Sta 111 - Summer II 2017
Probability and Statistical Inference

22. Model selection and conditions for MLR

Lu Wang

Duke University, Department of Statistical Science

August 7, 2017

Outline

1. Model selection

1. Model selection criterion depends on goal: significance vs. prediction
2. Backward-elimination
3. Forward-selection

2. Conditions for MLR

1. Checking model conditions using graphs

Professor rating

Data: Student evaluations of instructors' teaching quality for 463 courses at the University of Texas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty [1]	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [2]	0.1511	0.0749	2.02	0.04
lower.yes [3]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [4]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [5]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

[1] `beauty`: the beauty judgements were made by six students who had not attended the classes and were not aware of the course evaluations.

[2] `formal`: picture wearing tie&jacket/blouse, levels: `yes`, `no`

[3] `lower`: lower division course, levels: `yes`, `no`

[4] `students`: number of students

[5] `tenure`: tenure status, levels: `non-tenure track`, `tenure track`, `tenured`

Hypotheses

Just as the interpretation of the slope parameters take into account all other variables in the model, the hypotheses for testing for significance of a predictor also takes into account all other variables.

> $H_0 : \beta_i = 0$ when other explanatory variables are included in the model.
>
> $H_A : \beta_i \neq 0$ when other explanatory variables are included in the model.

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) If we keep all other variables in the model, there is strong evidence that professor's age is associated with their rating.

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.

Based on what we've learned so far, what are some ways you can think of that can be used to determine which variables to keep in the model and which to leave out?

- If the goal is to find the set of statistically significant 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.

Backward-elimination

1. R^2_{adj} approach:
   - Start with the full model
   - Drop one variable at a time and record R^2_{adj} of each smaller model
   - Pick the model with the highest increase in R^2_{adj}
   - Repeat until none of the models yield an increase in R^2_{adj}

2. p-value approach:
   - 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

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
best model —>	*beauty + gender + age + formal + native + tenure*	*0.0855*
	beauty + gender + age + formal + native + students	0.0713
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806
	beauty + gender + age + formal + tenure	0.0773
	beauty + gender + age + formal + native	0.0713

Selected model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6284	0.1673	27.66	0.00
beauty	0.1055	0.0328	3.21	0.00
gender.male	0.2081	0.0519	4.01	0.00
age	-0.0088	0.0032	-2.75	0.01
formal.yes	0.1324	0.0714	1.85	0.06
native:non english	-0.2430	0.1080	-2.25	0.02
tenure:tenure track	-0.2068	0.0839	-2.46	0.01
tenure:tenured	-0.1760	0.0641	-2.74	0.01

Backward-elimination: $p - value$ approach

Step	**Variables included & p-value**									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenu
		male		yes	yes	non english	yes		tenure track	tenur
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenu
		male		yes	yes	non english			tenure track	tenur
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.
Step 2	beauty	gender	age	formal		native		students	tenure	tenu
		male		yes		non english			tenure track	tenur
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.
Step 3	beauty	gender	age	formal		native			tenure	tenu
		male		yes		non english			tenure track	tenur
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.
Step 4	beauty	gender	age			native			tenure	tenu
		male				non english			tenure track	tenur
	0.00	0.00	0.01			*0.06*			0.01	0.
Step 5	beauty	gender	age						tenure	tenu
		male							tenure track	tenur
	0.00	0.00	0.01						0.01	0.

Best model: beauty + gender + age + tenure

Forward-selection

1. R^2_{adj} approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the model with the highest R^2_{adj}
   - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   - Repeat until the addition of any of the remaining variables does not result in a higher R^2_{adj}

2. $p - value$ approach:
   - Start with 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 does not have a significant p-value

In forward-selection the p-value approach isn't any simpler (you still need to fit a bunch of models), so there's almost no incentive to use it.

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

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

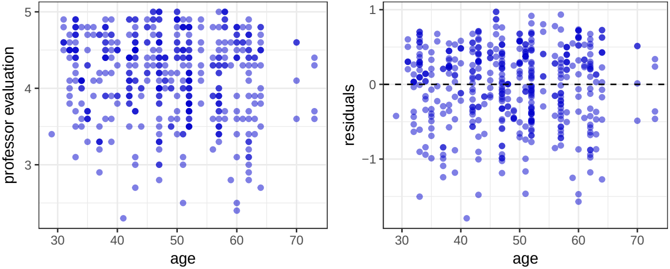
Important regardless of doing inference

- *Linearity* $\rightarrow$ each variable is linearly related to the outcome

Important for doing inference

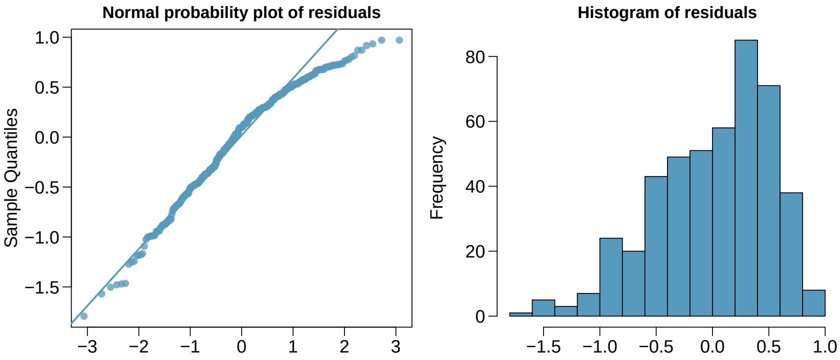
- *Nearly normally distributed residuals* $\rightarrow$ primary concern relates to residuals that are outliers
- *Constant variability of residuals* (*homoscedasticity*)
- *Independence of observations (and hence residuals)*
- Also important to make sure that your explanatory variables are *not collinear*

- For categorical variable, using boxplot of the residuals against each level to check whether variability fluctuates across levels.
- Using scatterplot of residuals vs. each numerical predictor to check if there is some possible structure such as curvature in the residuals.



Does this condition appear to be satisfied?

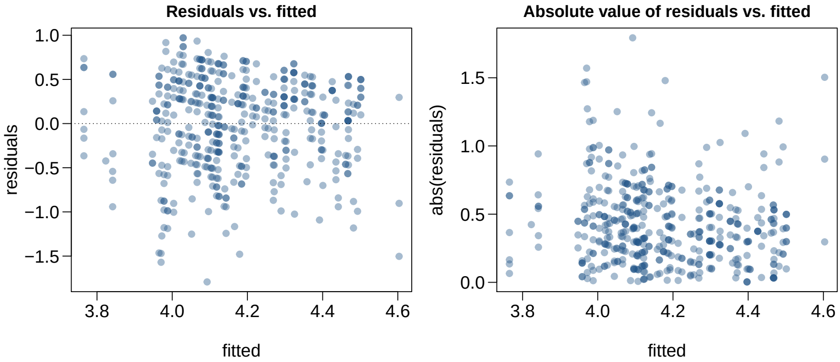
Q-Q plot and/or histogram of residuals:



Does this condition appear to be satisfied?

(3) constant variability in residuals

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):



Does this condition appear to be satisfied?

Checking constant variance - recap

- When we did simple linear regression (one predictor) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ predictors) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

scatterplot of residuals vs. order of data collection:

Does this condition appear to be satisfied?

- Checking for independent residuals allows us to indirectly check for independent observations.

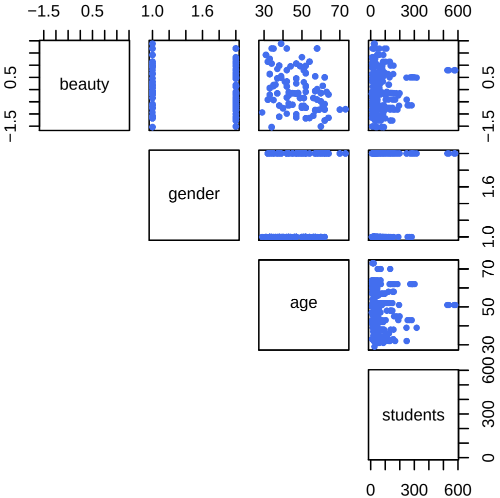
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.

- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(5) Checking collinearity among predictors

Use pairwise correlations to check collinearity.

```
pairs(~beauty+gender+age+students, pch = 19,
      col = "royalblue", lower.panel = NULL)
```



Practice

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) Q-Q 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.