Sta 111 - Summer II 2017
Probability and Statistical Inference

21. Introduction to multiple linear regression

Lu Wang

Duke University, Department of Statistical Science

August 3, 2017

Outline

1. In MLR everything is conditional on all other variables in the model

2. Categorical predictors with k levels need $k - 1$ dummy variables

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

4. Avoid collinearity in MLR

5. R^2 vs adjusted R^2
    1. Adjusted R^2 applies a penalty for additional variables

6. Homework 6

- Simple linear regression: Bivariate - two variables: y and x

- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Data from the ACS

A random sample of 783 observations from the 2012 ACS.

1. `income`: Yearly income (wages and salaries) —> *response variable*
2. `employment`: Employment status, not in labor force, unemployed, or employed
3. `hrs_work`: Weekly hours worked
4. `race`: Race, White, Black, Asian, or other
5. `age`: Age
6. `gender`: 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

- All estimates in a MLR for a given variable are conditional on all other variables being held constant in the model.

- **Slope:**
  - Numerical x: *All else held constant*, for one unit increase in x_j, y is expected to be higher / lower on average by b_j units.

  - Categorical x: *All else held constant*, the predicted difference in y for the given level of x_j and the baseline is b_j.
- **Intercept:** With all the numerical x's set at 0 and all the categorical x's set at their corresponding baseline levels, y is expected on average to be b_0.

  - The intercept often does not make sense in context. It only serves to adjust the height of the line.

1. Interpret the slope for hrs_work.
2. Interpret the slope for gender.
   - Which gender is the baseline level?

	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

- It only takes $k - 1$ columns to code a categorical variable with k levels as 0/1s $\rightarrow$ *dummy variables*.
- Each categorical variable, with k levels, added to the model results in $k - 1$ parameters being estimated.

Citizen: yes / no ($k = 2$)
Baseline: no

Respondent	citizen:yes
1, Citizen	1
2, Not-citizen	0

Race: ($k = 4$)
Baseline: White

Respondent	race:black	race:asian	race:other
1, White	0	0	0
2, Black	1	0	0
3, Asian	0	1	0
4, Other	0	0	1

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 mor

than those born Apr thru Jun.

Inference for MLR

- Inference for the model as a whole: F-test, $df_1 = p$, $df_2 = n - p - 1$
  - Testing if the predictors *collectively* have an effect on the response variable

    $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
    $H_A:$ At least one of the $\beta_i \neq 0$
  - p is the number of predictors (slopes) in the model
  - Note the difference between regression *inputs* and *predictors*.
    - For example, we have 4 predictors for `race` and `age` (`raceblack`, `raceasian`, `raceother`, `age`), but just 2 inputs: `race` and `age`.
  - When did we use F-test before?

- Inference for each slope: T-test, $df = n - p - 1$
  - HT:

    $H_0: \beta_1 = 0$, when all other variables are included in the model
    $H_A: \beta_1 \neq 0$, when all other variables are included in the model
  - CI: $b_1 \pm T^{\star}_{df} SE_{b_1}$

Model output

```
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)
Multiple R-squared:  0.3126,^^IAdjusted R-squared:  0.2982
F-statistic: 21.77 on 16 and 766 DF,  p-value: < 2.2e-16
```

Practice

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

(a) True

(b) False

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

Significance also depends on what else is in the 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
```

```
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 <----
```

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 <u>independent</u> 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 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 in observational data, experiments are usually designed to control for correlated predictors.

Use ANOVA to compute R^2 in MLR

R^2 is the percent of variability in y that is explained by the model

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} = \frac{SST - SSE}{SST}$$

- sum of squares of y: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- sum of squares of residuals: $SSE = \sum_{i=1}^{n} e_i^2$
- *explained variability* $= SST - SSE$

When did we calculate sum of squares before?

Using *ANOVA* we can calculate the explained variability and total variability in y.

Use ANOVA to compute R^2 in MLR

```
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
```

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{1.8147e+12}{2.6399e+12} = 0.3126$$

Adjusted R^2

- R^2 increases when <u>any</u> variable is added to the model .
- 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

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of observations and p is the number of predictors (slopes) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Calculate adjusted R^2

```
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
```

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

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

(a) True

(b) False

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

(a) True

(b) False

Homework 6

Practice questions:

- Regression with multiple predictors: 8.1, 8.3
- Inference for MLR: 8.5
- Model selection: 8.7, 8.9, 8.11
- Model diagnostics: 8.13