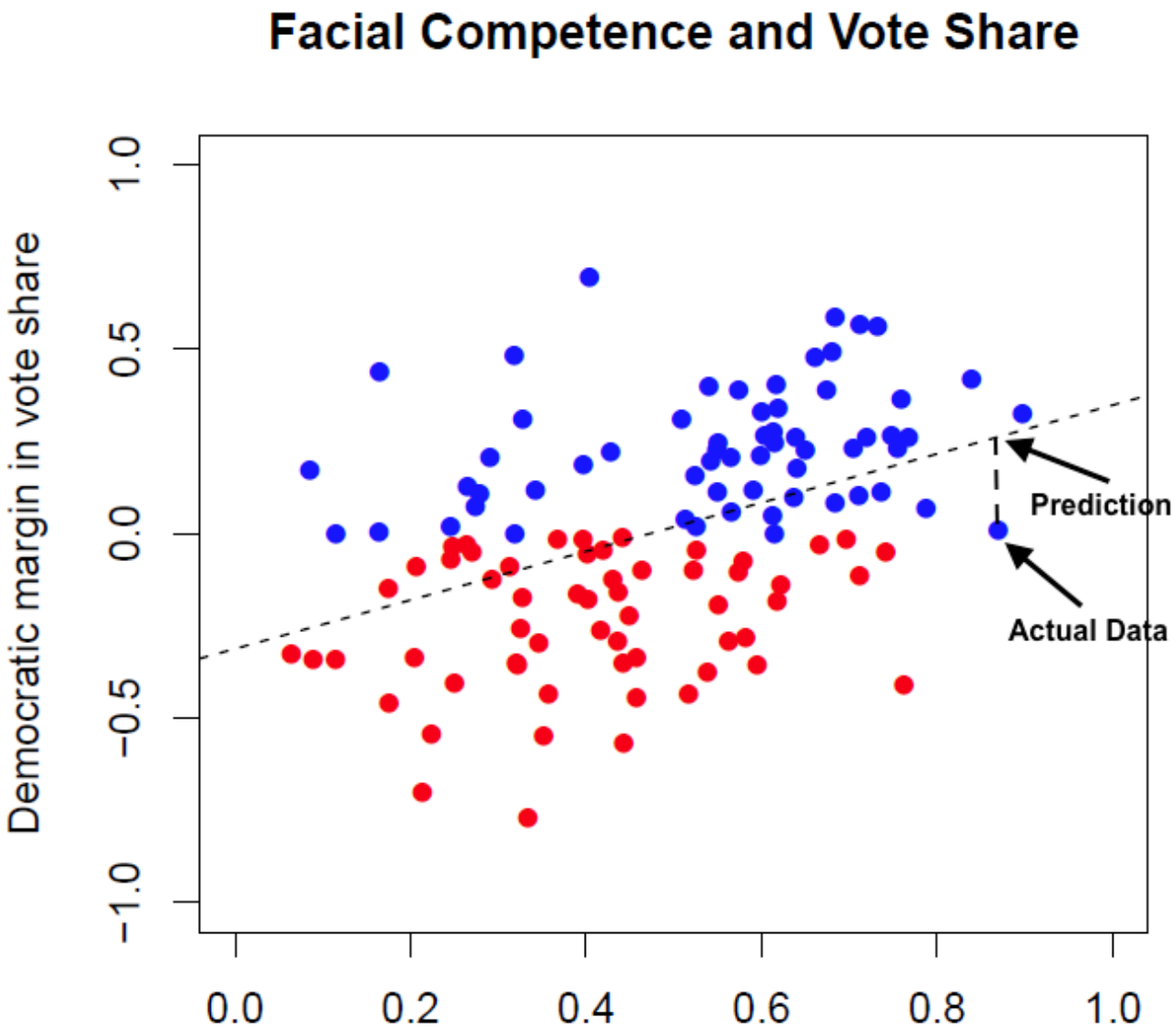
Least squares line always goes through $(\overline{X}, \overline{Y})$:

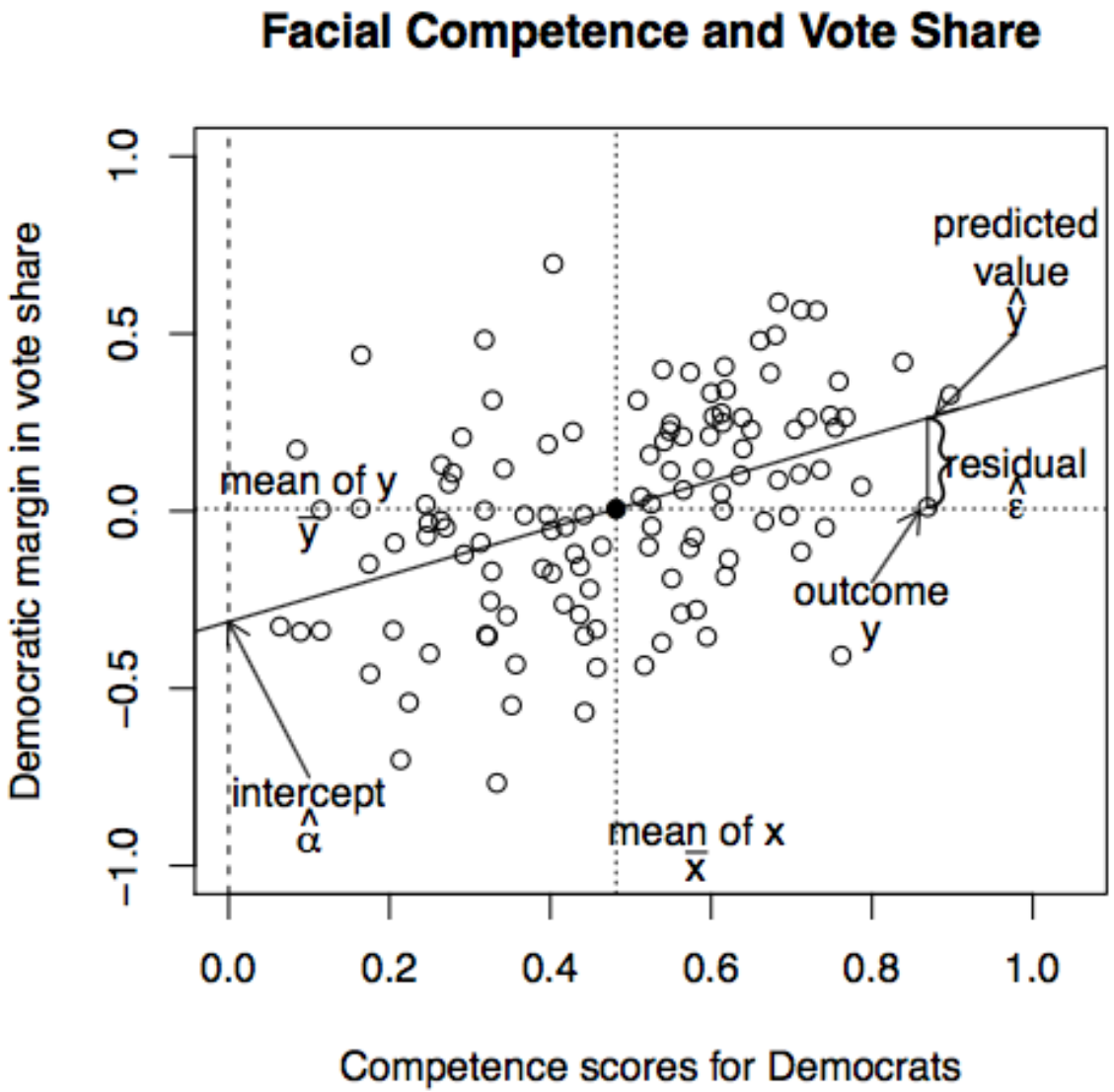
$$\widehat{Y} = (\overline{Y} - \hat{\beta}\overline{X}) + \hat{\beta}\overline{X} = \overline{Y}$$

Mean of residuals is always zero:

$$\text{mean of } \hat{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) = \overline{Y} - \hat{\alpha} - \hat{\beta}\overline{X} = 0$$

Residuals


Facial Competence and Vote Share
Democratic margin in vote share
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
1
Prediction
Actual Data


Facial Competence and Vote Share
Democratic margin in vote share
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
Competence scores for Democrats
predicted value
$\hat{y}$
residual
$\hat{\varepsilon}$
mean of y
$\bar{y}$
outcome
y
intercept
$\hat{\alpha}$
mean of x
$\bar{x}$

Assessing Goodness of Fit

How well does the model fit to the data?

How well does the model predict the outcome variable in the data?

Coefficient of determination or R^2:

$$R^2 = 1 - \frac{\text{SSR}}{\text{Total sum of squares (TSS)}} = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

Regression with Nominal Predictors

Because nominal predictors have no clear ordering, they must be *dummy coded*

- Coded 1 if attribute is present; 0 otherwise
- For m categories, need $m - 1$ dummy variables
- At least one category must be excluded (a.k.a., known as the reference, baseline, or excluded category
- All comparisons made to this group
- The value of the intercept is the group mean of the excluded category

Estimating SATE for Experimental Data

Recall: randomization of treatment assignment means we can approximate SATE with the difference in means between treatment and control conditions

Difference in Means Estimator

$$SATE = \underbrace{\bar{Y}_i(1)}_{\text{Treatment}} - \underbrace{\bar{Y}_i(0)}_{\text{Control}}$$

Estimating SATE Using Linear Regression

Linear Regression

$$\hat{y} = \beta_0 + \beta_1 \mathit{Treatment}$$

where β_0 is the group mean in the control condition,

β_1 is the difference in means between the treatment and control conditions, and

$\beta_0 + \beta_1$ is the group mean in the treatment condition.

Get Out the Vote Campaign Experiment

Get Out the Vote Difference in Means

```
## Find means by treatment condition
tapply(social$primary2006, social$messages, mean)
```

```
## Civic Duty    Control  Hawthorne  Neighbors
##  0.3145377  0.2966383  0.3223746  0.3779482
```

```
## Calculate SATE
mean(social$primary2006[social$messages == "Civic Duty"])
- mean(social$primary2006[social$messages == "Control"])
```

```
## [1] 0.01789934
```

```
mean(social$primary2006[social$messages == "Hawthorne"]) -
 mean(social$primary2006[social$messages == "Control"])
```

```
## [1] 0.02573631
```

```
mean(social$primary2006[social$messages == "Neighbors"]) -
 mean(social$primary2006[social$messages == "Control"])
```

```
## [1] 0.08130991
```

Get Out the Vote Regression Results

```
## [1] "Civic Duty" "Control"    "Hawthorne"  "Neighbors"
```

```
social$civic <- ifelse(social$messages == "Civic Duty", 1, 0)
social$hawthorne <- ifelse(social$messages == "Hawthorne", 1, 0)
social$neighbors <- ifelse(social$messages == "Neighbors", 1, 0)

## Estimate regression using lm()
results.1 <- lm(primary2006 ~ civic + hawthorne + neighbors, data = social)
summary(results.1)
```

```
##
## Call:
## lm(formula = primary2006 ~ civic + hawthorne + neighbors, data = social)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3780 -0.2966 -0.2966  0.6776  0.7034
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.296638   0.001058 280.393  < 2e-16 ***
## civic       0.017899   0.002592   6.905 5.03e-12 ***
## hawthorne   0.025736   0.002593   9.927  < 2e-16 ***
## neighbors   0.081310   0.002593  31.360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4627 on 305862 degrees of freedom
## Multiple R-squared:  0.003283,   Adjusted R-squared:  0.003273
## F-statistic: 335.8 on 3 and 305862 DF,  p-value: < 2.2e-16
```