

Lecture 27: Model Selection + Multiple Regression Conditions

Chapter 8.2-8.3

1 / 22

Question for Today

Recall the Mario Kart analysis

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.34	1.71	24.15	< 2e-16 ***
condused	-5.13	1.05	-4.88	2.91e-06 ***
stockPhotoyes	1.08	1.06	1.02	0.308
duration	-0.03	0.19	-0.14	0.888
wheels	7.29	0.55	13.13	< 2e-16 ***

Residual standard error: 4.901 on 136 degrees of freedom

Multiple R-squared: 0.719, Adjusted R-squared: 0.7108

2 / 22

Question for Today

This is the **full model**: every explanatory variable provided is included.

Recall Occam's Razor: **all other things being equal, simpler is better**.

In our case: simpler = less predictor variables included in the model.

The act of choosing which predictor variables to include in your model is **model selection**.

3 / 22

Two Common Strategies

There are two **stepwise regression** methods that add/subtract one variable at a time:

- ▶ Backward Elimination
- ▶ Forward Selection

The criteria used will be p -values.

4 / 22

Backward Elimination

1. Start with the **full model**
2. While there still exists statistically non-significant variables
 - 2.1 Identify the variable with the largest p-value and drop it
 - 2.2 Refit the model
3. Report model once there are no more non-significant variables

5 / 22

Backward Elimination

Starting here:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.3415	1.7117	24.15	0.0000
cond_used	-5.1306	1.0511	-4.88	0.0000
stockPhotoyes	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000

6 / 22

Backward Elimination

Drop duration.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.3415	1.7117	24.15	0.0000
cond_used	-5.1306	1.0511	-4.88	0.0000
stockPhotoyes	1.0803	1.0568	1.02	0.3085
duration	-0.0268	0.1904	-0.14	0.8882
wheels	7.2852	0.5547	13.13	0.0000

7 / 22

Backward Elimination

Drop stockPhotoyes.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.2245	1.4911	27.65	0.0000
cond_used	-5.1763	0.9961	-5.20	0.0000
stockPhotoyes	1.1177	1.0192	1.10	0.2747
wheels	7.2984	0.5448	13.40	0.0000

8 / 22

Backward Elimination

Done.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	42.3698	1.0651	39.78	0.0000
cond_used	-5.5848	0.9245	-6.04	0.0000
wheels	7.2328	0.5419	13.35	0.0000

9 / 22

Forward Selection

1. Start with the model with no variables
2. Fit all models with one possible additional variable
3. Add the additional variable with the smallest p-value if its significant
4. Repeat steps 2 and 3 until there are no significant additional variables.

10 / 22

Criticisms of the Techniques

Critics regard stepwise regression as **data dredging**, where intense computation is used as a substitute for subject area expertise when deciding on a model.

Data mining involves automatically testing huge numbers of hypotheses about a single data set by exhaustively searching for combinations of variables that might show a correlation.

11 / 22

Criticisms of the Techniques



12 / 22

Assumptions of Multiple Regression

- ▶ The residuals e_i of the model
 - ▶ are nearly normal
 - ▶ have nearly constant variance
 - ▶ are independent
- ▶ Each variable is linearly related to the outcome
- ▶ No pattern in residuals relative to dependent variables.

13 / 22

Example Model

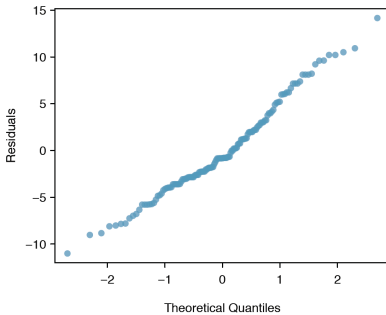
We investigate plots for the following model:

$$\widehat{\text{price}} = b_0 + b_1 \times \text{cond_new} + b_2 \times \text{wheels}$$

- ▶ Normal probability plot of residuals
- ▶ Absolute values of residuals against fitted values: look for non-constant variance
- ▶ Residuals against each predictor variable

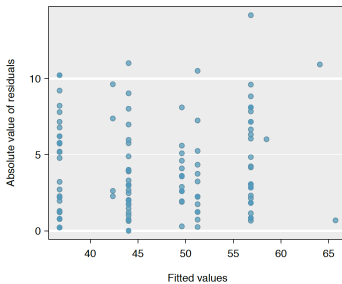
14 / 22

Normal Probability Plot of Residuals



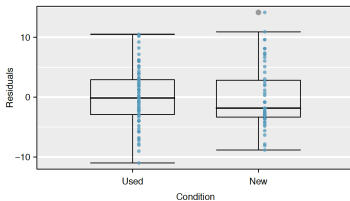
15 / 22

Absolute Values of Residuals Against Fitted Values



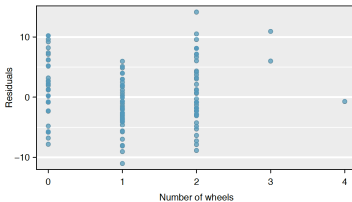
16 / 22

Residuals Against Each Predictor Variable: Condition



17 / 22

Residuals Against Each Predictor Variable: Wheels



18 / 22

George E.P. Box

There was a famous statistician named Box



famous for the Box/Cox Transformation.

George E.P. Box's Famous Quote

"All models are wrong, but some are useful."

Caution

We can tolerate a little leeway with model assumptions, but when they are grossly violated we have to be skeptical of any confidence intervals/ p -values. If model assumptions are clearly violated

- ▶ consider a new model
- ▶ get the assistance of someone who can help

21 / 22

Next Time

What if the outcome variable is not numerical, but rather a **yes/no** response variable?

- ▶ Was an email spam or not?
- ▶ Will someone develop cancer or not?
- ▶ Is a person female?

We use **logistic regression**.

22 / 22