4. The linear regression model

Frank Edwards

School of Criminal Justice, Rutgers - Newark

Our old friend

$$y = \mathsf{X}\beta + \varepsilon$$

$$\varepsilon \sim Normal(0, \sigma^2)$$

OR

$$\mu = \mathsf{X}\beta$$

$$y \sim Normal(\mu, \sigma^2)$$

Basics of the linear model

- We assume a *linear* functional relationship between an outcome y and set of predictors X
- We assume that residual errors follow a Normal distribution with constant variance, centered around the regression line $E(y) = a + bx$
- Regression does not *automatically* produce estimates of ‘effects’. It compares group means, conditional on predictor values

Read in election data

```
# read in GHV election data
hibbs <- read_delim("./data/hibbs.dat")

glimpse(hibbs)
```

```
## Rows: 16
## Columns: 5
## $ year                <dbl> 1952, 1956, 1960, 1964, 1968, 1972, 1976, 1980, 19~
## $ growth              <dbl> 2.40, 2.89, 0.85, 4.21, 3.02, 3.62, 1.08, -0.39, 3~
## $ vote                <dbl> 44.60, 57.76, 49.91, 61.34, 49.60, 61.79, 48.95, 4~
## $ inc_party_candidate <chr> "Stevenson", "Eisenhower", "Nixon", "Johnson", "Hu~
## $ other_candidate     <chr> "Eisenhower", "Stevenson", "Kennedy", "Goldwater",~
```

```
m0 <- lm(vote ~ growth, data = hibbs)
```

Fit a simple model

For an election in year *i*, let's assume

$$vote_i = \beta_0 + \beta_1 growth_i + \varepsilon_i$$

```r
## Fit a model with lm
m0 <- lm(vote ~ growth, data = hibbs)
coef(m0)
```

```
## (Intercept)      growth
##   46.247648    3.060528
```

Fit a simple model

For an election in year *i*, let's assume

$$vote_i = \beta_0 + \beta_1 growth_i + \varepsilon_i$$

```r
## Fit a model with lm
m0 <- lm(vote ~ growth, data = hibbs)
coef(m0)
```

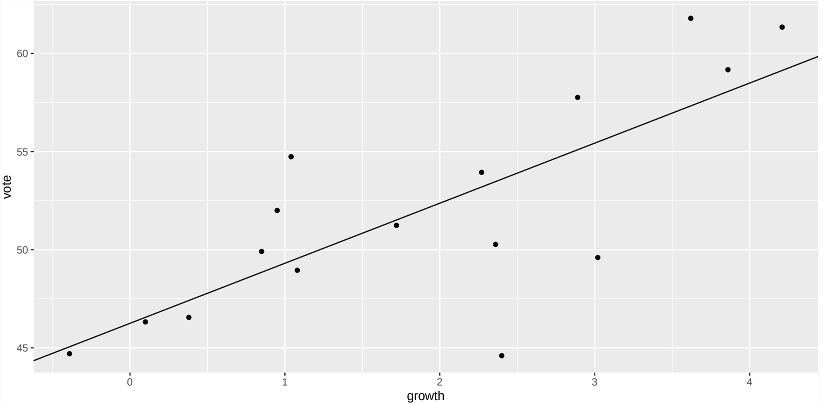
```
## (Intercept)      growth
##   46.247648    3.060528
```

Practice:

- What is the expected value of `vote` when growth = 3?
- When growth = 0?

Visualizing the fit

```r
ggplot(hibbs, aes(x = growth, y = vote)) + geom_point() + geom_abline(intercept = coef(m0)[1],
    slope = coef(m0)[2])
```



Continuous predictors in regression

- Parameters for continuous predictors act as *slopes*.
- Parameters for categorical predictors act as *intercepts*.

Adding a categorical predictor

For an election in year i, let's assume

$$vote_i = \beta_0 + \beta_1 war_i + \varepsilon_i$$

```
# add a predictor for major wars
hibbs <- hibbs %>%
    mutate(war = case_when(year >= 1950 & year <= 1953 ~ T, year >= 1964 & year <=
        1975 ~ T, year >= 2003 & year <= 2011 ~ T, T ~ F  # otherwise FALSE
))

m1 <- lm(vote ~ war, data = hibbs)
```

Adding a categorical predictor

For an election in year *i*, let's assume

$$vote_i = \beta_0 + \beta_1 war_i + \varepsilon_i$$

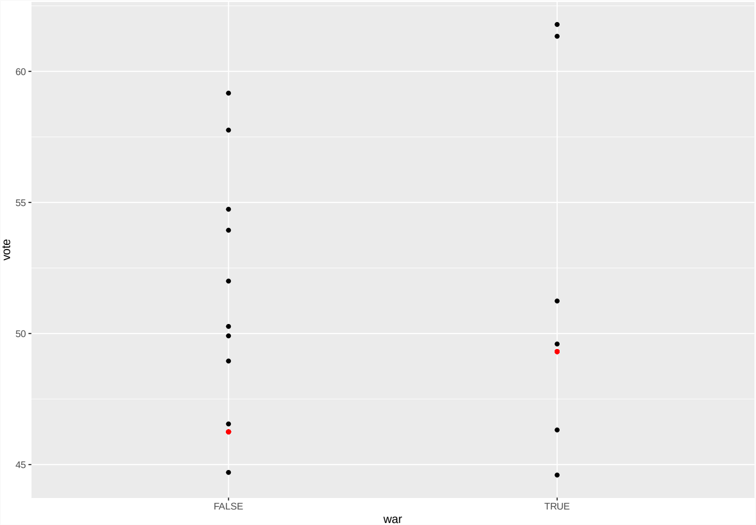
```
# add a predictor for major wars
hibbs <- hibbs %>%
    mutate(war = case_when(year >= 1950 & year <= 1953 ~ T, year >= 1964 & year <=
        1975 ~ T, year >= 2003 & year <= 2011 ~ T, T ~ F  # otherwise FALSE
))

m1 <- lm(vote ~ war, data = hibbs)
```

Practice:

- What is the expected value of `vote` when war = 1
- When war = 0

Visualizing a categorical predictor: black is observed, red is expected



Regression predictors

- Coefficients are not 'effects'
- Coefficients are differences in means of the outcome for different levels of the predictors

Regression with two predictors

$$vote_i = \beta_0 + \beta_1 growth_i + \beta_2 war_i + \varepsilon_i$$

```
m2 <- lm(vote ~ growth + war, data = hibbs)
coef(m2)
```

```
## (Intercept)      growth     warTRUE
##   46.607615    3.395281   -2.653763
```

Regression with two predictors

$$vote_i = \beta_0 + \beta_1 growth_i + \beta_2 war_i + \varepsilon_i$$

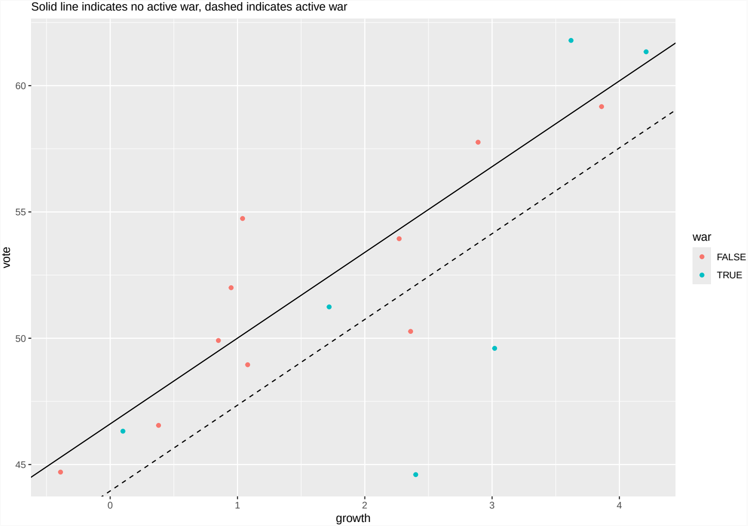
```r
m2 <- lm(vote ~ growth + war, data = hibbs)
coef(m2)
```

```
## (Intercept)      growth     warTRUE
##   46.607615    3.395281   -2.653763
```

Practice:

- What is the expected value of `vote` when war = 1 and growth = 2?
- When war = 0 and growth = 4?

Visualizing: two intercepts, one slope



We can specify that the relationship between growth and vote share may depend on whether there is a war. This model will have *two* slopes and two intercepts

$$vote_i = \beta_0 + \beta_1 growth_i + \beta_2 war_i + \beta_3 war_i \times growth_i + \varepsilon_i$$

```
m3 <- lm(vote ~ growth + war + growth * war, data = hibbs)
```

Continuous interactions

Maybe the relationship between growth and vote share changes over time?

$$vote_i = \beta_0 + \beta_1 growth_i + \beta_2 war_i + \beta_3 growth_i \times year_i + \varepsilon_i$$

```
m4 <- lm(vote ~ growth + war + growth * year, data = hibbs)
m4
```

```
##
## Call:
## lm(formula = vote ~ growth + war + growth * year, data = hibbs)
##
## Coefficients:
## (Intercept)       growth      warTRUE         year  growth:year
##  -304.39362    141.97857     -4.05982      0.17669     -0.06977
```

Continuous interactions

Maybe the relationship between growth and vote share changes over time?

$$vote_i = \beta_0 + \beta_1 growth_i + \beta_2 war_i + \beta_3 growth_i \times year_i + \varepsilon_i$$

```
m4 <- lm(vote ~ growth + war + growth * year, data = hibbs)
m4
```

```
##
## Call:
## lm(formula = vote ~ growth + war + growth * year, data = hibbs)
##
## Coefficients:
## (Intercept)       growth      warTRUE         year  growth:year
##  -304.39362    141.97857     -4.05982      0.17669     -0.06977
```

Practice:

- What is the expected value of `vote` when war = 1, growth = 2, and year = 1964?
- When war = 0, growth = 4, and year = 2012?

Interpretation when our specification is complicated

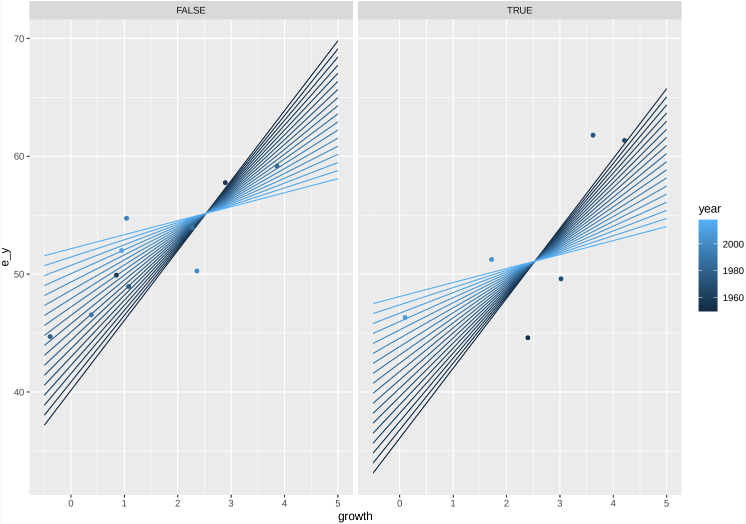
As our models get more complex, the parameters themselves start to become less meaningful on their own.

Rather than directly discussing parameter estimates, It can be helpful to discuss *expected values*

```
# let's simulate some data and generate predictions to better understand the
# model
year <- seq(1950, 2018, by = 4)
growth <- seq(-0.5, 5, by = 0.1)
war <- c(T, F)
# use expand_grid to make a data.frame with all unique combinations of these
# vectors
sim_dat <- expand_grid(year, growth, war)
# add expected values with predict
sim_dat <- sim_dat %>%
    mutate(e_y = predict(m4, sim_dat))
```

Visualization

```
ggplot(sim_dat, aes(x = growth, y = e_y, color = year, group = year)) + geom_line() +
    facet_wrap(~war) + geom_point(data = hibbs, aes(x = growth, y = vote))
```



Interpreting parameter estimate precision

```r
summary(m2)
```

```
##
## Call:
## lm(formula = vote ~ growth + war, data = hibbs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5025 -1.3303  0.0207  2.0617  5.5452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.6076     1.6056   29.03 3.31e-13 ***
## growth        3.3953     0.7255    4.68 0.000431 ***
## warTRUE      -2.6538     2.0251   -1.31 0.212722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 13 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.5718
## F-statistic: 11.01 on 2 and 13 DF,  p-value: 0.001592
```

- What does the standard error tell us?
- What about the t value?
- And the p value?
- Compute a 95% confidence interval for `growth`. What does this

Null hypothesis testing

We use t tests to compare what we observe in the data against a null hypothesis that there is no difference in the outcome y at different levels of x

- $H_0 : \beta_1 = 0$

Null hypothesis testing and the sampling distribution

Assume the null hypothesis is true. We model the *sampling distribution* of β_1 under this scenario, using the standard error we've estimated from the data

```r
library(broom)
tidy(m2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    46.6      1.61      29.0  3.31e-13
## 2 growth          3.40     0.726      4.68 4.31e- 4
## 3 warTRUE        -2.65     2.03      -1.31 2.13e- 1
```

- $H_0 : \beta_1 \sim N(0, 0.73)$

Null hypothesis testing and the sampling distribution

Assume the null hypothesis is true. We model the *sampling distribution* of β_1 under this scenario, using the standard error we've estimated from the data

```r
library(broom)
tidy(m2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    46.6      1.61      29.0  3.31e-13
## 2 growth          3.40     0.726      4.68 4.31e- 4
## 3 warTRUE        -2.65     2.03      -1.31 2.13e- 1
```

- $H_0 : \beta_1 \sim N(0, 0.73)$

How likely are we to observe $\beta_1 = 3.4$ if H_0 is true?

Using the Normal PDF

We can use the *probability density* of the Normal distribution to evaluate the probability of observing a value greater than or equal to 3.4 under the null hypothesis for the sampling distribution

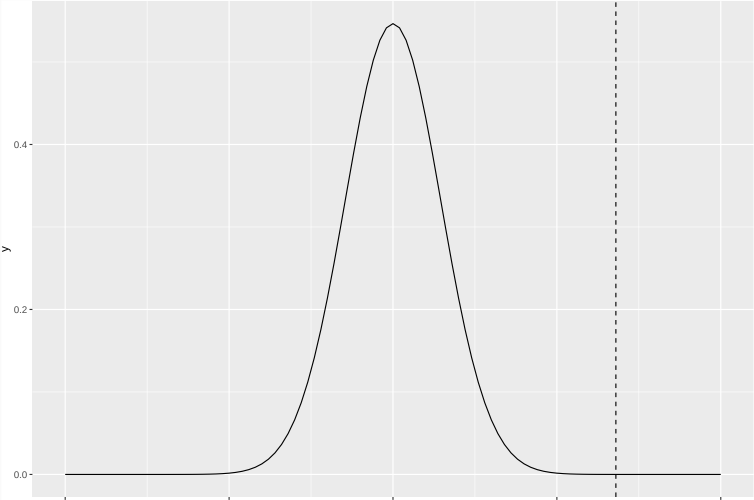
```
1 - pnorm(3.4, 0, 0.73)
```

```
## [1] 1.600096e-06
```

The Normal PDF for the Null and our data

```
plot_dat <- data.frame(x = seq(-5, 5, by = 0.1)) %>%
    mutate(y = dnorm(x, 0, 0.73))

ggplot(plot_dat, aes(x = x, y = y)) + geom_line() + geom_vline(xintercept = 3.4,
    lty = 2)
```



Interpretation

```r
summary(m2)
```

```
##
## Call:
## lm(formula = vote ~ growth + war, data = hibbs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5025 -1.3303  0.0207  2.0617  5.5452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.6076     1.6056   29.03 3.31e-13 ***
## growth        3.3953     0.7255    4.68 0.000431 ***
## warTRUE      -2.6538     2.0251   -1.31 0.212722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 13 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.5718
## F-statistic: 11.01 on 2 and 13 DF,  p-value: 0.001592
```

We conclude that it is very unlikely that we would observe a value like 3.4 if the null hypothesis were true, thus we say our estimate for β_1 is statistically significant

Interpretation

```r
summary(m2)
```

```
##
## Call:
## lm(formula = vote ~ growth + war, data = hibbs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5025 -1.3303  0.0207  2.0617  5.5452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.6076     1.6056   29.03 3.31e-13 ***
## growth        3.3953     0.7255    4.68 0.000431 ***
## warTRUE      -2.6538     2.0251   -1.31 0.212722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.67 on 13 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.5718
## F-statistic: 11.01 on 2 and 13 DF,  p-value: 0.001592
```

We conclude that it is very unlikely that we would observe a value like 3.4 if the null hypothesis were true, thus we say our estimate for β_1 is statistically significant

Practice

- Interpret and explain the statistical significance of our estimate for β_2
- What about the intercept β_0?

Parameter confidence intervals

We can construct confidence intervals for our parameters by using our point estimates and standard errors.

$$CI_{95}(\beta_1) = \beta_1 \pm 1.96 \times SE_{\beta_1}$$

```r
tidy(m2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    46.6      1.61      29.0  3.31e-13
## 2 growth          3.40     0.726      4.68 4.31e- 4
## 3 warTRUE        -2.65     2.03      -1.31 2.13e- 1
```

```r
## CI
3.4 + 1.96 * 0.73
```

```
## [1] 4.8308
```

```r
3.4 - 1.96 * 0.73
```

```
## [1] 1.9692
```

Parameter confidence intervals

We can construct confidence intervals for our parameters by using our point estimates and standard errors.

$$CI_{95}(\beta_1) = \beta_1 \pm 1.96 \times SE_{\beta_1}$$

```
tidy(m2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    46.6      1.61      29.0  3.31e-13
## 2 growth          3.40     0.726      4.68 4.31e- 4
## 3 warTRUE        -2.65     2.03      -1.31 2.13e- 1
```

```
## CI
3.4 + 1.96 * 0.73
```

```
## [1] 4.8308
```

```
3.4 - 1.96 * 0.73
```

```
## [1] 1.9692
```

If we were to repeat this experiment many times, 95 percent of our intervals would include the 'true' value of β_1. We have no guarantee of