Unit 7: Multiple linear regression

1. Introduction to multiple linear regression

Sta 101 – Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

Coming up…

- Project Stage 2 Materials due 4/17 11:55pm (html file, RMD file, slides)
- Project Stage 2 Presentations 4/18
- Don't forget to **ask/answer 2 questions on Piazza** before the final exam… part of your participation grade! Memes don't count ☺

Outline

1. Housekeeping
2. Main ideas: Multiple Linear Regression (MLR)
    1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
    2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
    3. Inference for MLR: model as a whole + individual slopes
    4. Selecting Predictors for MLR:
        1. Adjusted R^2 applies a penalty for additional variables
        2. Avoid collinearity in MLR
        3. Model selection criterion depends on goal: significance vs. prediction
    5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR
3. Summary

Outline

1. Housekeeping
2. Main ideas: Multiple Linear Regression (MLR)
    1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
    2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
    3. Inference for MLR: model as a whole + individual slopes
    4. Selecting Predictors for MLR:
        1. Adjusted R^2 applies a penalty for additional variables
        2. Avoid collinearity in MLR
        3. Model selection criterion depends on goal: significance vs. prediction
    5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR
3. Summary

Outline

When do we need a multiple linear regression model?

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots$$

Outline

When do we need a multiple linear regression model?

- **More than one explanatory variable**
  - Ex: Number of Dependents and Party Affiliation

$$\widehat{income} = b_0 + b_1(\text{num. dependents}) + b_2(party: dem) + b_3(party: ind)$$

Outline

When do we need a multiple linear regression model?

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots$$

*more than 1 slope or predictor is needed.

Outline

When do we need a multiple linear regression model?

- **More than one explanatory variable**
  - Ex: Number of Dependents and Party Affiliation

$$\widehat{income} = b_0 + b_1(\text{num. dependents}) + b_2(party: dem) + b_3(party: ind)$$

- **One categorical explanatory variable with >2 levels**
  - Ex: Party Affiliation (Dem./Ind./Rep.)

$$\widehat{income} = b_0 + b_1(party: dem) + b_2(party: ind)$$

Outline

When do we need a multiple linear regression model?

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots$$

*more than 1 slope or predictor is needed.

*# of explanatory variables *not always equal* to the # of slopes/predictors in the linear regression equation

Outline

What is different about interpreting a slope in a *simple linear regression* vs. *multiple linear regression*?

*assume x_1 is numerical

$$\hat{y} = b_0 + \boldsymbol{b_1}x_1$$

$$\hat{y} = b_0 + \boldsymbol{b_1}x_1 + b_2x_2 + \cdots$$

Outline

1. Housekeeping

2. Main ideas: Multiple Linear Regression (MLR)
   1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
   2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
   3. Inference for MLR: model as a whole + individual slopes
   4. Selecting Predictors for MLR:
      1. Adjusted R^2 applies a penalty for additional variables
      2. Avoid collinearity in MLR
      3. Model selection criterion depends on goal: significance vs. prediction
   5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR

3. Summary

Outline

What is different about interpreting a slope in a *simple linear regression* vs. *multiple linear regression*?

*assume x_1 is numerical

$$\hat{y} = b_0 + \boldsymbol{b_1}x_1$$

"For one unit increase in x1, we would expect, y to increase/decrease, on average, by |b1|."

$$\hat{y} = b_0 + \boldsymbol{b_1}x_1 + b_2x_2 + \cdots$$

Outline

What is different about interpreting a slope in a *simple linear regression* vs. *multiple linear regression*?

*assume x_1 is numerical

$$\hat{y} = b_0 + \boldsymbol{b_1} x_1$$

"For one unit increase in x1, we would expect, y to increase/decrease, on average, by |b1|."

$$\hat{y} = b_0 + \boldsymbol{b_1} x_1 + b_2 x_2 + \cdots$$

"For one unit increase in x1, we would expect, y to increase/decrease, on average, by |b1|….. **BUT WHAT ARE THE OTHER VARIABLES GOING TO DO? HOW DO THEY AFFECT Y? If we deleted or added new explanatory variables to the model and recreated the multiple regression line, the slope may change.**

Outline

What is different about interpreting predictor slope for a level of a cat. variable vs. a slope for a numerical variable in a multiple linear regression?

Numerical Variable

$$\hat{y} = b_0 + \boldsymbol{b_1} x_1 + b_2 x_2 + \cdots$$

"**ALL ELSE HELD CONSTANT**, for one unit increase in x1, we would expect, y to increase/decrease, on average, by |b1|

Categorical Variable Predictor

$$\hat{y} = b_0 + \boldsymbol{b_1}(cat: level) + b_2 x_2 + \cdots$$

"**ALL ELSE HELD CONSTANT**, the predicted difference in y for this level and the baseline level is b1.

Outline

What is different about interpreting a slope in a *simple linear regression* vs. *multiple linear regression*?

*assume x_1 is numerical

$$\hat{y} = b_0 + \boldsymbol{b_1} x_1$$

"For one unit increase in x1, we would expect, y to increase/decrease, on average, by |b1|."

$$\hat{y} = b_0 + \boldsymbol{b_1} x_1 + b_2 x_2 + \cdots$$

"**ALL ELSE HELD CONSTANT**, for one unit increase in x1, we would expect, y to increase/decrease, on average, by |b1|

Data from the ACS

A random sample of 783 observations from the 2012 ACS.

1. `income`: Yearly income (wages and salaries)
2. `employment`: Employment status, not in labor force, unemployed, or employed
3. `hrs_work`: Weekly hours worked
4. `race`: White, Black, Asian, or other
5. `age`: Age
6. `gender`: male or female
7. `citizens`: Whether respondent is a US citizen or not
8. `time_to_work`: Travel time to work
9. `lang`: Language spoken at home, English or other
10. `married`: Whether respondent is married or not
11. `edu`: Education level, hs or lower, college, or grad
12. `disability`: Whether respondent is disabled or not
13. `birth_qrtr`: Quarter in which respondent is born, jan thru mar, apr thru jun, jul thru sep, or oct thru dec

Activity: MLR interpretations

1. Interpret the intercept.
2. Interpret the slope for hrs_work.
3. Interpret the slope for gender.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

Activity: MLR interpretations

Intercept: Income for the following type of person is expected to be on average -$1534.76:

a white, ____, non-citizen, that speaks English at home, that is not married, has HS or lower education level, without disabilities, born in _____, who works 0 hours/week, that is 0 years old, that takes 0 hours to get to work.

→ Baseline observation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

$$\widehat{income} = -15342.76$$

$$+1048.96(hrs_{work})$$

$$-7998.99(race: black) + 29909.8(race: asian) - 6756.32(race: other)$$

$$+565.07(age)$$

$$-17135.05(gender: female)$$

$$-12907.34(citizen: yes) +$$

$$+90.04(lang: other)$$

$$-10510.44(time_{to_{work}})$$

$$+5409.24(married: yes)$$

$$+15993.85(edu: college) + 59658.52(edu: grad)$$

$$-14142.79(disability: yes)$$

$$-2043.42(birth\_qrt: apr - jun) + 3036.02(birth\_qrt: jul - sep) + 2674.11(birth\_qrt: oct - dec)$$

Activity: MLR interpretations

Intercept: Income for the following type of person is expected to be on average -$1534.76:

a white, male, non-citizen, that speaks English at home, that is not married, has HS or lower education level, without disabilities, born in January-March, who works 0 hours/week, that is 0 years old, that takes 0 hours to get to work.

→ baseline

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

Activity: MLR interpretations

Slope for Hours Worked: All else held constant, if we increase the weekly hours worked by one hour, we would expect yearly income to increase on average by $1048.96.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

Outline

1. Housekeeping
2. Main ideas: Multiple Linear Regression (MLR)
    1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
    2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
    3. Inference for MLR: model as a whole + individual slopes
    4. Selecting Predictors for MLR:
        1. Adjusted R^2 applies a penalty for additional variables
        2. Avoid collinearity in MLR
        3. Model selection criterion depends on goal: significance vs. prediction
    5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR
3. Summary

Activity: MLR interpretations

Slope for Gender: All else held constant, the predicted difference in monthly income for females and males is -$17,135.05.

All else held constant, the model predicts that females earn $17,135.05 less than people that are not female ("male in this example.")

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

Outline

How many slopes would we need to model income with age and hours to get to work?

Outline

How many slopes would we need to model income with age and hours to get to work?

Response Variable:
- Income

2 Explanatory Variables:
- Age (numerical)
- Hours to get to work (numerical)

2 slopes being estimated in the model: b_1, b_2

$$\widehat{income} = b_0 + b_1 age + b_2 \, hrs\_work$$

Outline

How many slopes would we need to model income with citizen status (citizen or not) and race (white/black/asian/other)?

Response Variable:
- Income

2 Explanatory Variables:
- Citizen Status (categ. With 2 levels)
- Race (categ. With 4 levels)

4 slopes being estimated in the model: b_1, b_2, b_3, b_4

$$\widehat{income} = b_0 + b_1 citizen{:}\, yes + b_2 \, race{:}\, black + b_3 \, race{:}\, asian + b_4 \, race{:}\, other$$

Outline

How many slopes would we need to model income with citizen status (citizen or not) and race (white/black/asian/other)?

(2) Categorical predictors and slopes for (almost) each level

- Each **categorical variable, with w levels**, added to the model results in $w - 1$ **slopes** being estimated.
- It only takes $w - 1$ columns to code a categorical variable with w levels as 0/1s.
- Baseline: Level that's left out (ie: non-citizen)

$$\widehat{income} = b_0 + b_1 citizen{:}\, yes + b_2 \, race{:}\, black + b_3 \, race{:}\, asian + b_4 \, race{:}\, other$$

	citizen:yes
Obs 1	1
Obs 2	0
Obs 3	1
Obs 4	0
Obs 5	0
...	...

Represented by

	Original Data
Obs 1	Citizen
Obs 2	Non-citizen
Obs 3	Citizen
Obs 4	Non-citizen
Obs 5	Non-citizen
...	...

(2) Categorical predictors and slopes for (almost) each level

- Each **categorical variable, with w levels**, added to the model results in $w-1$ **slopes** being estimated.
- It only takes $w-1$ columns to code a categorical variable with w levels as 0/1s.
- Baseline: Level that's left out (ie: white)

$$\widehat{income} = b_0 + b_1 citizen{:}\ yes + b_2\ race{:}\ black + b_3\ race{:}\ asian + b_4\ race{:}\ other$$

Represented by

	Race
Obs 1	White
Obs 2	Black
Obs 3	Asian
Obs 4	White
Obs 5	Other
...	...

	race:black	race:asian	race:other
Obs 1	0	0	0
Obs 2	1	0	0
Obs 3	0	1	0
Obs 4	0	0	0
Obs 5	0	0	1
...	...	...	...

Clicker question

All else held constant, how do incomes of those born January thru March compare to those born April thru June?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

baseline

All else held constant, those born Jan thru Mar make, on average,

(a) $2,043.42 less (b) *$2,043.42 more* (c) $4978.12 less (d) $4978.12 more

than those born Apr thru Jun.

Clicker question

All else held constant, how do incomes of those born January thru March compare to those born April thru June?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

All else held constant, those born Jan thru Mar make, on average,

(a) $2,043.42 less (b) $2,043.42 more (c) $4978.12 less (d) $4978.12 more

than those born Apr thru Jun.

Outline

1. Housekeeping
2. Main ideas: Multiple Linear Regression (MLR)
   1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
   2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
   3. Inference for MLR: model as a whole + individual slopes
   4. Selecting Predictors for MLR:
      1. Adjusted R^2 applies a penalty for additional variables
      2. Avoid collinearity in MLR
      3. Model selection criterion depends on goal: significance vs. prediction
   5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR
3. Summary

What is different about conducting inference with a simple linear regression model and a multiple linear regression model?

How do we conduct inference for a intercept or individual coefficient in a regression model?

What is different about conducting inference with a simple linear regression model and a multiple linear regression model?

The inferences we conduct with a MLR must consider the presence of:

- *all of the variables* and
- this *specific combination* of variables.

Hypotheses:

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

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

Test Statistic: $T_{n-k-1} = \frac{b_i - 0}{SE_{b_i}}$ **k=# of slopes being estimated*

p-value=$\boldsymbol{p}(T_{n-k-1} > |test\ statistic|)$

(3) Inference for MLR: model as a whole + individual slopes

Hypotheses:

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

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

Test Statistic: $T_{n-k-1} = \frac{b_i - 0}{SE_{b_i}}$ **k=# of slopes being estimated*

p-value= $p(T_{n-k-1} > |test\ statistic|)$

In R:

```
library(oilabs)
data(acs12)
mod=lm(income~ hrs_work+race+age+gender+
                citizen+time_to_work+lang+
                married+edu+disability
                +birth_qrtr,data=acs12)
summary(mod)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -15342.76   11716.57  -1.309 0.190760
hrs_work         1048.96     149.25   7.028 4.63e-12 ***
raceblack       -7998.99    6191.83  -1.292 0.196795
raceasian       29909.80    9154.92   3.267 0.001135 **
raceother       -6756.32    7240.08  -0.933 0.351019
age               565.07     133.77   4.224 2.69e-05 ***
genderfemale    -17135.05    3705.35  -4.624 4.41e-06 ***
```

Clicker question

What is the **degrees of freedom** you would use to construct a confidence interval for the slope of the **age** variable? (Assume n=959).

10 Explanatory Variables

- Hrs_work
- Race
- Age
- Gender
- Citizen
- Lang
- Married
- Edu
- Disability
- birthqrtr

→ ***16 slope parameters***

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

(a) df = 959 – 10 - 1

(b) df = 959 – 16 - 1

(3) Inference for MLR: model as a whole + individual slopes

Confidence Interval for β_i

$b_i \pm T^*_{n-k-1} SE_{b_1}$

**k=# of slopes being estimated*

In R:

```
library(oilabs)
data(acs12)
mod=lm(income~ hrs_work+race+age+gender+
                citizen+time_to_work+lang+
                married+edu+disability
                +birth_qrtr,data=acs12)
summary(mod)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -15342.76   11716.57  -1.309 0.190760
hrs_work         1048.96     149.25   7.028 4.63e-12 ***
raceblack       -7998.99    6191.83  -1.292 0.196795
raceasian       29909.80    9154.92   3.267 0.001135 **
raceother       -6756.32    7240.08  -0.933 0.351019
age               565.07     133.77   4.224 2.69e-05 ***
genderfemale    -17135.05    3705.35  -4.624 4.41e-06 ***
```

Clicker question

What is the **degrees of freedom** you would use to construct a confidence interval for the slope of the **age** variable? (Assume n=959).

10 Explanatory Variables

- Hrs_work
- Race
- Age
- Gender
- Citizen
- Lang
- Married
- Edu
- Disability
- birthqrtr

→ ***16 slope parameters***

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

(a) df = 959 – 10 - 1

(b) df = 959 – 16 – 1 = n - k -1

**k = number of slopes in linear regression equation.*

Outline

How do we conduct inference on the whole multiple linear regression model?

(3) Inference for MLR: model as a whole + individual slopes

Hypotheses:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_A : At least one of the $\beta_i \neq 0$

F-statistic with $df_1 = k$, $df_2 = n - k - 1$

p-value

Outline

How do we conduct inference on the whole multiple linear regression model?

F-Test

(3) Inference for MLR: model as a whole + individual slopes

Hypotheses:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_A : At least one of the $\beta_i \neq 0$

F-statistic with $df_1 = k$, $df_2 = n - k - 1$

p-value

In R:

```
library(oilabs)
data(acs12)
mod=lm(income~ hrs_work+race+age+gender+
                citizen+time_to_work+lang+
                married+edu+disability
                +birth_qrtr,data=acs12)
summary(mod)
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -15342.76   11716.57  -1.309 0.190760
hrs_work                  1048.96     149.25   7.028 4.63e-12 ***
raceblack                -7998.99    6191.83  -1.292 0.196795
raceasian                29909.80    9154.92   3.267 0.001135 **
raceother                -6756.32    7240.08  -0.933 0.351019
age                        565.07     133.77   4.224 2.69e-05 ***
genderfemale            -17135.05    3705.35  -4.624 4.41e-06 ***
citizenyes              -12907.34    8231.66  -1.568 0.117291
time_to_work               90.04      79.83   1.128 0.259716
langother               -10510.44    5447.45  -1.929 0.054047 .
marriedyes                5409.24    3900.76   1.387 0.165932
educollege               15993.85    4098.99   3.902 0.000104 ***
edugrad                  59658.52    5660.26  10.540  < 2e-16 ***
disabilityyes           -14142.79    6639.40  -2.130 0.033479 *
birth_qrtrapr thru jun   -2043.42    4978.12  -0.410 0.681569
birth_qrtrjul thru sep    3036.02    4853.19   0.626 0.531782
birth_qrtroct thru dec    2674.11    5038.45   0.531 0.595752

Residual standard error: 48670 on 766 degrees of freedom
  (60 observations deleted due to missingness)
F-statistic: 21.77 on 16 and 766 DF,  p-value: < 2.2e-16
```

Outline

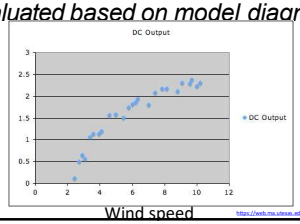
What does the F-Test for the whole linear regression model tell us?

Clicker question

True / False: The F test yielding a significant result means the model fits the data well.

(a) True
(b) False

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. Whether or not the model fit the data well is evaluated based on model diagnostics.



F-Test
p-value=7.5455E-12

Clicker question

True / False: The F test yielding a significant result means the model fits the data well.

(a) True
(b) False

Clicker question

True / False: The F test not yielding a significant result means individual variables included in the model are not good predictors of y.

(a) True
(b) False

Clicker question

True / False: The F test not yielding a significant result means individual variables included in the model are not good predictors of y.

(a) True

(b) False

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.

Outline

1. Housekeeping

2. Main ideas: Multiple Linear Regression (MLR)
   1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
   2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
   3. Inference for MLR: model as a whole + individual slopes
   4. Selecting Predictors for MLR:
      1. Adjusted R^2 applies a penalty for additional variables
      2. Avoid collinearity in MLR
      3. Model selection criterion depends on goal: significance vs. prediction
   5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR

3. Summary

Significance also depends on what else is in the model

A predictor may not be significant in one model…

```
Model 1:                 Estimate Std. Error t value Pr(>|t|)
(Intercept)           -15342.76   11716.57  -1.309 0.190760
hrs_work                1048.96     149.25   7.028 4.63e-12
raceblack              -7998.99    6191.83  -1.292 0.196795
raceasian              29909.80    9154.92   3.267 0.001135
raceother              -6756.32    7240.08  -0.933 0.351019
age                      565.07     133.77   4.224 2.69e-05
genderfemale          -17135.05    3705.35  -4.624 4.41e-06
citizenyes            -12907.34    8231.66  -1.568 0.117291
time_to_work             90.04      79.83   1.128 0.259716
langother             -10510.44    5447.45  -1.929 0.054047
marriedyes             5409.24    3900.76   1.387 0.165932 <----
educollege            15993.85    4098.99   3.902 0.000104
edugrad               59658.52    5660.26  10.540  < 2e-16
disabilityyes        -14142.79    6639.40  -2.130 0.033479
birth_qrtrapr thru jun  -2043.42    4978.12  -0.410 0.681569
birth_qrtrjul thru sep   3036.02    4853.19   0.626 0.531782
birth_qrtroct thru dec   2674.11    5038.45   0.531 0.595752
```

… but significant in another.

```
Model 2:     Estimate Std. Error t value Pr(>|t|)
(Intercept) -22498.2     8216.2  -2.738  0.00631
hrs_work       1149.7      145.2   7.919 7.60e-15
raceblack     -7677.5     6350.8  -1.209  0.22704
raceasian     38600.2     8566.4   4.506 7.55e-06
raceother     -7907.1     7116.2  -1.111  0.26683
age             533.1      131.2   4.064 5.27e-05
genderfemale -15178.9     3767.4  -4.029 6.11e-05
marriedyes     8731.0     3956.8   2.207  0.02762 <----
```

Outline

Should we have as many variables as possible in our multiple linear regression model?

Outline

Should we have as many variables as possible in our multiple linear regression model?

Goal: **Parsimonious Model** – *small* amount of variables which has *highest* predictive power.

Outline

What metric helps us find a parsimonious model?

$$R^2_{adj}$$

Outline

What metric helps us find a parsimonious model?

(4) Adjusted R^2 applies a penalty for additional variables

- When any variable is added to the model R^2 increases.

(4) Adjusted R^2 applies a penalty for additional variables

- When <u>any</u> variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

(4) Adjusted R^2 applies a penalty for additional variables

Multiple Linear Regression ANOVA

	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
Explanatory Variable 1	# of slopes for this expl. var. in model	◯	◯	◯	◯
Explanatory Variable 2	# of slopes for this expl. var. in model	◯	◯	◯	◯
...	...	...	...	...	...
Explanatory Variable w	# of slopes for this expl. var. in model	◯	◯	◯	◯
Residuals	n-k-1	$SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$MS_{Res}=SS_{Res}/Df_{Res}$		
Total	n-1	$SS_{Tot}=\sum_{i=1}^{n}(y_i-\bar{y})^2$			

$$R^2_{adj} = 1 - \left(\frac{SS_{error}}{SS_{total}} \times \frac{n-1}{n-k-1}\right)$$

*k=# of slopes in the model

$$R^2_{adj} = 1 - \left(\frac{SS_{resid}}{SS_{total}} \times \frac{n-1}{n-k-1}\right)$$

(4) Adjusted R^2 applies a penalty for additional variables

- When <u>any</u> variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1}\right)$$

The higher k gets, the higher the "penalty."

where n is the number of cases and k is the number of slopes <u>estimated in the model</u>.

The new (potentially smaller) $\frac{SSE}{SST}$ you get by adding a predictor to the model needs to be <u>small enough</u> to counterbalance the penalty of increasing k (ie: adding a new predictor).

```
Analysis of Variance Table

Response: income
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
hrs_work        1 3.0633e+11 3.0633e+11 129.3025 < 2.2e-16 ***
race            3 7.1656e+10 2.3885e+10  10.0821 1.608e-06 ***
age             1 7.6008e+10 7.6008e+10  32.0836 2.090e-08 ***
gender          1 4.8665e+10 4.8665e+10  20.5418 6.767e-06 ***
citizen         1 1.1135e+09 1.1135e+09   0.4700   0.49319
time_to_work    1 3.5371e+09 3.5371e+09   1.4930   0.22213
lang            1 1.2815e+10 1.2815e+10   5.4094   0.02029 *
married         1 1.2190e+10 1.2190e+10   5.1453   0.02359 *
edu             2 2.7867e+11 1.3933e+11 58.8131 < 2.2e-16 ***
disability      1 1.0852e+10 1.0852e+10   4.5808   0.03265 *
birth_qrtr      3 3.3060e+09 1.1020e+09   0.4652   0.70667
Residuals     766 1.8147e+12 2.3691e+09
Total         782 2.6399e+12
```

ANOVA Table for a Regression
- Use to calculate R^2
- Use to calculate Adjusted R^2

*R^2 and Adjusted R^2 *also* given in **coefficient output table** summary().

$$R^2_{adj} = 1 - \left(\frac{1.8147e+12}{2.6399e+12} \times \frac{783-1}{783-16-1}\right) \approx 1 - 0.7018 = 0.2982$$

Clicker question

True / False: For a model with at least one predictor, R^2_{adj} will always be smaller than R^2.

(a) True
(b) False

Clicker question

True / False: For a model with at least one predictor, R^2_{adj} will always be smaller than R^2.

(a) True
(b) False

Because k is positive, R^2_{adj} will always be smaller than R^2.

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right)$$

Always ≥ 1

Always larger than...

$$R^2 = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \right)$$

Clicker question

True / False: For a model with at least one predictor, R^2_{adj} will always be smaller than R^2.

(a) True
(b) False

Because k is positive, R^2_{adj} will always be smaller than R^2.

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right)$$

Always ≥ 1

$$R^2 = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \right)$$

Clicker question

True / False: For a model with at least one predictor, R^2_{adj} will always be smaller than R^2.

(a) True
(b) False

Because k is positive, R^2_{adj} will always be smaller than R^2.

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-k-1} \right)$$

Always ≥ 1

Always larger than...

Always smaller than...

$$R^2 = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \right)$$

Clicker question

True / False: Adjusted R^2 tells us the percentage of variability in the response variable explained by the model.

(a) True
(b) False

Outline

1. Housekeeping

2. Main ideas: Multiple Linear Regression (MLR)
   1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
   2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
   3. Inference for MLR: model as a whole + individual slopes
   4. Selecting Predictors for MLR:
      1. Adjusted R^2 applies a penalty for additional variables
      2. Avoid collinearity in MLR
      3. Model selection criterion depends on goal: significance vs. prediction
   5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR

3. Summary

Clicker question

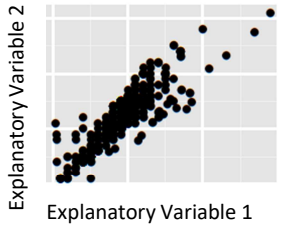
True / False: Adjusted R^2 tells us the percentage of variability in the response variable explained by the model.

(a) True
(b) False

R^2 tells us the percentage of variability in the response variable explained by the model, adjusted R^2 is only useful for model selection.

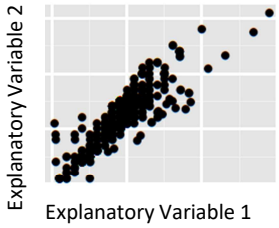
Outline

Why else should we be careful about choosing which explanatory variables to have in our model?



Outline

Why else should we be careful about choosing which explanatory variables to have in our model?



*Watch out for two explanatory variables (or more) that are **collinear**!*

(5) Avoid collinearity in MLR

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* (also called *multicollinearity*) complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables, so they should be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

(5) Avoid collinearity in MLR

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* (also called *multicollinearity*) complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables, so they should be independent of each other.

(5) Avoid collinearity in MLR

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* (also called *multicollinearity*) complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables, so they should be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

- In addition, **addition of collinear variables** can result in unreliable estimates of the slope parameters.

(5) Avoid collinearity in MLR

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* (also called *multicollinearity*) complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables, so they should be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- In addition, **addition of collinear variables** can result in unreliable estimates of the slope parameters.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to control for correlated predictors.

Outline

1. Housekeeping

2. Main ideas: Multiple Linear Regression (MLR)
   1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
   2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
   3. Inference for MLR: model as a whole + individual slopes
   4. Selecting Predictors for MLR:
      1. Adjusted R^2 applies a penalty for additional variables
      2. Avoid collinearity in MLR
      3. Model selection criterion depends on goal: significance vs. prediction
   5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR

3. Summary

Outline

What algorithms can/should we use to help select what explanatory variables should be in the model?

Outline

What algorithms can/should we use to help select what explanatory variables should be in the model?

Depends on your goal for making a model!

(6) Model selection criterion depends on goal: significance vs. prediction

backwards elimination - adjusted R^2

- Start with the full model
- Drop one variable at a time and record adjusted R^2 of each smaller model
- Pick the model with the highest increase in adjusted R^2
- Repeat until none of the models yield an increase in adjusted R^2

backwards elimination - p-value

- Start with the full model
- Drop the variable with the highest p-value and refit a smaller model
- Repeat until all variables left in the model are significant

forward selection - adjusted R^2

- Start with single predictor regressions of response vs. each explanatory variable
- Pick the model with the highest adjusted R^2
- Add the remaining variables one at a time to the existing model, and pick the model with the highest adjusted R^2
- Repeat until the addition of any of the remaining variables does not result in a higher adjusted R^2

forward selection - p-value

- Start with single predictor regressions of response vs. each explanatory variable
- Pick the variable with the lowest significant p-value
- Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
- Repeat until any of the remaining variables do not have a significant p-value

(6) Model selection criterion depends on goal: significance vs. prediction

- If the goal is to find the set of statistically significant predictors of y →use **p-value selection**.
- If the goal is to do better prediction of y→ use **adjusted R^2 selection**.

(6) Model selection criterion depends on goal: significance vs. prediction

- If the goal is to find the set of statistically significant predictors of y →use **p-value selection**.

(6) Model selection criterion depends on goal: significance vs. prediction

- If the goal is to find the set of statistically significant predictors of y →use **p-value selection**.
- If the goal is to do better prediction of y→ use **adjusted R^2 selection**.
- Either way, can use *backward elimination* or *forward selection*.

(6) Model selection criterion depends on goal: significance vs. prediction

- If the goal is to find the set of statistically significant predictors of y →use **p-value selection**.
- If the goal is to do better prediction of y→ use **adjusted R^2 selection**.
- Either way, can use *backward elimination* or *forward selection*.
- Expert opinion and focus of research might also demand that a particular variable be included in the model.

Clicker question

Using the p-value approach, which variable would you remove from the model below first?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14022.48	11137.08	-1.26	0.21
hrs_work	1045.85	149.05	7.02	0.00
raceblack	-7636.32	6177.50	-1.24	0.22
raceasian	29944.35	9137.13	3.28	0.00
raceother	-7212.57	7212.25	-1.00	0.32
age	559.51	133.27	4.20	0.00
genderfemale	-17010.85	3699.19	-4.60	0.00
citizenyes	-13059.46	8219.99	-1.59	0.11
time_to_work	88.77	79.73	1.11	0.27
langother	-10150.41	5431.15	-1.87	0.06
marriedyes	5400.41	3896.12	1.39	0.17
educollege	16214.46	4089.17	3.97	0.00
edugrad	59572.20	5631.33	10.58	0.00
disabilityyes	-14201.11	6628.26	-2.14	0.03

(a) married
(b) race
(c) race:other
(d) race:black
(e) time_to_work

Outline

Be careful about adding/deleting predictors that correspond to categorical explanatory variables with >2 levels….

Clicker question

Using the p-value approach, which variable would you remove from the model the model below first?

*In the p-value approach, the level with the SMALLEST p-value is chosen as representative for the categorical variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14022.48	11137.08	-1.26	0.21
hrs_work	1045.85	149.05	7.02	0.00
raceblack	-7636.32	6177.50	-1.24	0.22
raceasian	29944.35	9137.13	3.28	0.00
raceother	-7212.57	7212.25	-1.00	0.32
age	559.51	133.27	4.20	0.00
genderfemale	-17010.85	3699.19	-4.60	0.00
citizenyes	-13059.46	8219.99	-1.59	0.11
time_to_work	88.77	79.73	1.11	0.27
langother	-10150.41	5431.15	-1.87	0.06
marriedyes	5400.41	3896.12	1.39	0.17
educollege	16214.46	4089.17	3.97	0.00
edugrad	59572.20	5631.33	10.58	0.00
disabilityyes	-14201.11	6628.26	-2.14	0.03

(a) married
(b) race
(c) race:other
(d) race:black
(e) time_to_work

Clicker question

Using the p-value approach, which variable would you remove from the model below first?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

(a) race:other
(b) race
(c) time_to_work
(d) birth_qrtr:apr thru jun
(e) birth_qrtr

Outline

1. Housekeeping
2. Main ideas: Multiple Linear Regression (MLR)
   1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
   2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
   3. Inference for MLR: model as a whole + individual slopes
   4. Selecting Predictors for MLR:
      1. Adjusted R^2 applies a penalty for additional variables
      2. Avoid collinearity in MLR
      3. Model selection criterion depends on goal: significance vs. prediction
   5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR
3. Summary

Clicker question

Using the p-value approach, which variable would you remove from the model below first?

*Always remove ALL of the levels of a given categorical variable (for **p-value method** and **Adjusted R^2 method**)!

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15342.76	11716.57	-1.31	0.19
hrs_work	1048.96	149.25	7.03	0.00
raceblack	-7998.99	6191.83	-1.29	0.20
raceasian	29909.80	9154.92	3.27	0.00
raceother	-6756.32	7240.08	-0.93	0.35
age	565.07	133.77	4.22	0.00
genderfemale	-17135.05	3705.35	-4.62	0.00
citizenyes	-12907.34	8231.66	-1.57	0.12
time_to_work	90.04	79.83	1.13	0.26
langother	-10510.44	5447.45	-1.93	0.05
marriedyes	5409.24	3900.76	1.39	0.17
educollege	15993.85	4098.99	3.90	0.00
edugrad	59658.52	5660.26	10.54	0.00
disabilityyes	-14142.79	6639.40	-2.13	0.03
birth_qrtrapr thru jun	-2043.42	4978.12	-0.41	0.68
birth_qrtrjul thru sep	3036.02	4853.19	0.63	0.53
birth_qrtroct thru dec	2674.11	5038.45	0.53	0.60

(a) race:other
(b) race
(c) time_to_work
(d) birth_qrtr:apr thru jun
(e) birth_qrtr

(7) Conditions for MLR are (almost) the same as conditions for SLR

Important regardless of doing inference

- Linearity → randomly scattered residuals around 0 in the residuals plot

(7) Conditions for MLR are (almost) the same as conditions for SLR

Important regardless of doing inference

- Linearity → randomly scattered residuals around 0 in the residuals plot

Important for doing inference

- Nearly normally distributed residuals → histogram or normal probability plot of residuals
- Constant variability of residuals (*homoscedasticity*) → no fan shape in the residuals plot
- Independence of residuals (and hence observations) → depends on data collection method, often violated for time-series data

Clicker question

Which of the following is the appropriate plot for checking the homoscedasticity condition in MLR?

(a) scatterplot of residuals vs. $\hat{y}$
(b) scatterplot of residuals vs. x
(c) histogram of residuals
(d) normal probability plot of residuals
(e) scatterplot of residuals vs. order of data collection

(7) Conditions for MLR are (almost) the same as conditions for SLR

Important regardless of doing inference

- **Linearity** → randomly scattered residuals around 0 in the residuals plot (or use scatter plot) *slightly different for MLR (see next slides

Important for doing inference

- **Nearly normally distributed residuals** → histogram or normal probability plot of residuals
- **Constant variability of residuals** (*homoscedasticity*) → no fan shape in the residuals plot *slightly different for MLR (see next slides
- **Independence of residuals** (and hence observations) → depends on data collection method, often violated for time-series data
- Also important to make sure that your **explanatory variables are not *collinear*** (make sure each pair of explanatory variables does not have high correlation… can check in scatter plot or with R) *just for multiple linear regression

Clicker question

Which of the following is the appropriate plot for checking the homoscedasticity condition in MLR?

(a) scatterplot of residuals vs. $\hat{y}$
(b) scatterplot of residuals vs. x
(c) histogram of residuals
(d) normal probability plot of residuals
(e) scatterplot of residuals vs. order of data collection

Plotting residuals against $\hat{y}$(predicted, or fitted, values of y) allows us to evaluate the whole model as a whole as opposed to homoscedasticity with regards to just one of the explanatory variables in the model.

(7) Conditions for MLR are (almost) the same as conditions for SLR

Important regardless of doing inference

- **Linearity** → randomly scattered residuals around 0 in the residuals plot vs. fitted values plot (or use scatter plot)

Important for doing inference

- **Nearly normally distributed residuals** → histogram or normal probability plot of residuals
- **Constant variability of residuals** (*homoscedasticity*) → no fan shape in the residuals vs. fitted values plot
- **Independence of residuals** (and hence observations) → depends on data collection method, often violated for time-series data

*just for multiple linear regression

- Also important to make sure that your **explanatory variables are not *collinear*** (make sure each pair of explanatory variables does not have high correlation… can check in scatter plot or with R)

Final Regression Equation (after deleting some variables)

$$\widehat{CO} = -0.0586 + 0.7344(TAR) + 0.0267(LEN) - 6.1949(FILTER{:}NF) + 0.5597(PACK{:}SOFT) + 1.9077(STRENGTH{:}LIGHT) + 0.7900(STRENGTH{:}MEDIUM) + 0.5664(STRENGTH{:}REGULAR) + 3.0920(STRENGTH{:}FLAVOR)$$

Summary of main ideas

1. Interpreting Slopes/Coefficients: In MLR everything is conditional on all other variables in the model
2. Setting up the MLR Equation: Categorical predictors and slopes for (almost) each level
3. Inference for MLR: model as a whole + individual slopes
4. Selecting Predictors for MLR:
    1. Adjusted R^2 applies a penalty for additional variables
    2. Avoid collinearity in MLR
    3. Model selection criterion depends on goal: significance vs. prediction
5. Additional Conditions for MLR: Conditions for MLR are (almost) the same as conditions for SLR

Final Regression Equation (after deleting some variables)

$$\widehat{CO} = -0.0586 + 0.7344(TAR) + 0.0267(LEN) - 6.1949(FILTER{:}NF) + 0.5597(PACK{:}SOFT) + 1.9077(STRENGTH{:}LIGHT) + 0.7900(STRENGTH{:}MEDIUM) + 0.5664(STRENGTH{:}REGULAR) + 3.0920(STRENGTH{:}FLAVOR)$$

Prediction Input

$0.7344(TAR)$	12
$0.0267(LEN)$	80
$-6.1949(FILTER{:}NF)$	0
$0.5597(PACK{:}SOFT)$	0
$1.9077(STRENGTH{:}LIGHT)$	1
$0.7900(STRENGTH{:}MEDIUM)$	0
$0.5664(STRENGTH{:}REGULAR)$	0
$3.0920(STRENGTH{:}FLAVOR)$	0

Final Regression Equation (after deleting some variables)

$$\widehat{CO} = -0.0586 + 0.7344(12) + 0.0267(80) - 6.1949(0) + 0.5597(0) + 1.9077(1) + 0.7900(0) + 0.5664(0) + 3.0920(0)$$

=12.79 (predicted CO)