# Regression with Uncertainty

Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

# Linear Regression Model

- Recall the model:

$$y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i,$$

  where $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{V}(\epsilon_i) = \sigma^2$
- Estimation of parameters via least squares:

$$\text{minimize SSR} \quad \text{where} \quad \text{SSR} \;=\; \sum_{i=1}^{n} \hat{\epsilon}_i^2 \;=\; \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Key Assumptions:
  1. Exogeneity: the mean of $\epsilon_i$ does not depend on $x_i$

$$\mathbb{E}(\epsilon_i \mid x_i) \;=\; \mathbb{E}(\epsilon_i) \;=\; 0$$

  2. Homoskedasticity: the variance of $\epsilon_i$ does not depend on $x_i$

$$\mathbb{V}(\epsilon_i \mid x_i) \;=\; \mathbb{V}(\epsilon_i) \;=\; \sigma^2$$

- When is each assumption violated?
- There is an easy fix for heteroskedasticity but not for endogeneity

# Statistical Properties of Least Squares

- Repeated hypothetical data generation:
  1. sample $(y_i, x_i)$ according to the model
  2. equivalently sample $(x_i, \epsilon_i)$ and then construct $y_i$
  3. run regression and obtain $(\hat{\beta}_0, \hat{\beta}_1)$
  4. repeat
- Under exogeneity, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased
- Under the two assumptions, standard errors are unbiased
- 95% confidence intervals:

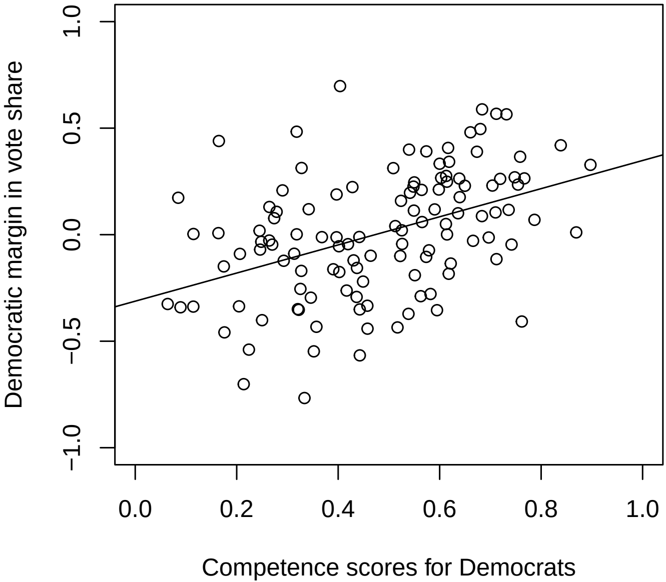
$$[\hat{\beta}_0 - z_{0.025} \cdot (\text{standard error of } \hat{\beta}_0),\ \hat{\beta}_0 + z_{0.025} \cdot (\text{standard error of } \hat{\beta}_0)]$$
$$[\hat{\beta}_1 - z_{0.025} \cdot (\text{standard error of } \hat{\beta}_1),\ \hat{\beta}_1 + z_{0.025} \cdot (\text{standard error of } \hat{\beta}_1)]$$

- Hypothesis test: $H_0 : \hat{\beta}_1 = \beta_1^*$
- test statistic: $\frac{\hat{\beta}_1 - \mathbb{E}(\hat{\beta}_1)}{\sqrt{\widehat{\mathbb{V}(\hat{\beta}_1)}}} = \frac{\hat{\beta}_1 - \beta_1^*}{\text{standard error of } \hat{\beta}_1} \overset{\text{approx.}}{\sim} \mathcal{N}(0,1)$
- Often, $t$-distribution is used

# Recall the Study of Facial Apperance and Politics


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

# Facial Competence and Vote Share



```
##
## Call:
## lm(formula = diff.share ~ d.comp, data = face)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.675 -0.166  0.014  0.177  0.743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.312      0.066   -4.73  6.2e-06 ***
## d.comp         0.660      0.127    5.19  8.9e-07 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.266 on 117 degrees of freedom
## Multiple R-squared:  0.187,Adjusted R-squared:  0.18
## F-statistic:   27 on 1 and 117 DF,  p-value: 8.85e-07
```

# Expected and Predicted Values

- Interpretation of $\beta_1$: the average increase in $Y_i$ associated with one unit increase in $X_i$
- Expected value: the average outcome given $X_i = x$
- Predicted value: the prediction of the outcome given $X_i = x$

- Point estimate: $\hat{\beta}_0 + \hat{\beta}_1 x$
- Standard error for expected value:

$$\sqrt{\mathbb{V}(\hat{\beta}_0 + \hat{\beta}_1 x)} = \sqrt{\mathbb{V}(\hat{\beta}_0) + 2x\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + x^2\mathbb{V}(\hat{\beta}_1)}$$

- Standard error for predicted value: $\sqrt{\mathbb{V}(\hat{\beta}_0 + \hat{\beta}_1 x) + \mathbb{V}(\epsilon)}$
- We can construct confidence intervals and conduct hypothesis testing in the same manner as before

```r
predict(fit, newdata = data.frame(d.comp = c(0.1, 0.5, 0.9)),
        se.fit = TRUE)
```

```
## $fit
##      1      2      3
## -0.246  0.018  0.282
##
## $se.fit
##      1      2      3
## 0.0544 0.0245 0.0585
##
## $df
## [1] 117
##
## $residual.scale
## [1] 0.266
```

`residual.scale` is the residual standard deviation

```r
predict(fit, newdata = data.frame(d.comp = c(0.1, 0.5, 0.9)),
        interval = "confidence", level = 0.9)
```

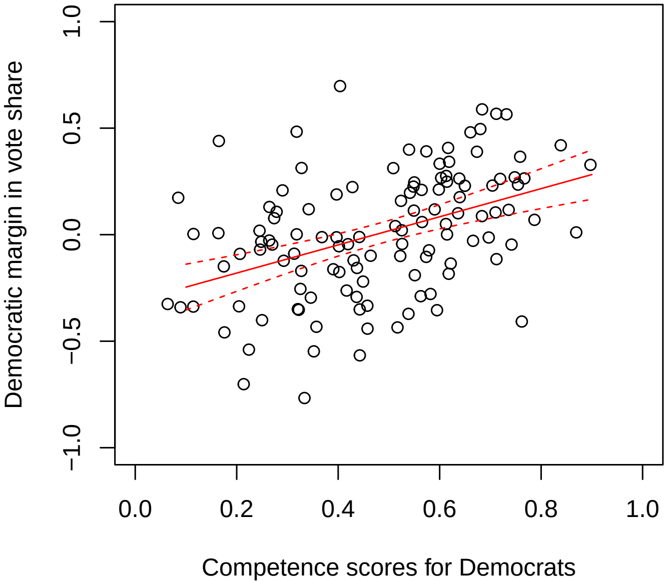
```
##      fit     lwr     upr
## 1 -0.246 -0.3363 -0.1561
## 2  0.018 -0.0227  0.0586
## 3  0.282  0.1851  0.3792
```

```r
predict(fit, newdata = data.frame(d.comp = c(0.1, 0.5, 0.9)),
        interval = "prediction", level = 0.9)
```

```
##      fit    lwr   upr
## 1 -0.246 -0.697 0.205
## 2  0.018 -0.426 0.462
## 3  0.282 -0.170 0.734
```

```r
x.pred <-
    data.frame(d.comp = seq(from = 0.1, to = 0.9, by = 0.01))
pred <- predict(fit, interval = "confidence",
                newdata = x.pred)
plot(face$d.comp, face$diff.share,
     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")
lines(x.pred$d.comp, pred[, "fit"], col = "red")
lines(x.pred$d.comp, pred[, "lwr"], col = "red",
      lty = "dashed")
lines(x.pred$d.comp, pred[, "upr"], col = "red",
      lty = "dashed")
```

# Facial Competence and Vote Share



# Statistical Inference with Multiple Regression

- Correlation does not imply causation
- Omitted variables $\Longrightarrow$ violation of exogeneity
- You can adjust for multiple confounding variables

$$y_i \;=\; \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Interpretation of $\beta_j$: an increase in the outcome associated with one unit increase in $x_{ij}$ when other variables take the same value
- Confidence intervals for $\hat{\beta}_j$, expected values, and predicted values can be constructed in the same manner
- Hypothesis testing for $\beta_j$, expected values, and predicted values, etc. can also conducted in the same manner

# Electoral Costs of Iraq War Casualties

- Outcome: Change in Bush's vote share from 2000 to 2004
- Multiple regression from Karol and Miguel (*J. of Politics* 2007)
- The average number of casualties per 100,000 = 3.39

| Variables | coef. | s.e. |
|---|---|---|
| Total Iraq deaths and wounded per 100,000 | −0.0055 | 0.0023 |
| Proportion active armed forces in 2000 | 0.43 | 0.26 |
| Proportion veterans in 2000 | −0.29 | 0.20 |
| Change in unemployment, 9/2003 – 8/2004 | −0.05 | 0.65 |
| Change in Black pop. prop., 2000 – 2003 | 2.15 | 0.66 |
| Change in White (non-Hispanic) pop. prop. | −0.35 | 0.60 |
| Proportional change in total population | −0.12 | 0.16 |
| Number of observations | 51 | |
| $R^2$ (coefficient of determination) | 0.41 | |