Model selection I

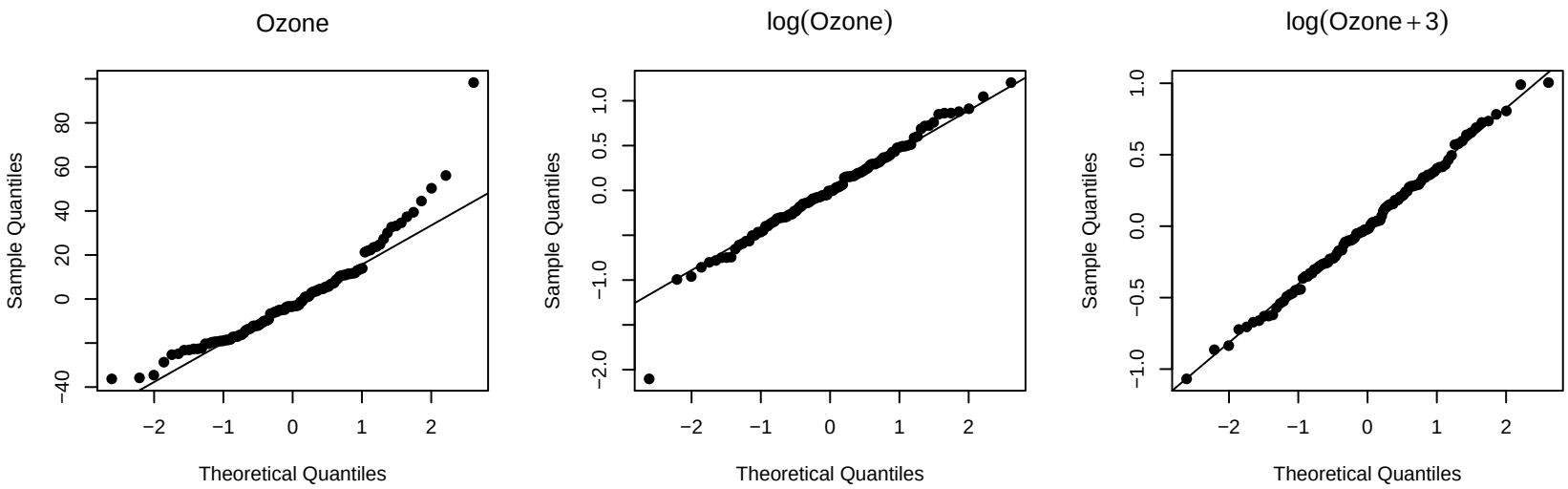
Patrick Breheny

February 17

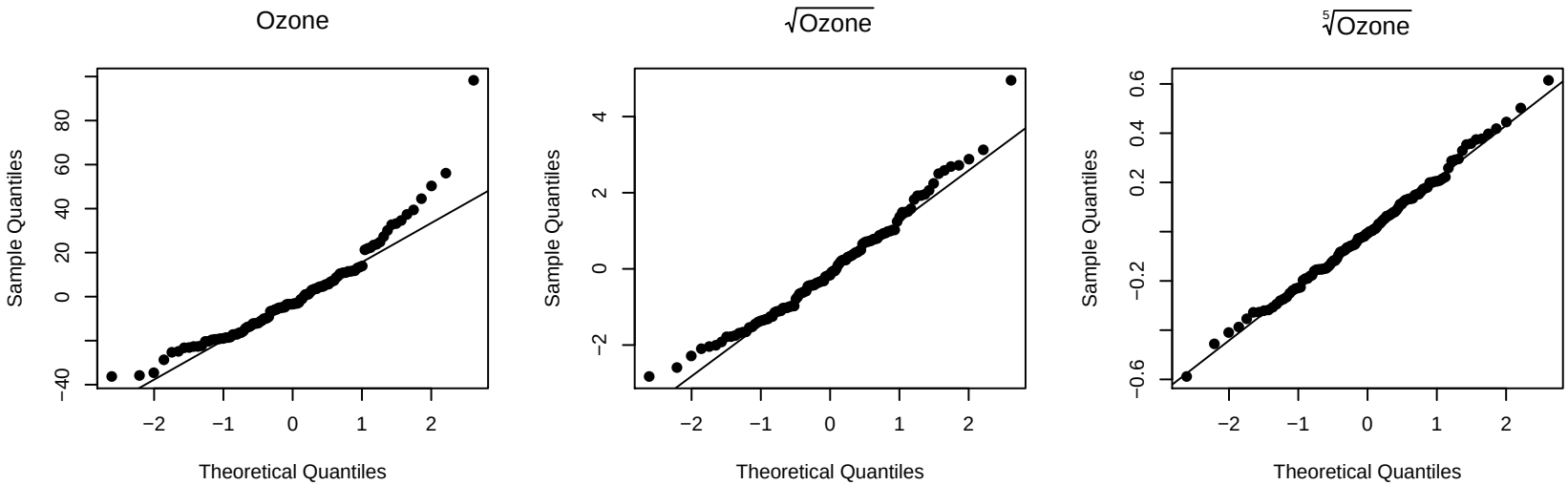
Remedial measures

- Suppose one of your diagnostic plots indicates a problem with the model's fit or assumptions; what options are available to you?
- Generally speaking, you have three avenues down which you could proceed:
  - Transforming the outcome variable
  - Using a different method
  - Fitting a more flexible model

Example: Ozone model



Example: Ozone model again



This approach is known as the *Box-Cox procedure*, after two statisticians who identified an automatic method for identifying the optimal normalizing exponent to which y should be raised

Central limit theorem for regression (informal)

- You may be wondering: how important is it if y does not follow a normal distribution?
- For simpler methods like the t-test, the central limit theorem guarantees that we have approximate normality regardless of the distribution of y; is there a similar result for regression
- There is, but it requires that the diagonal elements of $\mathbf{H}$ are small – *i.e.*, that no one point "takes over" the regression fit
- **Central limit theorem for regression (informal)**: If n is reasonably large and none of the values $\{H_{ii}\}$ are too large, then $(\hat{\beta} - \beta)/\widehat{\mathrm{SE}}$ follows an approximately normal distribution

Ozone example: Effect of transformation on p-values

- In the ozone example, what was the effect of transformation on hypothesis testing:

	Ozone	$\log(\text{Ozone} + 3)$	$\sqrt[5]{\text{Ozone}}$
Solar	.03	.0001	.0002
Wind	1×10^{-6}	2×10^{-5}	1×10^{-5}
Temp	1×10^{-9}	7×10^{-13}	7×10^{-13}
Day	.1	.2	.2

- Generally speaking, transformations which normalize the outcome result in more powerful tests and smaller standard errors (although not always for all coefficients)

SE and flexible models

- Now to the more complicated issue of fitting more flexible and complicated models to the data (but still a linear regression model)
- Let's compare the standard errors for the following two models fit to the alcohol metabolism data:

Coefficient	SE
Male	0.55
Alcoholic	0.60
Gastric	0.29

Coefficient	SE
Male	1.33
Alcoholic	3.94
Gastric	0.52
Male·Alcoholic	4.39
Male·Gastric	0.62
Alcoholic·Gastric	2.81
Male·Alcoholic·Gastric	3.00

The bias-variance tradeoff

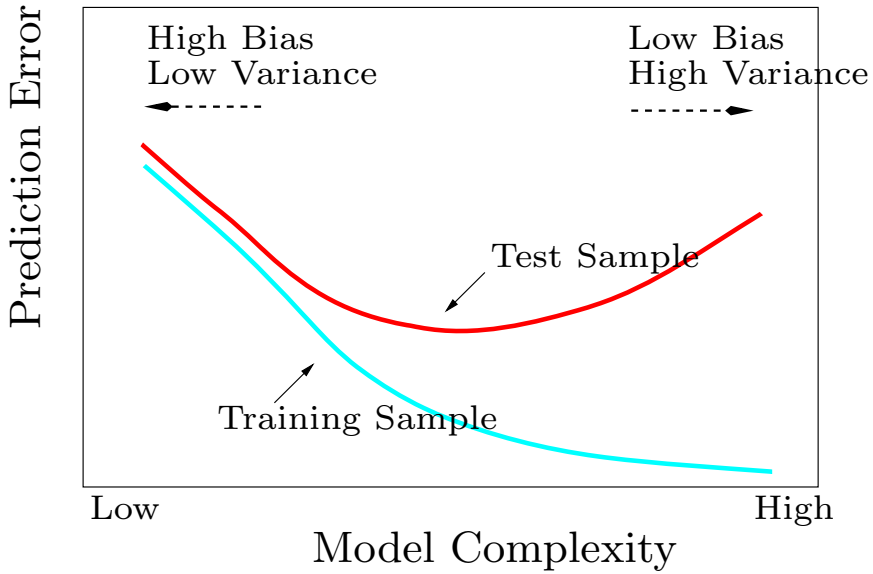
- This example illustrates what is perhaps the central concept in statistical modeling: the *bias-variance tradeoff*
- As we fit more flexible and complicated models with larger numbers of parameters and adjust for ever-larger numbers of confounders, bias becomes less of an issue
- However, the variances of our estimates become very large

Overfitting

- This is a fundamental idea in the notion of statistical *inference* – just because you can describe the sample well does not mean that this description can be generalized to the population
- As we have seen, RSS always goes down as you add more parameters to a model
- But this does not mean that more complicated models are more successful at predicting outcomes that lie *outside* our sample
- This phenomenon is referred to as *overfitting*; a model that describes the sample very well, but generalizes poorly, is said to be *overfit*

Bias-variance tradeoff – illustration

An illustration of this phenomenon, courtesy of *The Elements of Statistical Learning*, by Hastie, Tibshirani, and Friedman:



Here *training sample* refers to the data used to fit the model, and *test sample* on an external sample used to test the accuracy of the model

Parsimony

- A related notion is that of *parsimony*: given two models that explain the outcome roughly equally well, the simpler model is better (this is also referred to as *Occam's razor*)
- Statistically speaking, simpler models with fewer variables are desirable because they lead to lower variance and are easier to interpret
- Either way, the take-home message is the same: overly simple and overly complex models are both bad (for different reasons), and the best model usually lies somewhere in the middle between these two extremes
- In the words of Einstein, "Everything should be made as simple as possible, but no simpler"

Model selection criteria

- So how do we find this "sweet spot" in the middle?
- The most common approach is to use some sort of *model-selection criterion* which provides a measure of the overall quality of a model
- To be useful, such a criterion must punish models that are overly simple, as well as enforce parsimony and punish models that are overly complex
- The idea is that we can fit a number of different models, and then compare them in terms of some criterion to identify a model or models that seem to appropriately balance bias and variance

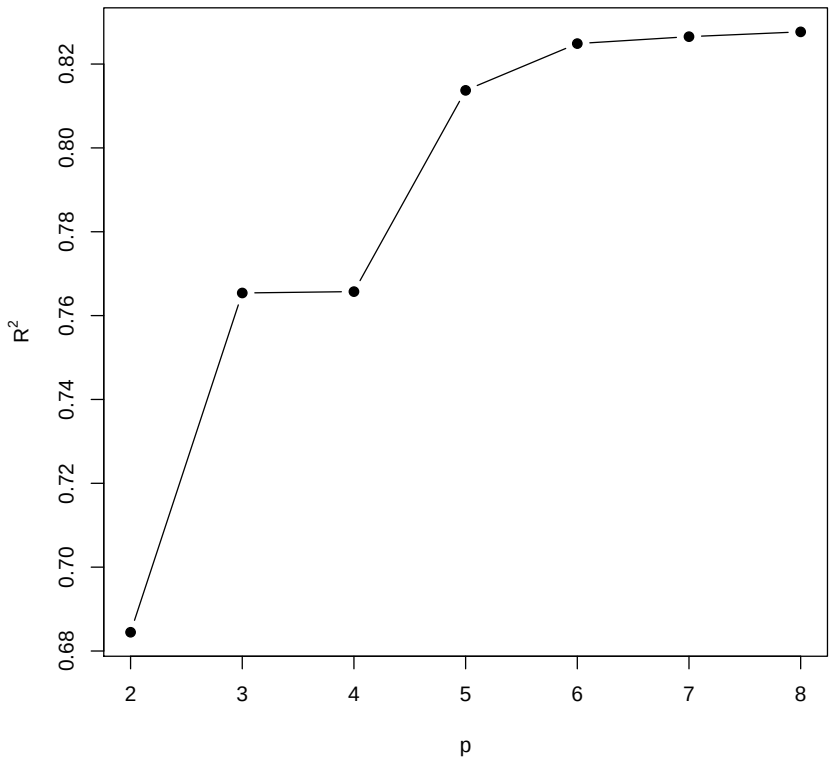
Sequence of models

In what follows, we will consider the following nested sequence of ever more complex models for the alcohol data set:

- 1: Intercept only
- 2: Add `Gastric`
- 3: Add `Male`
- 4: Add `Alcoholic`
- 5: Add `MaleGastric`
- 6: Add `AlcoholicGastric`
- 7: Add `MaleAlcoholic`
- 8: Add `MaleAlcoholicGastric`

R^2

We know that R^2 is not a good model selection criterion – it will always choose the most complex model and therefore drive us towards overfitting:



Adjusted R^2

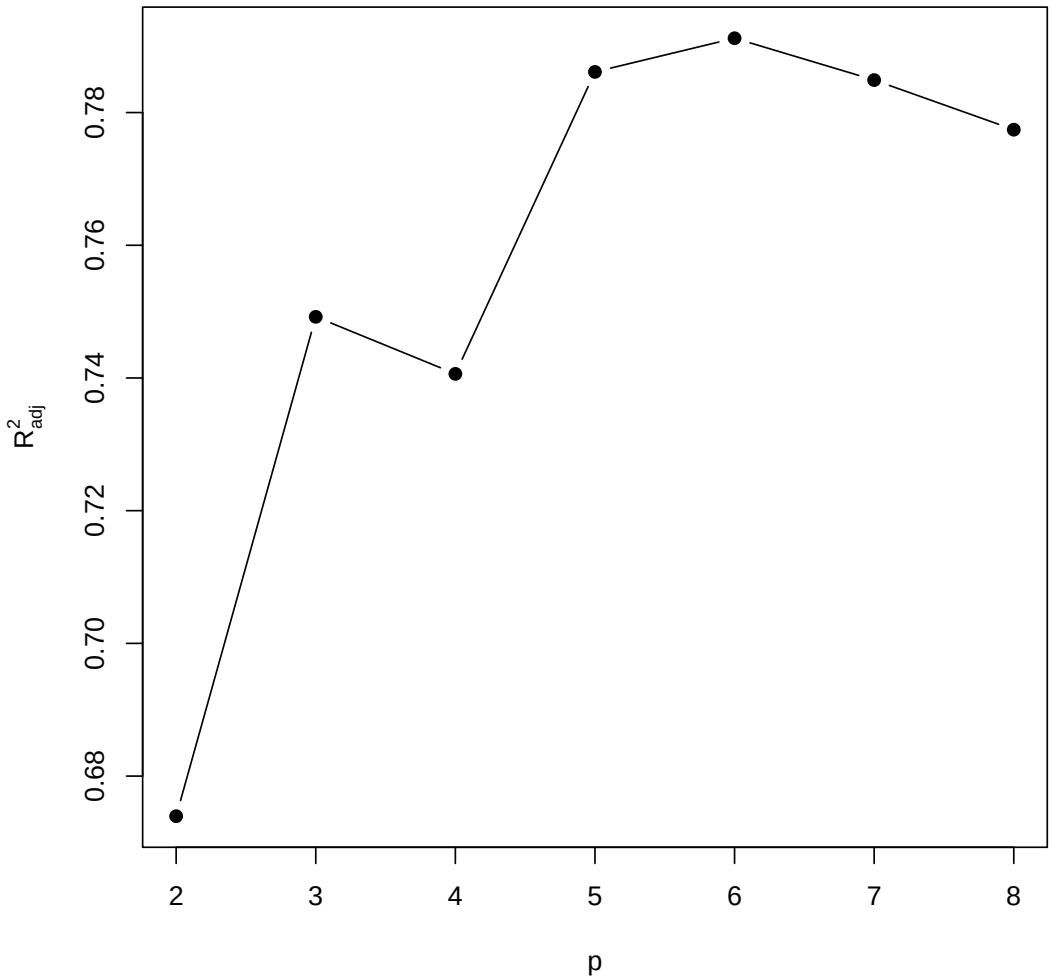
- One approach is to simply adjust R^2 by dividing RSS and TSS by their degrees of freedom:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/(n-p)}{\mathrm{TSS}/(n-1)}$$

- This criterion is called the *adjusted* R^2
- If a more complex model does not fit any better than you would expect by random chance, then its $\mathrm{R}^2_{\mathrm{adj}}$ will not be any higher than the simpler model

Adjusted R^2: Illustration



Mean squared error

- However, although $\mathrm{R}^2_{\mathrm{adj}}$ does not reward overfitting, it doesn't really penalize it either
- More sophisticated approaches attempt to directly estimate quantities which measure both the bias and variance of a model
- The *total mean squared error* of a model's fit is defined as the expected value of

$$\sum_i (\hat{\mu}_i - \mu_i)^2$$

Estimating total mean squared error

- By a similar calculation to how we decomposed the total sum of squares, it can be shown that

$$\begin{aligned} \mathrm{E}\sum_i (\hat{\mu}_i - \mu_i)^2 &= \sum_i \left\{ (\mathrm{E}\hat{\mu}_i - \mu_i)^2 + \mathrm{Var}(\hat{\mu}_i) \right\} \\ &= \mathrm{BSS} + p\sigma^2, \end{aligned}$$

  where BSS stands for "bias sum of squares"
- It can also be shown that if the $\{\hat{\mu}_i\}$ are not unbiased,

$$\mathrm{E}(\mathrm{RSS}) = \mathrm{BSS} + (n-p)\sigma^2$$

- Thus, a reasonable estimator of the total mean squared error is

$$\mathrm{RSS} - (n-p)\sigma^2 + p\sigma^2 = \mathrm{RSS} - n\sigma^2 + 2p\sigma^2$$

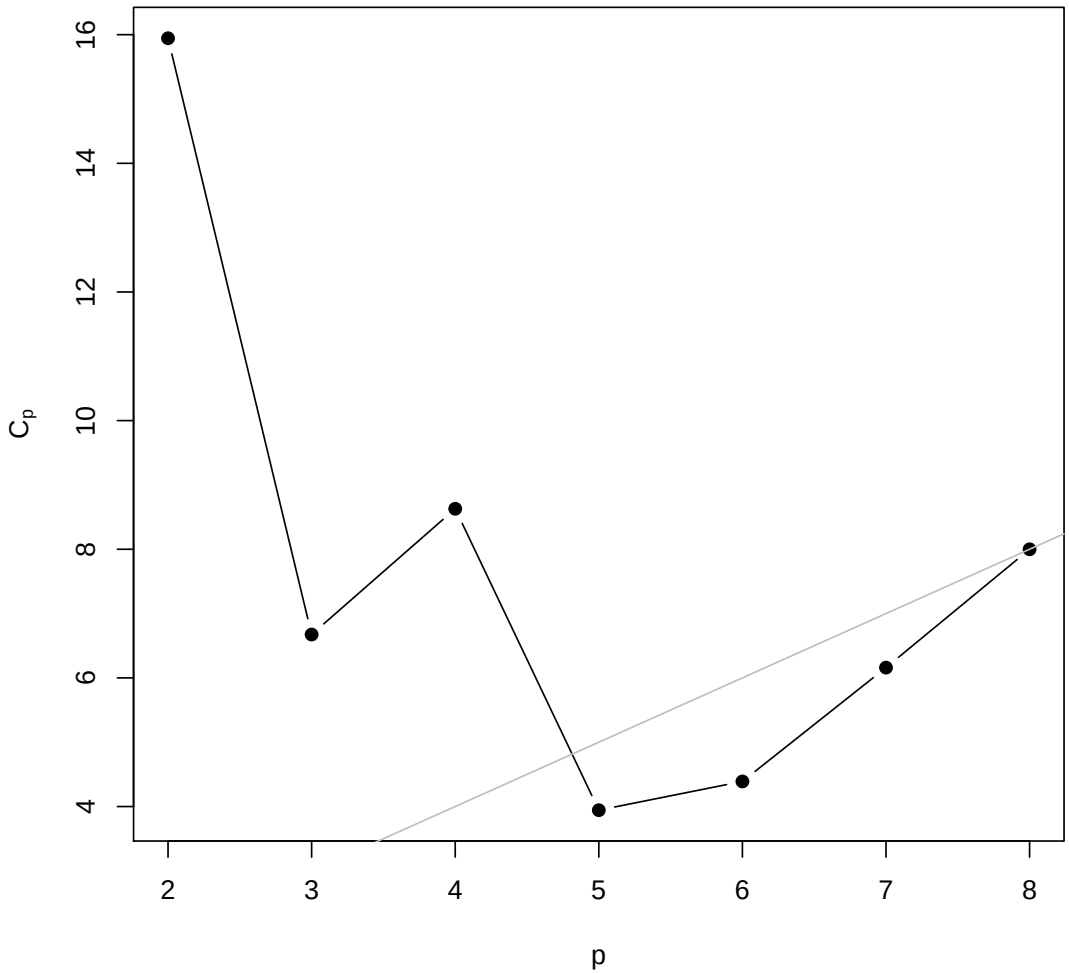
Mallows' C_p

- This idea was originally proposed by Mallows, who divided both sides by σ^2 to obtain

$$C_p = \frac{\text{RSS}}{\sigma^2} - n + 2p,$$

  which is referred to as *Mallows' C_p*
- Obviously, to use this criterion, we need an estimate for σ^2
- Customary practice is to use the largest model under consideration to estimate the error variance, and then use this $\hat{\sigma}^2$ to calculate C_p for all the models
- Note that if the model has no bias, then $C_p \approx p$

C_p: Illustration



Expected prediction error

- A related concept is the *expected prediction error* of a model, defined as the expected value of

$$\sum_i (Y_i - \hat{\mu}_i)^2$$

- Note that in the above, we are drawing two separate sets of y's (*i.e.*, the expectation is a double expectation):
  - One set is used to fit the model
  - The other set is used to evaluate the fit
  - The two sets have the same $\{\mathbf{x}_i\}$ values, however

Estimating prediction error

- Using the same sort of decomposition as before,

$$\begin{aligned} \mathrm{E}\sum_i (Y_i - \hat{\mu}_i)^2 &= \sum_i \left\{ \mathrm{Var}(Y_i) + (\mathrm{E}\hat{\mu}_i - \mu_i)^2 + \mathrm{Var}(\hat{\mu}_i) \right\} \\ &= n\sigma^2 + \mathrm{BSS} + p\sigma^2 \end{aligned}$$

- Thus, a reasonable estimator of the prediction error is

$$n\sigma^2 + \mathrm{RSS} - (n-p)\sigma^2 + p\sigma^2 = \mathrm{RSS} + 2p\sigma^2$$

- This produces the same criterion as C_p, up to the constant term of $-n$, which is the same for all models

AIC

- Consider, however, evaluating the accuracy of the predictions using the log likelihood – *i.e.*, trying to estimate

$$\mathrm{E} \sum_i \log \Pr_{\hat{\theta}}(Y_i)$$

where $\hat{\theta}$ is the maximum likelihood estimate of the parameters of the distribution function of y (and once again, the y's used to fit the model are different from the y's used to evaluate the fit)

- This idea was originally proposed by Akaike, who showed that

$$-2\mathrm{E} \sum_i \log \Pr_{\hat{\theta}}(Y_i) \approx -2\mathrm{E}(\mathrm{loglik}) + 2p,$$

where loglik is the log-likelihood of the fitted model

AIC (cont'd)

- This suggests the following criterion, named the *Akaike information criterion*:

$$\text{AIC} = -2\text{loglik} + 2p$$

- However, because $-2\text{loglik} = n\log(\text{RSS})$ plus some other constants, we can write

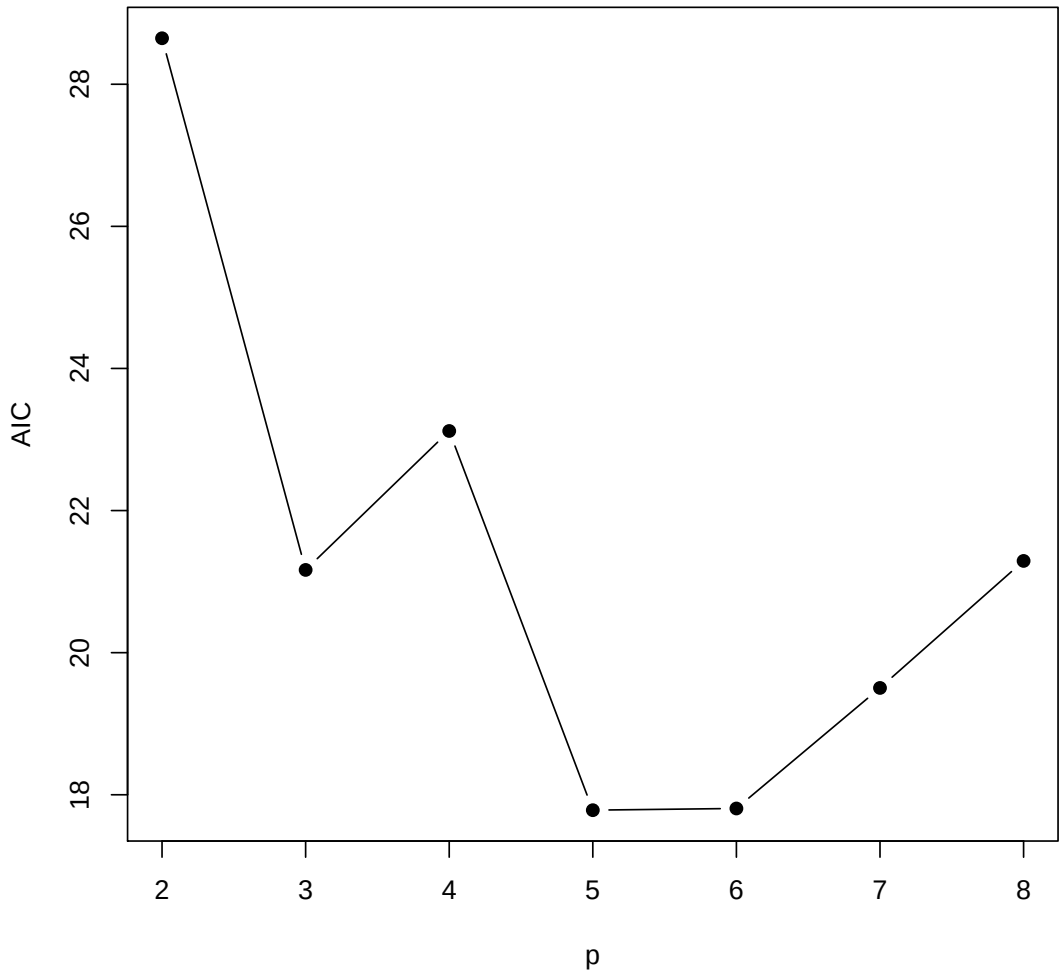
$$\text{AIC} = n\log(\text{RSS}) + 2p$$

- Note that this is not actually equal to the above AIC, but as long as you use the same definition to evaluate all the models, the relative ordering will be the same

Advantages and drawbacks of AIC

- A drawback of this criterion is that it only holds approximately (asymptotically)
- However, AIC has two considerable advantages:
  - It does not require a reference model which is assumed to be able to estimate σ^2
  - It is readily extended to other distributions besides the normal, which will come in handy later on when we look at generalized linear models

AIC: Illustration



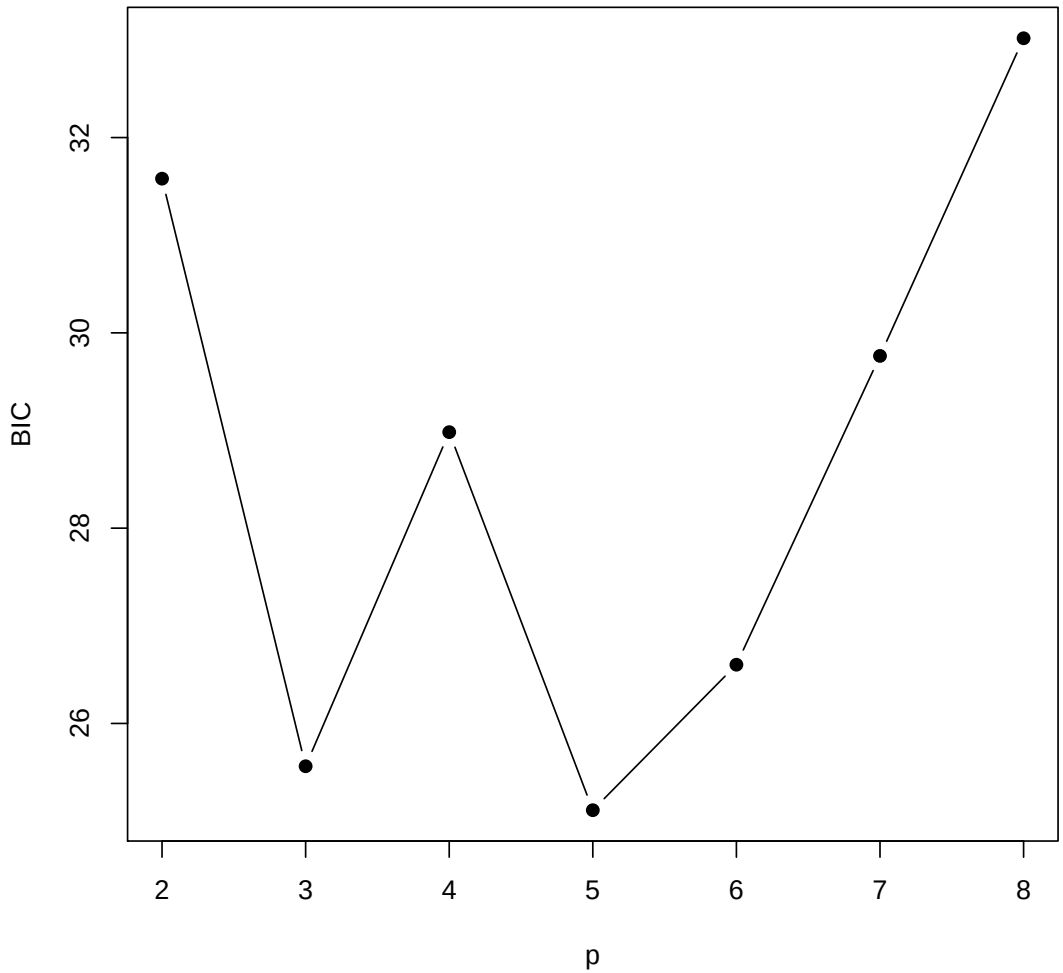
BIC

- A somewhat related criterion is called the *Bayesian information criterion*, or BIC
- As you might guess, its derivation is Bayesian and beyond the scope of this course
- However, its form is very similar to AIC:

$$\mathrm{BIC} = n \log(\mathrm{RSS}) + p \log(n)$$

- Note that because $\log(n)$ is bigger than 2 (unless the sample size is impractically small), BIC penalizes model complexity more heavily than AIC, and thus tends to favor highly parsimonious models

BIC: Illustration



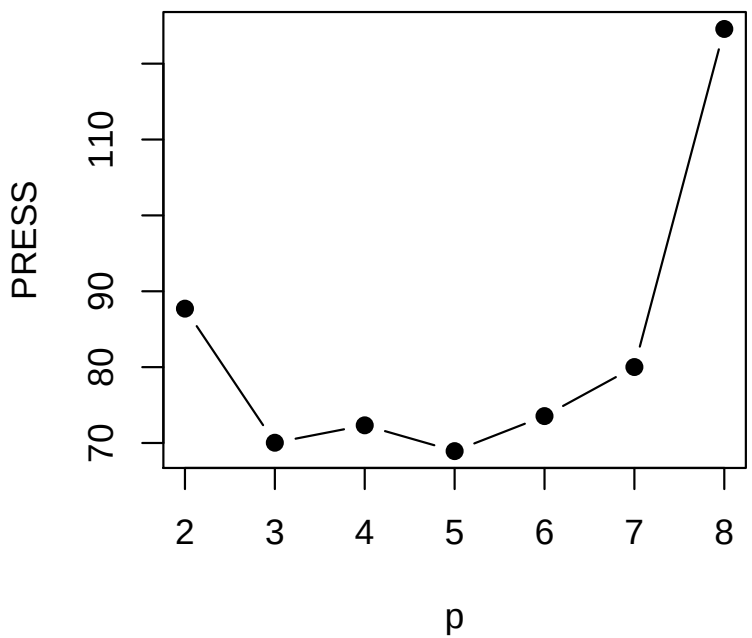
PRESS

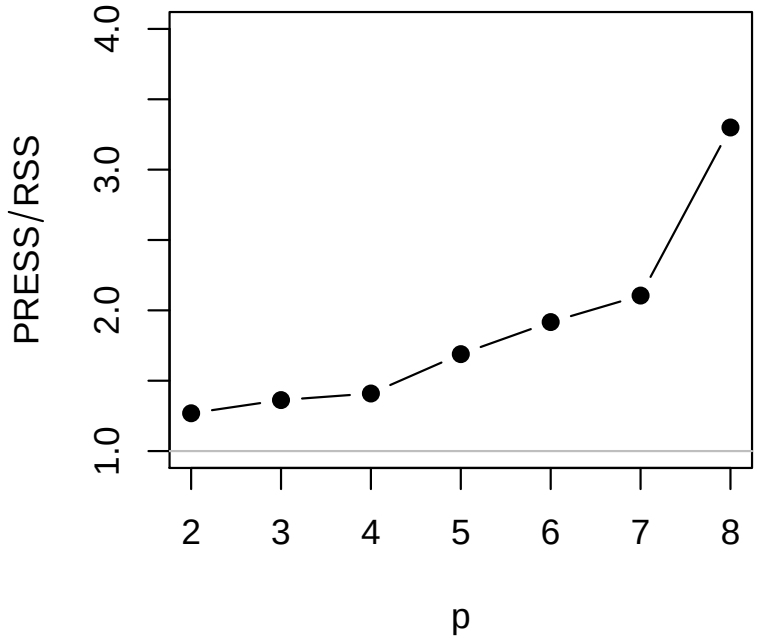
- One final measure that can be used to estimate the generalization error of a model is to re-fit the model without the ith observation and measure how far off its prediction is from y_i:

$$\sum_i (y_i - \hat{\mu}_{i(-i)})^2 = \sum_i \left(\frac{r_i}{1 - H_{ii}} \right)^2$$

- This statistic is called PRESS (for prediction sum of squares)
- PRESS can also be used to check for overfitting: if PRESS is much higher than RSS, then overfitting has occurred

PRESS: Illustration





Remarks

- Each of the model selection criteria we have talked about have their own strengths and weaknesses
- For example, none of AIC, BIC, or PRESS are invariant to a change of scale in the outcome variable, making them useless for choosing transformations
- On the other hand, R^2 and $\mathrm{R}^2_{\mathrm{adj}}$ are invariant and can be used to select transformations:

	R^2	$\mathrm{R}^2_{\mathrm{adj}}$
`Ozone`	0.616	0.602
$\log($`Ozone`$)$	0.667	0.654
$\log($`Ozone+3`$)$	0.688	0.676
$\sqrt{\texttt{Ozone}}$	0.679	0.667
$\sqrt[5]{\texttt{Ozone}}$	0.686	0.675

Remarks (cont'd)

- However, as we have seen, R^2 and $\mathrm{R}^2_{\mathrm{adj}}$ are not particularly useful for choosing between models with different numbers of variables
- Among the rest, PRESS and C_p are exact (although C_p relies on a shaky assumption that you can reliably estimate the true error variance), although they do not generalize well to non-linear models
- Meanwhile, AIC and BIC are applicable to all likelihood-based models, and as a result are the most popular of the model selection criteria, although their derivations rely on asymptotic approximations which may not be valid at smaller sample sizes