MULTIPLE REGRESSION

Interpretation, R^2, and F-test

OUTLINE OF TODAY

Interpretation of parameter estimates

Joint hypotheses

Change in R^2 (R^2-adjusted)

INTERPRETATION

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

β_1: The change in Y for a one-unit increase in X_1, holding X_2 and X_3 constant.

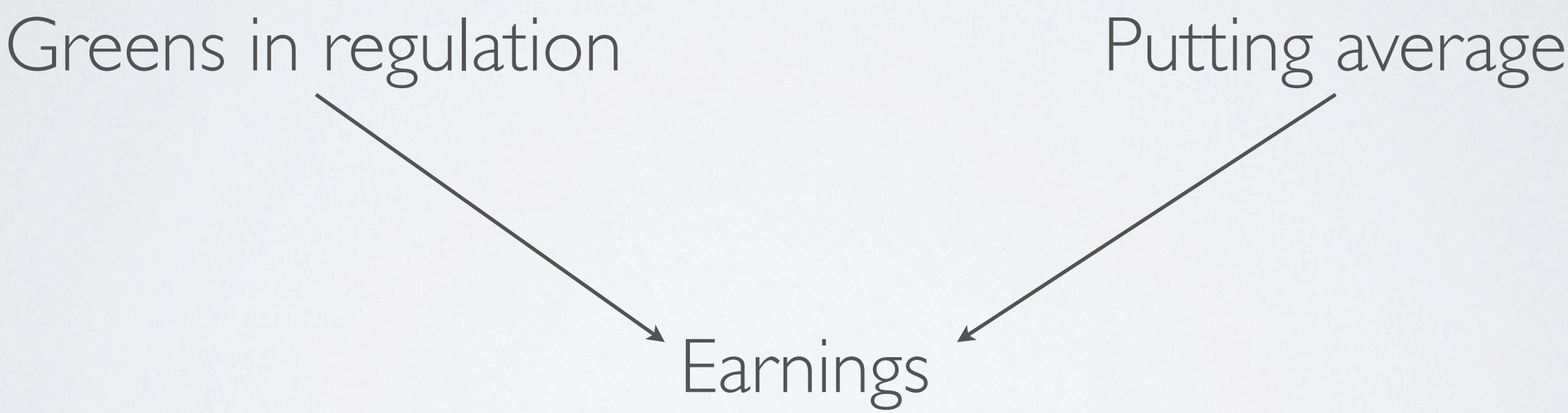
β_2: The change in Y for a one-unit increase in X_2, holding X_1 and X_3 constant.

β_3: The change in Y for a one-unit increase in X_3, holding X_1 and X_2 constant.

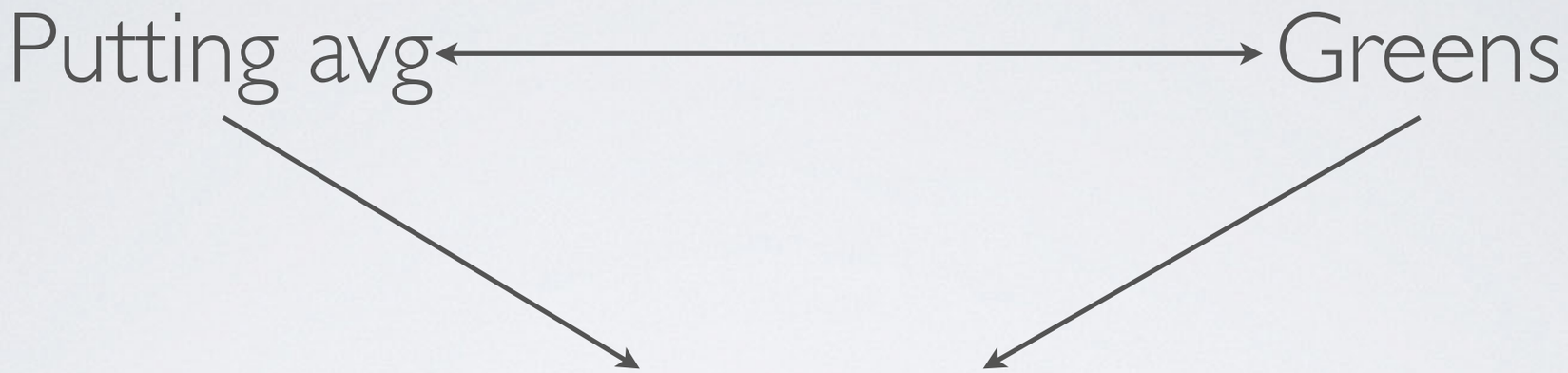
EXAMPLE - INTERPRETATION

LPGA players' earnings

Factors:



PARTIAL VS MARGINAL



"Overlap" between putting avg and greens in regulation

Why do we get that?

LPGA DATA

Statistic	N	Mean	St. Dev.	Min	Max
earnings.usd	30	812,099.700	425,880.700	451,981	2,588,240
scoring.avg	30	71.597	0.641	69.330	73.160
greens.in.reg	30	0.700	0.027	0.631	0.772
putting.avg	30	1.795	0.028	1.750	1.860

MODEL

Y: earnings

X: regulations, putting

$$Y = \beta_0 + \beta_1 \text{ regulations} + \varepsilon$$

$$Y = \beta_0 + \beta_2 \text{ putting} + \varepsilon$$

$$Y = \beta_0 + \beta_1 \text{ regulations} + \beta_2 \text{ putting} + \varepsilon$$

EARNINGS AND REGULATIONS

Expected effect?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5397213    1710971  -3.154  0.00382 **
Greens.in.Reg.  8875519    2443862   3.632  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 357300 on 28 degrees of freedom
Multiple R-squared:  0.3202, Adjusted R-squared:  0.2959
F-statistic: 13.19 on 1 and 28 DF,  p-value: 0.001117
```

EARNINGS AND PUTTING

Expected effect?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  14525433    4409895   3.294  0.00268 **
Putting.Avg. -7638322    2456016  -3.110  0.00427 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 373700 on 28 degrees of freedom
Multiple R-squared:  0.2567, Adjusted R-squared:  0.2302
F-statistic: 9.672 on 1 and 28 DF,  p-value: 0.004271
```

FULL RESULTS

```
> myReg1 <- lm(earnings.usd ~ greens.in.reg + putting.avg, data = LPGA)
> summary(myReg1)

Call:
lm(formula = earnings.usd ~ greens.in.reg + putting.avg, data = LPGA)

Residuals:
    Min      1Q  Median      3Q     Max
-782440 -204465    5995  162807  967070

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6349501    4538234   1.399  0.17316
greens.in.reg 7430690    2261177   3.286  0.00282 **
putting.avg  -5979899    2173234  -2.752  0.01046 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 321600 on 27 degrees of freedom
Multiple R-squared:  0.4691,    Adjusted R-squared:  0.4298
F-statistic: 11.93 on 2 and 27 DF,  p-value: 0.000194
```

T TEST

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

b_1 : parameter estimate

β_1 : null hypothesis (normally zero)

s_{b1} : standard error of b_1

PUTTING AND REGULATION?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.9644     0.1339  14.670 1.14e-14 ***
Greens.in.Reg. -0.2416     0.1913  -1.263    0.217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02797 on 28 degrees of freedom
Multiple R-squared:  0.05392, Adjusted R-squared:  0.02014
F-statistic: 1.596 on 1 and 28 DF,  p-value: 0.2169
```

ADJUSTED R^2

Problem: R^2 always increases with extra X variables

Solution: "Penalize" for adding X variable if it does not improve the model substantially

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{\sum_i^n (y_i - \bar{y})^2/(n-1)} = 1 - \frac{MSE}{s_y^2}$$

GOLF EXAMPLE

	R^2	R^2-adjusted
Only regulation	0.3201	0.2959
Only putting	0.2567	0.2302
Regulation and Putting	0.4691	0.4298

Note: summing R^2 for regulation and putting does not add to combined regression R^2 because non-zero correlation between regulation and putting!

R PRACTICE

- load LPGA data from Canvas
  - Note this is a R data set so use load!
- Run the 3 regressions with earnings as dependent variable
- Try to divide earnings with 100,000 and run multiple regression again. What happens to your result?

JOINT HYPOTHESES

WHEN?

- Model check
- Restrictions on coefficients

WHY NOT T-TEST?

- One (bad) approach: Test restrictions one at the time
- Test joint null hypothesis: $\beta_1=0$ and $\beta_2= 0$
- t_1 t-statistics for first
- t_2 t-statistics for second
- Reject null if either t_1 or $t_2 > 1.96$ (absolute values)

WHY NOT T-TEST?

Special case to make it easy: t-statistics are uncorrelated

H_0 only rejected if $|t_1| \leq 1.96 \;\&\; |t_2| \leq 1.96$

$$\Pr(|t_1| \leq 1.96) \times \Pr(|t_2| \leq 1.96)$$
$$= 0.95^2 = 0.9025$$

Probability of rejecting H_0 when true: 9.75%!

MODEL "VALID"

Example from above:

H_0: $\beta_{regulation} = \beta_{putting} = 0$

H_a: At least one β not equal to 0

HOW TO TEST

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

Degrees of freedom: k and n-k- l

Full model example:

F = 11.93

DF: 2, 27

Critical value for α=0.05: $F_{0.05,2,27} = 3.35$

Reject H_0: Model is valid

EXCLUSION RESTRICTIONS

F test can be used for testing whether a subset of variables jointly have a statistically significant effect

$$F = \frac{(SSE_r - SSE_{ur})/q}{SSE_{ur}/(n-k-1)}$$

SSE is sum of squared residuals

ALSO IN R^2 VERSION!

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k_u - 1)}$$

Problem: works only under homoskedasticity

EXAMPLE

- Use built-in data: mtcars
- Run two regression and compare
- Need package `car` as well
- Package option is heteroskedasticity-robust
- Only need one regression for package

```r
car_ur <- lm(mpg ~ cyl + disp + hp + wt, data = mtcars)
summary(car_ur)

# disp and hp closely correlated
cor(mtcars$disp,mtcars$hp)

# Neither are statistically significant but
# are they jointly significant?

# Restricted model
car_r <- lm(mpg ~ cyl + wt, data = mtcars)
summary(car_r)
```

ASIDE ON RESULTS IN R

- Most things can be saved as objects
- For example `car_ur_summary <- summary(car_ur)`
- Inspect both `car_ur` and `car_ur_summary`
- What are different and what are the same?
- Often no need to save intermediate objects

```r
# The R-squared version
r2_ur <- summary(car_ur)$r.squared
r2_r <- summary(car_r)$r.squared
q <- length(car_ur$coefficients) -
  length(car_r$coefficients)
n_k_1 <- (length(car_r$residuals) -
            length(car_ur$coefficients))
      # no need for -1 since this counts all coefficients
((r2_ur - r2_r)/q)/((1 - r2_ur) / n_k_1)
qf(.9, df1 = q, df2 = n_k_1) # CV for 10% sig level
```

PACKAGE FOR F-TEST

```r
# The function version
library(car)
Hnull <- c("disp = 0", "hp = 0")
linearHypothesis(car_ur, Hnull)
```

```
Linear hypothesis test

Hypothesis:
disp = 0
hp = 0

Model 1: restricted model
Model 2: mpg ~ cyl + disp + hp + wt

  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     29 191.17
2     27 170.44  2    20.728 1.6417 0.2124
```

Why different from our version?

OTHER HYPOTHESES

- Not just equal to zero joint tests!
- $\beta_1 = \beta_2$ (equal to $\beta_1 - \beta_2 = 0$)
- $\beta_1 = -\beta_2$ (equal to $\beta_1 + \beta_2 = 0$)
- $\beta_1 + \beta_2 = 1$ (equal to $\beta_1 - \beta_2 - 1 = 0$)

```r
# The silly equal to version
Hnull <- c("disp = hp")
linearHypothesis(car_ur, Hnull)
```

```
Linear hypothesis test

Hypothesis:
disp - hp = 0

Model 1: restricted model
Model 2: mpg ~ cyl + disp + hp + wt

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     28 189.38
2     27 170.44  1    18.934 2.9994 0.09471
```

IN-CLASS PROBLEM SET 2

- Same drill as the previous ones
- Getting correlation matrix on only some variables:

```
cor(autos[c("curb.weight", "horsepower", "speed.quarter.mile")])
```

- Or subset first (using select, for example)
- You can also use `ggpairs` (from `GGally` package)

FOR NEXT TIME

- ISL: Chapter 3.3.3
- Wooldridge: Chapters 8, 9.4-5