inference for MLR

- significant predictors
- HT and CI for the slope
- interpretations

Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

modeling cognitive test scores of children

Data: Cognitive test scores of three- and four-year-old children and characteristics of their mothers (from a subsample from the National Longitudinal Survey of Youth).

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
…	…	…	…	…	…
6	98	no	107.90	no	18
…	…	…	…	…	…
434	70	yes	91.25	yes	25

R

```
# load data
> cognitive = read.csv("http://bit.ly/dasi_cognitive")

# full model
> cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive)
> summary(cog_full)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.59241    9.21906   2.125   0.0341 *
mom_hs:yes   5.09482    2.31450   2.201   0.0282 *
mom_iq       0.56147    0.06064   9.259   <2e-16 ***
mom_work:yes 2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
Multiple R-squared:  0.2171, Adjusted R-squared:  0.2098
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

inference for the model as a whole

H_0: $\beta_1 = \beta_1 = \ldots = \beta_k = 0$

H_A: At least one β_i is different than 0

```
F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-16
```

Since p-value < 0.05, the model as a whole is significant.

- The F test yielding a significant result doesn't mean the model fits the data well, it just means at least one of the βs is non-zero.
- The F test not yielding a significant result doesn't mean individuals variables included in the model are not good predictors of *y*, it just means that the combination of these variables doesn't yield a good model.

hypothesis testing for slopes

Is whether or not the mother went to high school a significant predictor of the cognitive test scores of children, given all other variables in the model?

H_0: $\beta_1 = 0$, when all other variables are included in the model

H_A: $\beta_1 \neq 0$, when all other variables are included in the model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

Whether or not mom went to high school is a significant predictor of the cognitive test scores of children, given all other variables in the model.

testing for the slope - mechanics

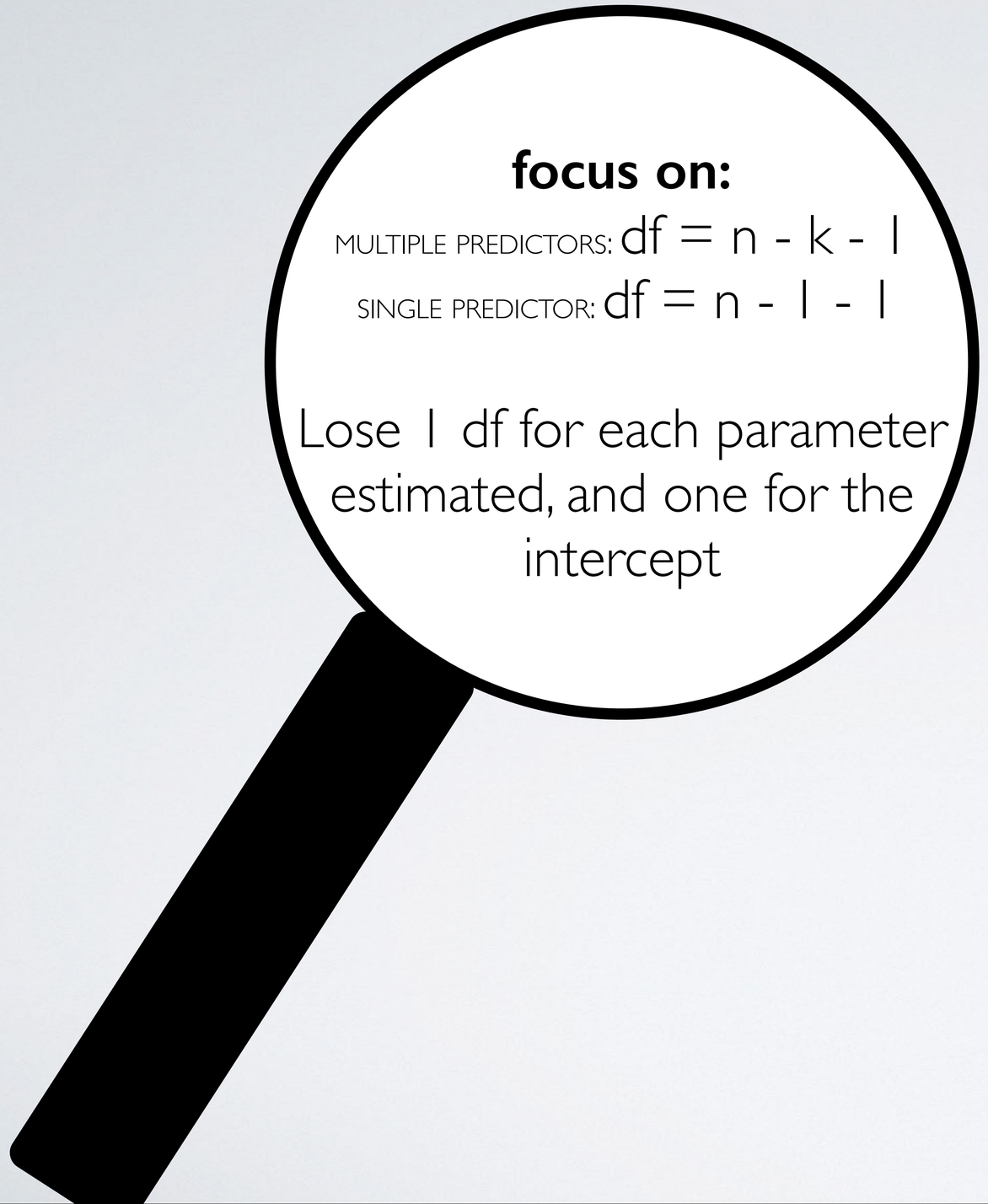
use a t-statistic in inference for regression

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

b_1 SE_{b_1}

t-statistic for the slope:

$$T = \frac{b_1 - 0}{SE_{b_1}} \qquad df = n - k - 1$$


focus on:
MULTIPLE PREDICTORS: df = n - k - 1
SINGLE PREDICTOR: df = n - 1 - 1
Lose 1 df for each parameter estimated, and one for the intercept

practice

Verify the T score and the p-value for the slope of **mom_hs**.

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     19.59241    9.21906   2.125   0.0341
mom_hs:yes       5.09482    2.31450   2.201   0.0282
mom_iq           0.56147    0.06064   9.259   <2e-16
mom_work:yes     2.53718    2.35067   1.079   0.2810
mom_age          0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = \frac{5.095 - 0}{2.315}$$

$$= 2.201$$

$$df = n - k - 1$$

$$= 434 - 4 - 1$$

$$= 429$$

R

```
> pt(2.201,df = 429, lower.tail = FALSE) * 2
[1] 0.0282
```

confidence intervals for slopes

point estimate ± margin of error

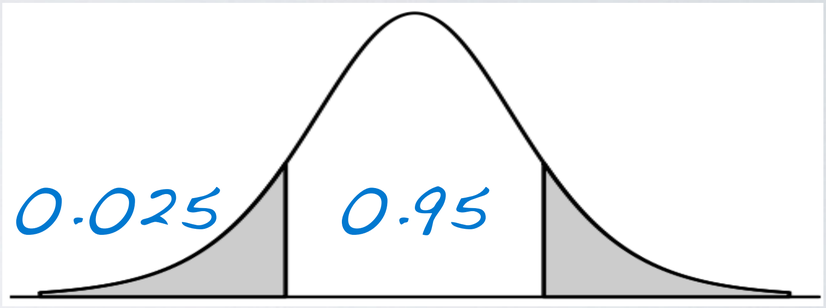
$$b_1 \pm t^\star_{df} SE_{b_1}$$

practice

Calculate the 95% confidence interval for the slope of **mom_work**.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hs:yes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_work:yes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

Residual standard error: 18.14 on 429 degrees of freedom



R

```
> qt(0.025, df = 429)
[1] -1.97
```

$df = 434 - 4 - 1 = 429$

$t^*_{429} = 1.97$

$2.54 \pm 1.97 \times 2.35 \approx (-2.09, 7.17)$

practice

Interpret the 95% confidence interval for the slope of **`mom_work`**.
CI: (-2.09, 7.17)

We are 95% confident that, all else being equal, the model predicts that children whose moms worked during the first three years of their lives score 2.09 points lower to 7.17 points higher than those whose moms did not work.