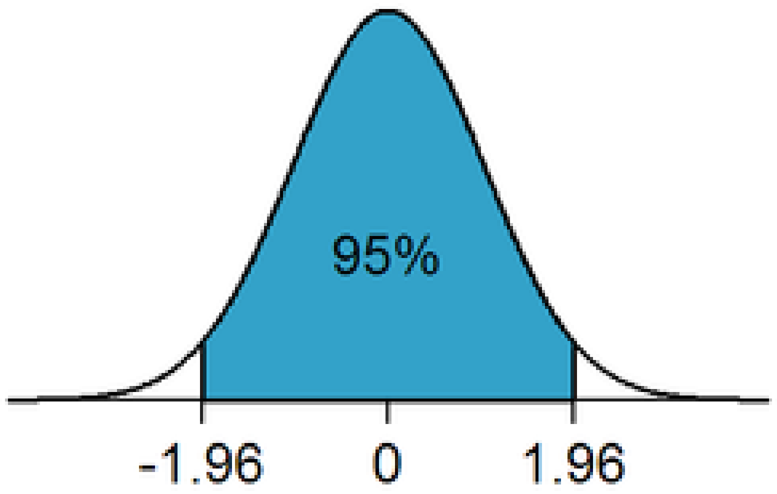
Multiple Regression

Understanding Political Numbers

March 11, 2019

Review

Review



A result is *statistically significant* if is was unlikely to have occurred by chance

We want to make inferences about the "true" parameters, but we only observe a sample of data.

Assuming that the null hypothesis is true, what would be the *probability* of observing our slope

An estimate is *significant* if the probability of getting it, under the null, is "sufficiently low"

Null relationships can still "pop" as significant, and "non-null" relationships may fail to show insignificance

What is a confidence interval?

All estimates are uncertain

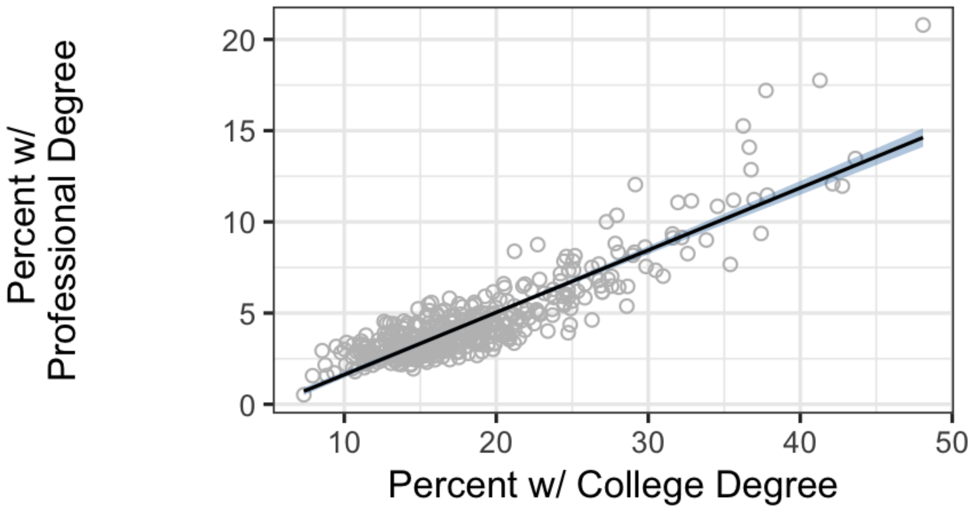
95% intervals contain "true parameter" 95% of the time

$\hat{y} = \alpha + \beta x$

Interval is $\text{Estimate} \pm \text{MOE}$

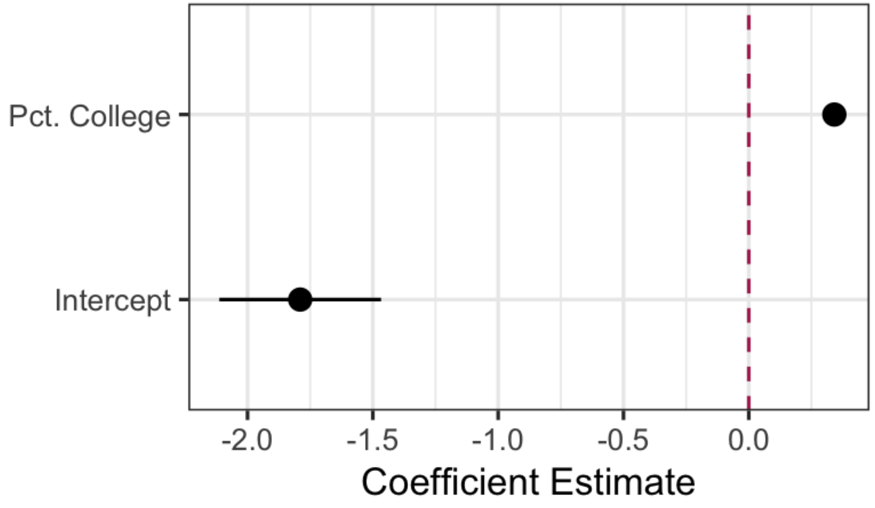
$b \pm (1.96 \times se(b))$

Software calculates CIs for you



Shaded area is 95% confidence interval

Coefficient Plot



Bars show 95% confidence intervals

Inference issues with p values

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

Insignificance does *not* mean "no relationship," only that there wasn't enough data to reject the null hypothesis

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

Insignificance does *not* mean "no relationship," only that there wasn't enough data to reject the null hypothesis

It takes *lots* of data to estimate small effects w/ statistical significance

Inference issues with *p* values

Null hypothesis testing: Higher quality learning by rejecting inconsistent ideas (*falsifying* the null? Probabilistically?)

If we want to be 95% confident, 5% of the "null models" will appear significant

Insignificance does *not* mean "no relationship," only that there wasn't enough data to reject the null hypothesis

It takes *lots* of data to estimate small effects w/ statistical significance

Relationships are everywhere, we just need enough data to make confident inferences about what they are

Multiple Regression

"Controlling for" other factors

"Controlling for" other factors

y affected by many potential w, x, z variables

"Controlling for" other factors

y affected by many potential w, x, z variables

Partial effect: what would happen to y if I *only* changed w

Or, the effect of w, "controlling for" x and z

"Controlling for" other factors

y affected by many potential w, x, z variables

Partial effect: what would happen to y if I *only* changed w

Or, the effect of w, "controlling for" x and z

SES and voting: Income or education?

"Controlling for" other factors

y affected by many potential w, x, z variables

Partial effect: what would happen to y if I *only* changed w

Or, the effect of w, "controlling for" x and z

SES and voting: Income or education?

Experiments!

Multiple regression

"Simple" or "bivariate" regression (two variables)

$$y = \alpha + \beta x + \epsilon$$

"Multiple regression" (many independent variables)

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \epsilon$$

Multiple regression

"Simple" or "bivariate" regression (two variables)

$$y = \alpha + \beta x + \epsilon$$

"Multiple regression" (many independent variables)

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \epsilon$$

Predicted value $\hat{y}$ a function of multiple x variables

β_1: the effect of x_1, *all else constant*

β_2: the effect of x_2, *all else constant*

α: value of $\hat{y}$ when *all* x variables are 0

ϵ: still leftover error

Interpreting Multiple Regression

```r
library("tidyverse")

# show the car data, convert to 'tibble'
mtcars %>%
  as_tibble(rownames = "model") %>%
  select(model, mpg, wt, disp)
```

```
## # A tibble: 32 x 4
##    model               mpg    wt  disp
##    <chr>             <dbl> <dbl> <dbl>
##  1 Mazda RX4          21    2.62  160
##  2 Mazda RX4 Wag      21    2.88  160
##  3 Datsun 710         22.8  2.32  108
##  4 Hornet 4 Drive     21.4  3.22  258
##  5 Hornet Sportabout  18.7  3.44  360
##  6 Valiant            18.1  3.46  225
##  7 Duster 360         14.3  3.57  360
##  8 Merc 240D          24.4  3.19  147.
##  9 Merc 230           22.8  3.15  141.
## 10 Merc 280           19.2  3.44  168.
## # … with 22 more rows
```

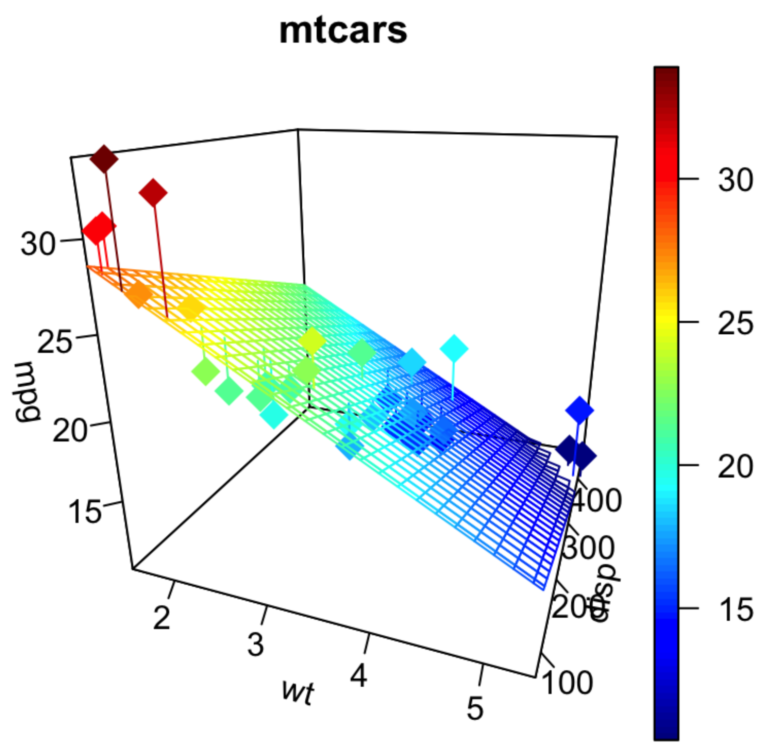
Interpreting Multiple Regression

```r
library("tidyverse")

# show the car data, convert to 'tibble'
mtcars %>%
  as_tibble(rownames = "model") %>%
  select(model, mpg, wt, disp)
```

```
## # A tibble: 32 x 4
##    model               mpg    wt  disp
##    <chr>             <dbl> <dbl> <dbl>
##  1 Mazda RX4          21    2.62  160
##  2 Mazda RX4 Wag      21    2.88  160
##  3 Datsun 710         22.8  2.32  108
##  4 Hornet 4 Drive     21.4  3.22  258
##  5 Hornet Sportabout  18.7  3.44  360
##  6 Valiant            18.1  3.46  225
##  7 Duster 360         14.3  3.57  360
##  8 Merc 240D          24.4  3.19  147.
##  9 Merc 230           22.8  3.15  141.
## 10 Merc 280           19.2  3.44  168.
## # … with 22 more rows
```

$$\text{Miles per gallon} = \alpha + \beta_1 \text{weight} + \beta_2 \text{displacement} + \epsilon$$



Multiple Regression in R

```r
library("broom") # for tidy() function

# add independent variables with `+`
car_model <- lm(mpg ~ wt + disp,
                data = mtcars)

tidy(car_model, conf.int = TRUE)
```

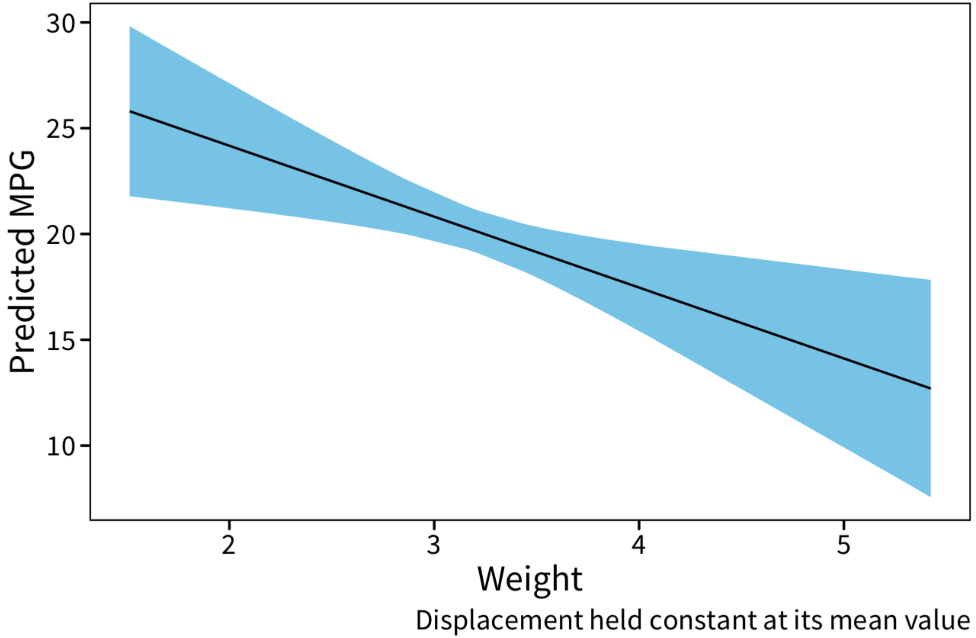
```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  35.0      2.16        16.2  4.91e-16  30.5     39.4
## 2 wt           -3.35     1.16        -2.88 7.43e- 3  -5.73    -0.970
## 3 disp         -0.0177   0.00919     -1.93 6.36e- 2  -0.0365   0.00107
```

Predictions from Multiple Regression

Conventionally: plot partial effect of one variable, holding everything else at their mean

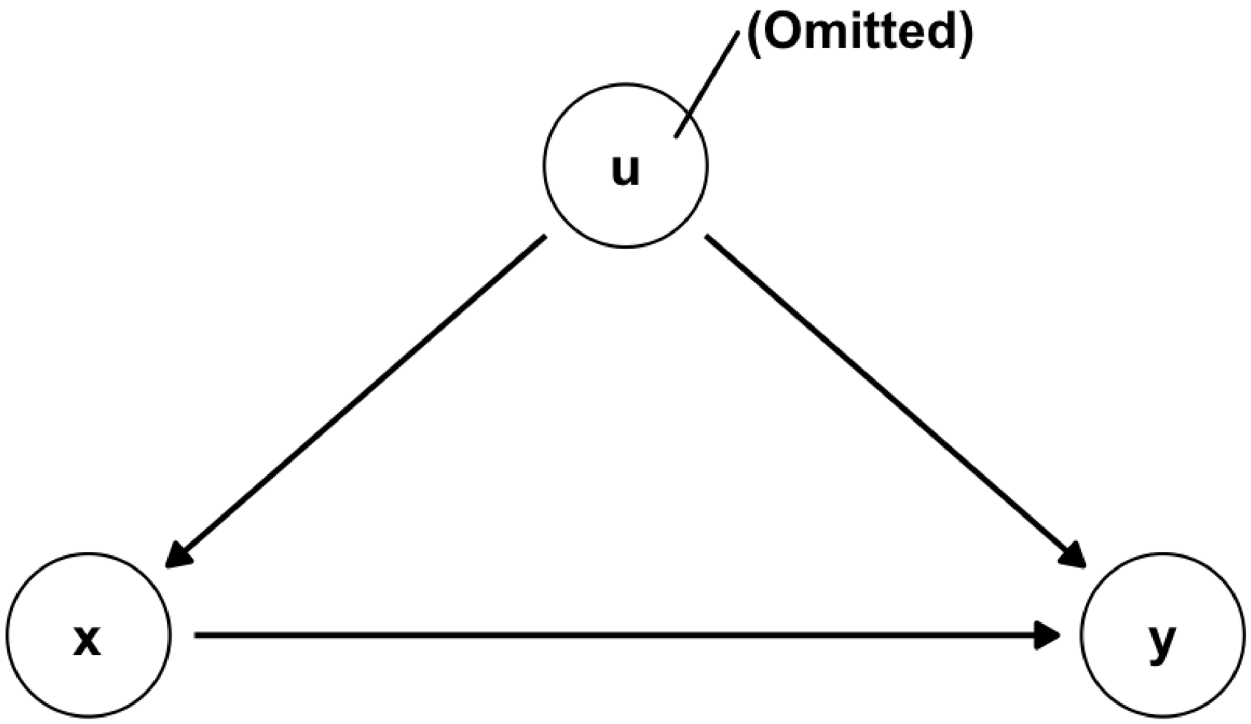
```
# new data frame; disp held at mean
vary_wt <- mtcars %>%
  select(mpg, wt, disp) %>%
  mutate(disp = mean(disp))

# predictions using augment()
wt_predictions <-
  augment(car_model, newdata = vary_wt) %>%
  mutate(MOE = 1.96 * .se.fit,
         lower_bound = .fitted - MOE,
         upper_bound = .fitted + MOE) %>%
  print()
```




```
## # A tibble: 32 x 8
##      mpg    wt  disp .fitted .se.fit   MOE lower_bound upper_bound
##    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>       <dbl>       <dbl>
##  1  21    2.62  231.    22.1   0.866  1.70        20.4        23.8
##  2  21    2.88  231.    21.2   0.652  1.28        20.0        22.5
##  3  22.8  2.32  231.    23.1   1.16   2.28        20.8        25.4
##  4  21.4  3.22  231.    20.1   0.516  1.01        19.1        21.1
```

(Spooky voice) Omitted Variable Bias



Causality Advice

Correlation ≠ causation

Bad controls

Better causality: control "upstream" variables

- Back-door paths
- Post-treatment bias

For advanced advice: [1] and [2]