Lecture 27: Model Selection + Multiple Regression Assumption Verification

Chapter 8.2-8.3

Question for Today

Recall the Mario Kart analysis

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.34153    1.71167  24.153  < 2e-16 ***
condused      -5.13056    1.05112  -4.881 2.91e-06 ***
stockPhotoyes  1.08031    1.05682   1.022    0.308
duration      -0.02681    0.19041  -0.141    0.888
wheels         7.28518    0.55469  13.134  < 2e-16 ***
---

Residual standard error: 4.901 on 136 degrees of freedom
Multiple R-squared:  0.719, Adjusted R-squared:  0.7108
```

Question for Today

This was the full model: we included every explanatory variable provided.

Recall the principle inspired by Occam's Razor: all other things being equal, simpler is better. In our case: less predictor variables included in the model!

Is there a systematic (or should I say, less unsystematic) way to pick which predictor variables to include?

Via model selection techniques.

Two Common Strategies

The following are two common stepwise regression methods because they add/subtract one variable at a time:

- Backward Elimination
- Forward Selection

We will discuss this in terms of a p-value approach. We can also use R^2_{adj} as a criterion.

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

Backward Elimination

	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

Drop `duration`.

Backward Elimination

	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

Drop `stockPhotoyes`.

Backward Elimination

	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

Done.

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.

Criticisms of the Techniques

Data dredging is the use of data mining to uncover relationships in data.

Critics regard stepwise regression as a paradigmatic example of data dredging, intense computation often being an inadequate substitute for subject area expertise.

The process of 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. Think of multiple testing issues!

Criticisms of the Techniques

DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
scottadams@aol.com
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 1999 United Feature Syndicate, Inc.
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

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

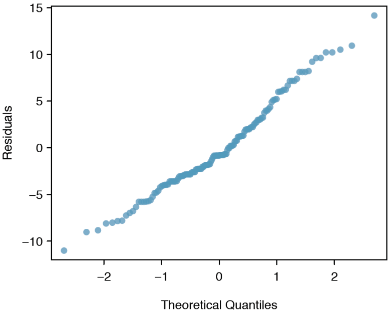
Example Model

We investigate plots for the following model:

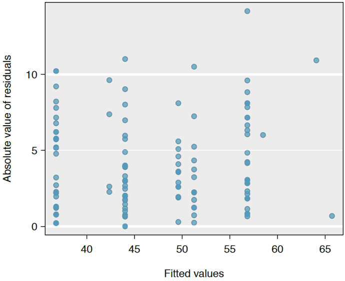
$$\widehat{\texttt{price}} = b_0 + b_1 \times \texttt{cond\_new} + b_2 \times \texttt{wheels}$$

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

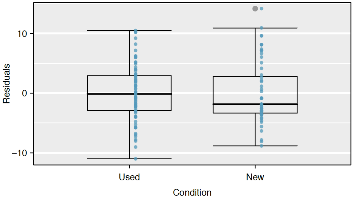
Normal Probability Plot of Residuals



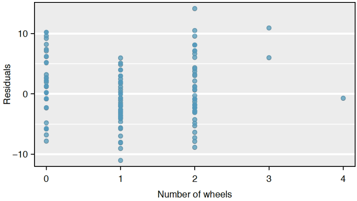
Absolute Values of Residuals Against Fitted Values



Residuals Against Each Predictor Variable: Condition



Residuals Against Each Predictor Variable: Wheels



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

That being said, while we can tolerate a little leeway with model assumptions, don't report results when the assumptions are grossly violated. If model assumptions are clearly violated

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

Next Time

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

- Was an email spam or not?
- Will someone develop cancer or not?
- Will a car pass an emission test?

We use logistic regression.