Unit 7: Multiple linear regression

1. Introduction to multiple linear regression

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Abrahamsen

Slides posted at *https://stat.duke.edu/courses/Spring19/sta101.002*

Announcements

- Project questions?
- TEAMMATES Survey Due Tonight

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

- All estimates in a MLR for a given variable are conditional on all other variables being in the model.
- **Slope:**
  - Numerical x: *All else held constant*, for one unit increase in x_i, y is expected to be higher / lower on average by b_i units.
  - Categorical x: *All else held constant*, the predicted difference in y for the baseline and given levels of x_i is b_i.

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`: 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

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

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

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

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

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.

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

- Inference for the model as a whole: F-test, $df_1 = p$, $df_2 = n - k - 1$
  $H_0: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0$
  $H_A:$ At least one of the $\beta_i \neq 0$
- Inference for each slope: T-test, $df = n - k - 1$
  - 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
educcollege          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
```


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

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 <----
educcollege             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 <----
```

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

- When any 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

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

where n is the number of cases and k is the number of sloped estimated in the model.


```
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
```

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

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

(a) True
(b) False

(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.

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

- If the goal is to find the set of statistically predictors of $y \rightarrow$ use p-value selection.
- If the goal is to do better prediction of $y \rightarrow$ 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 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

Clicker question

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

	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

Given k predictors, there are 2^k possible models that can be fit. For small k, we can compare all possible models; however, when k is large fitting all possible models becomes computationally infeasible.

k	2^k
1	2
2	4
3	8
$\vdots$	$\vdots$
10	1024
$\vdots$	$\vdots$
20	1048576
$\vdots$	$\vdots$
100	1.267651e+30

There are several values, in addition to R^2_{adj}, that are commonly used for model selection. Like R^2_{adj}, they adjust the reduction in SSE (MSE) to account for the number of predictors in the model.

- Mallow's C_p:
$$C_p = \frac{1}{n}\left(SSE + 2ps^2\right),$$
- *AIC*:
$$AIC = \frac{1}{n\hat{\sigma}^2}\left(SSE + 2ps^2\right),$$
- *BIC*:
$$BIC = \frac{1}{n}\left(SSE + \log(n)ps^2\right),$$

where $s = \sqrt{MSE}$. Unlike R^2_{adj}, smaller values are better for C_p, *AIC* and *BIC*.

Step-wise model selection methods provide a computationally efficient alternative to trying to fit all possible models. For large k, step-wise methods fit a subset of models that is much smaller than the 2^k possible models.

Forward Step-wise Selection begins with a model consisting of no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In particular, at each step the variable that gives the greatest **additional** improvement to the fit of the model.

Forward Step-wise Selection Algorithm:

1. Let M_0 denote the *null* model, which contains no predictors

2 For $m = 0, 1, \ldots, k-1$:
   (a) Fit the $k - m$ models that augment the predictors in M_k with one additional predictor.
   (b) Choose the **best** among these $k - m$ models, and call it M_{m+1}. Here **best** is defined as having the smallest *SSE* or highest R^2.

3. Select a single best model from among $M_1, \ldots, M_p$, using R^2_{adj}, C_p, *AIC* or *BIC*.

Step 1: Fit all single variable models

- NOB:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0204410	79.98%	78.55%	70.51%

- Weight:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0451373	2.40%	0.00%	0.00%

- M:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0447214	4.19%	0.00%	0.00%

NOB results in the largest R^2, so we add that variable to the model.

Step 2: Fit all 2 variable models with NOB as one of the predictors

- NOB, Weight:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0104104	95.18%	94.44%	92.01%

- NOB, M:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0181633	85.32%	83.07%	74.61%

The model with NOB and Weight has the highest R^2, so we add Weight to the model.

Step 3: Fit all 3 variable models that include NOB and Weight.

- NOB, Weight, M:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0107174	95.28%	94.10%	91.08%

This is the only left to fit, so we are finished.

Summary:

Predictors	R^2_{adj}
NOB	78.55%
NOB, Weight	94.44%
NOB, Weight, M	94.10%

The model with NOB and Weight as predictors has the largest R^2_{adj}, and thus is the model we would select using forward selection.

Backward Step-wise Selection is another step-wise method, which is similar to forward selection. However, unlike forward step-wise selection, it begins will the full least squares model containing all k predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Backwards Step-wise Selection Algorithm:

1. Let M_k, denote the full model, which contains all p predictors.
2. For $m = k, k-1, \ldots, 1$:
   (a) Consider all the m models that contain all but one of the predictors in M_m, for a total of $m-1$ predictors.
   (b) Choose the **best** among these m models, and call it M_{m-1}. Here **best** is defined as having the smallest SSE or highest R^2
3. Select a single best model among $M_0, \ldots, M_m$ using R^2_{adj}, C_p, AIC or BIC.

Step 1: Fit the model with all predictors

- NOB, Weight, M:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0107174	95.28%	94.10%	91.08%

Step 2: Fit each model by removing removing one predictor

- NOB, Weight:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0104104	95.18%	94.44%	92.01%

- NOB, M:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0181633	85.32%	83.07%	74.61%

- Weight, M

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0470333	10.39%	0.00%	0.00%

The model with NOB and Weight has the largest R^2 so we keep that model.

Step 3: Fit two models by removing each predictor from the NOB, Weight model

- NOB:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0204410	79.98%	78.55%	70.51%

- Weight:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0451373	2.40%	0.00%	0.00%

NOB has the largest R^2 so that is the model we keep. We are down to the single predictor model, so we are done.

Summary:

Predictors	R^2_{adj}
NOB, Weight, M	94.10%
NOB, Weight	94.44%
NOB	78.55%

NOB, Weight has the largest R^2_{adj}, thus it is the model chose by backward selection.

Remarks:

- Best subset (fitting all models), forward step-wise, and backward step-wise selection approaches generally give similar but not identical models.
- As another alternative, hybrid version of forward and backward stepwise selection are available, in which variables are added to the model sequentially, similar to forward selection.

  However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

  Hybrid methods more closely mimic best subset selection while retaining the computational advantages of forward and backward stepwise selection.

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
- Also important to make sure that your explanatory variables are not *collinear*

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

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