Goodness-of-Fit and Model Selection

November 13, 2019

Example

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	26.50	1.773	14.95	$<$ 2e-16
trt	-48.20	0.159	-303.71	$<$ 2e-16
sexM	-0.63	0.159	-3.99	7.23e-05
age	0.02	0.012	1.29	0.197
weight	-0.01	0.003	-2.45	0.015
sys.bp	-0.15	0.006	-27.44	$<$ 2e-16
dia.bp	-0.11	0.010	-10.84	$<$ 2e-16
incomelow	-0.23	0.227	-1.01	0.314
incomemed	-0.05	0.238	-0.19	0.848
ldl.pre	0.99	0.008	126.16	$<$ 2e-16

Write out the regression model.

Example

- Interpret the coefficients corresponding to `sexM` and `age`.
- Calculate the residual for the first patient.

	ldl.post	trt	sex	age	weight	sys.bp	dia.bp	income	ldl.pre
1	176	0	M	47	186	119	64	low	178

Example

The estimated linear regression line for $\texttt{LDL.post} = \beta_0 + \beta_1 \texttt{trt}$ is

$$\hat{\texttt{LDL.post}} = 175.0309 - 49.0756 \times \texttt{trt}.$$

with $SE(b_1) = 0.6657$.

Why is this different from the estimate and standard error for `trt` in the multiple regression model?

Correlation Between Predictor Variables

- We say the two predictor variables are **collinear** when they are correlated.
- This complicates model estimation.
- We can't always prevent collinearity, but we do want to control it.
- Ex: height and arm span give us essentially the same information. We wouldn't want to use both in a model.

Goodness-of-Fit

Recall that we used R^2 to determine the amount of variability explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{total variability}} = 1 - \frac{SS_{residuals}}{SS_{total}}$$

We can continue to use this in multiple regression.

Goodness-of-Fit

- R^2 will always increase when we include more variables in the model.
- This is true even if the variables aren't very useful!
- We want a measure that will help us balance model efficacy with model size.

Adjusted R^2

Adjusted R^2 is computed as

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

where n is the number of observations and k is the number of predictor variables in the model.

Note that k includes the $p-1$ predictor variables for categorical variables with p levels.

Adjusted R^2

Notice that

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

and

$$R^2 = 1 - \frac{SS_{resid}}{SS_{total}}$$

Since $k \geq 1$, $R^2_{adj} < R^2$.

Adjusted R^2

- The idea here lies with degrees of freedom.
- We adjust R^2 based on model and error df.
- This balances efficacy and model size (what we wanted).
- This will also help us compare models.

Model Selection

- We want models to balance efficacy and size.
- In multiple linear regression, **model selection** refers to "pruning" variables that are less important.
- Models that have been optimized in this way are referred to as **parsimonious**.
  - (Think parsimonious = "frugal".)
- The model that includes all possible variables is called the **full model**.

Identifying Unhelpful Variables

The full model:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	26.50	1.773	14.95	$<$ 2e-16
trt	-48.20	0.159	-303.71	$<$ 2e-16
sexM	-0.63	0.159	-3.99	7.23e-05
age	0.02	0.012	1.29	0.197
weight	-0.01	0.003	-2.45	0.015
sys.bp	-0.15	0.006	-27.44	$<$ 2e-16
dia.bp	-0.11	0.010	-10.84	$<$ 2e-16
incomelow	-0.23	0.227	-1.01	0.314
incomemed	-0.05	0.238	-0.19	0.848
ldl.pre	0.99	0.008	126.16	$<$ 2e-16

Multiple R-squared: 0.9913, Adjusted R-squared: 0.9912

Identifying Unhelpful Variables

Removing `income`:

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	26.41	1.767	14.95	$<$ 2e-16
trt	-48.20	0.159	-303.84	$<$ 2e-16
sexM	-0.62	0.159	-3.91	9.82e-05
age	0.02	0.012	1.29	0.196
weight	-0.01	0.003	-2.49	0.013
sys.bp	-0.15	0.006	-27.62	$<$ 2e-16
dia.bp	-0.10	0.010	-10.83	$<$ 2e-16
ldl.pre	0.99	0.008	126.21	$<$ 2e-16

Multiple R-squared: 0.9913, Adjusted R-squared: 0.9912

Identifying Unhelpful Variables

- We find that the models have the same R^2_{adj}!
- Which one should we choose?
- Should we remove more variables?

Identifying Unhelpful Variables

What if we remove `trt`?

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-22.61	17.060	-1.33	0.185
sexM	-0.29	1.537	-0.19	0.853
age	0.05	0.113	0.47	0.636
weight	0.002	0.026	0.07	0.947
sys.bp	-0.15	0.054	-2.79	0.005
dia.bp	-0.08	0.094	-0.86	0.389
ldl.pre	1.10	0.076	14.46	$<$ 2e-16

Multiple R-squared: 0.1783, Adjusted R-squared: 0.1733

Model Selection Strategies

We will discuss two common model selection approaches:

1. Forward Selection
2. Backward Elimination

These are referred to as **step-wise** model selection.

Backward Elimination

- **Backward elimination** starts with the full model.
- Variables are removed one-at-a-time until R^2_{adj} stops improving.
- At each step, we want to remove the least useful variable.

Example

Consider a data set on various loans. We want to predict the interest rate. The available variables are

- `interest_rate`: loan interest rate
- `income_var`: whether income source & amount verified. Takes values verified, source only, and not.
- `debt_to_income`: ratio of debt to income
- `credit_util`: proportion of credit being utilized
- `bankruptcy`: whether borrower has previous bankruptcy
- `term`: length of loan (months)
- `issued`: month and year loan issued
- `credit_checks`: number of credit checks in last 12 months

Example: Backward Selection

The full model is

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001

There are $n = 10000$ cases in this data set. The variance of the residuals is 18.53, and the variance of the total price is 25.01. Calculate R^2 and R^2_{adj}.

Example: Backward Selection

Can we drop a variable and improve R^2_{adj}?

- We got a baseline R^2_{adj} on the previous slide.
- Variables are eliminated one-at-a-time from the full model.
- R^2_{adj} is checked each time.
- We move forward with the model with the highest R^2_{adj}

Exclude ...				
	`income_ver`	`debt_to_income`	`credit_util`	`bankruptcy`
	$R^2_{adj} = 0.22380$	$R^2_{adj} = 0.25468$	$R^2_{adj} = 0.19063$	$R^2_{adj} = 0.25787$
	`term`	`issued`	`credit_checks`	
	$R^2_{adj} = 0.14581$	$R^2_{adj} = 0.25854$	$R^2_{adj} = 0.24689$	

Example: Backward Selection

After checking the R^2_{adj} for each potential variable removal, we remove `issued`.

- Now we repeat the process with the model with `issued` removed.
- Our new baseline R^2_{adj} is 0.25854.

Exclude `issued` and ...			
	`income_ver` $R^2_{adj} = 0.22395$	`debt_to_income` $R^2_{adj} = 0.25479$	`credit_util` $R^2_{adj} = 0.19074$
	`bankruptcy` $R^2_{adj} = 0.25798$	`term` $R^2_{adj} = 0.14592$	`credit_checks` $R^2_{adj} = 0.24701$

Example: Backward Selection

So our final model is

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9213	0.1982	9.69	<0.0001
income_ver: *source_only*	0.9740	0.0991	9.83	<0.0001
income_ver: *verified*	2.5355	0.1172	21.64	<0.0001
debt_to_income	0.0211	0.0029	7.19	<0.0001
credit_util	4.8958	0.1619	30.25	<0.0001
bankruptcy	0.3869	0.1324	2.92	0.0035
term	0.1537	0.0039	38.97	<0.0001
credit_checks	0.2283	0.0182	12.51	<0.0001

Write the regression model for these results.