

Model Selection & Criteria

R-squared and Adjusted R-squared

$$1 - \frac{SSE}{SST}$$
$$1 - \frac{nSSE}{SST}$$

- ▶ R-squared is defined the same as before:

$$R^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO}$$

Which measure the proportion of the total variation in y that associated with the regression model with predictors x_1, \dots, x_{p-1} .

- ▶ Adding more predictors to the model would only increase R^2 and never reduce it.
- ▶ Adjusted coefficient of multiple determination, denoted by $R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$, might decrease when more predictors are introduced to the model.

$$R_a^2 < R^2$$

Potential
pool of preds
 $X_1 \rightarrow X_5$

What if the Regression Equation Contains "Wrong" Predictors?

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Before we can go off and learn about variable selection methods, we first need to understand the consequences of a regression equation containing the "wrong" or "inappropriate" variables. There are four possible outcomes when formulating a regression model for a set of data:

- ▶ The regression model is "correctly specified." ←
- ▶ The regression model is "underspecified."
- ▶ The regression model contains one or more "extraneous variables."

not collinear

← multicollinearity



- ▶ The regression model is "overspecified."

rank deficient
not full rank

$$X = \begin{pmatrix} 1 & 180 & 180 \\ \vdots & 170 & 170 \\ \vdots & \vdots & \vdots \end{pmatrix}$$


$$\text{rank}(X) = 2$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$n > p$$

$$\text{rank}(X^T X) = n < p$$

The four possible outcomes

- 
- ▶ **Correctly specified (outcome 1):** if the regression equation contains all of the relevant predictors, including any necessary transformations and interaction terms. That is, there are no missing, redundant or extraneous predictors in the model. Of course, this is the best possible outcome and the one we hope to achieve.
 - ▶ **Underspecified (outcome 2):** if the regression equation is missing one or more important predictor variables. This situation is perhaps the worst-case scenario, because an underspecified model yields biased regression coefficients and biased predictions of the response. That is, in using the model, we would consistently underestimate or overestimate the population slopes and the population means. To make already bad matters even worse, the mean square error MSE tends to overestimate σ^2 , thereby yielding wider confidence intervals than it should.

Example of an underspecified model

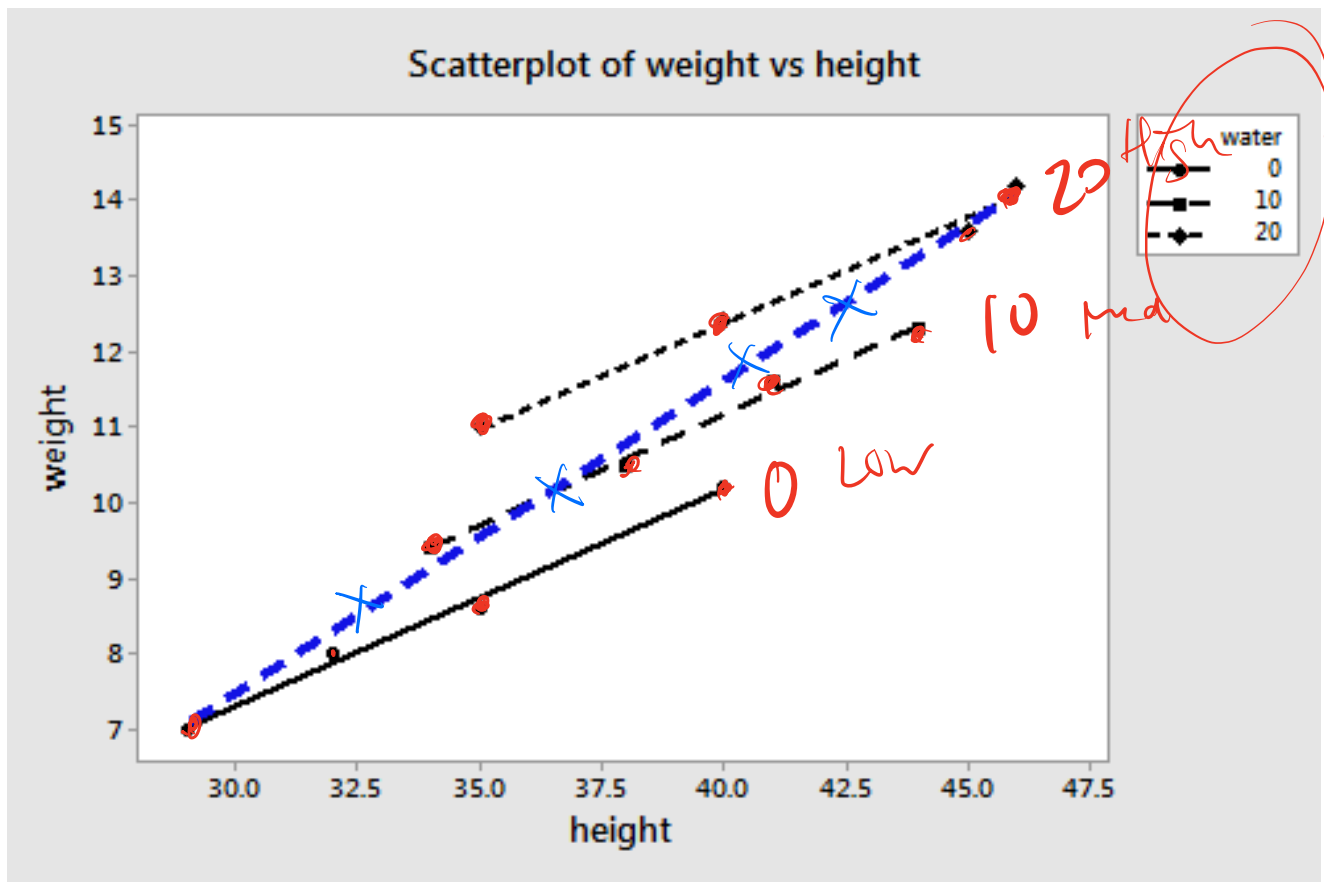
A data set contains $y = \text{weight}$ on the predictors $x_1 = \text{height}$ and $x_2 = \text{water}$, where "water" is the amount of daily water consumption (0, 10 or 20 cups per day) of 12 martians. Let's fit two models out of this data set:

- ▶ $\text{Weight}_i = \beta_0 + \beta_1 * \text{Height}_i + \epsilon_i$ ✓
- ▶ $\text{Weight}_i = \beta_0 + \beta_1 * \text{Height}_i + \beta_2 * \text{Water}_i + \epsilon_i$ ✓ ✗

Which has the results:

- ▶ $\hat{\text{Weight}}_i = -4.14 + 0.3889 * \text{Height}$ with $MSE = 0.653$
- ▶ $\hat{\text{Weight}}_i = -1.22 + 0.28344 * \text{Height} + 0.11121 * \text{Water}$ with $MSE = 0.017$

Example of an underspecified model



Example of an underspecified model

The first — in which water is left out of the model — is likely an underspecified model. Now, what is the effect of leaving water consumption out of the regression model?

- ▶ The slope of the line (0.3889) obtained when height is the only predictor variable (1) is much steeper than the slopes of the three parallel lines (0.28344) obtained by including the effect of water consumption, as well as height(2). That is, the slope likely overestimates the actual slope.
- ▶ The intercept of the line (-4.14) obtained in (1) is smaller than the intercepts of the three parallel lines (-1.220, $-1.220 + 0.11121(10) = -0.108$, and $-1.220 + 0.11121(20) = 1.004$) obtained by (2). That is, the intercept likely **underestimates** the actual intercepts.
- ▶ The estimate of the error variance σ^2 (MSE = 0.653) obtained in (1) is about 38 times larger than the estimate obtained (MSE = 0.017) in (2). That is, MSE likely **overestimates** the actual error variance σ^2 .

The four possible outcomes

- ▶ **The regression model contains one or more extraneous variables (outcome 3):** the regression equation contains extraneous variables that are neither related to the response nor to any of the other predictors. It is as if we went overboard and included extra predictors in the model that we didn't need.
- ▶ The good news is that such a model does yield unbiased regression coefficients, unbiased predictions of the response, and an unbiased MSE. The bad news is that — because we have more parameters in our model — MSE has fewer degrees of freedom associated with it. When this happens, our confidence intervals tend to be wider and our hypothesis tests tend to have lower power. It's not the worst thing that can happen, but it's not too great either. By including extraneous variables, we've also made our model more complicated and hard to understand than necessary.

The four possible outcomes

multicollinearity



- ▶ **The regression model is overspecified (outcome 4):** then the regression equation contains one or more redundant predictor variables. That is, part of the model is correct, but we have gone overboard by adding predictors that are redundant. Redundant predictors lead to problems such as inflated standard errors for the regression coefficients.

A goal and a strategy

1. **Know your goal, know your research question.**

Knowing how you plan to use your regression model can assist greatly in the model building stage. Do you have a few particular predictors of interest? If so, you should make sure your final model includes them. Are you just interested in predicting the response? If so, then multicollinearity should worry you less. Are you interested in the effects that specific predictors have on the response? If so, multicollinearity should be a serious concern. Are you just interested in summary description? What is it that you are trying to accomplish?

2. Identify all of the possible candidate predictors.

3. Use variable selection procedures to find the middle ground between an underspecified model and a model with extraneous or redundant variables.

4. Fine-tune the model to get a correctly specified model.

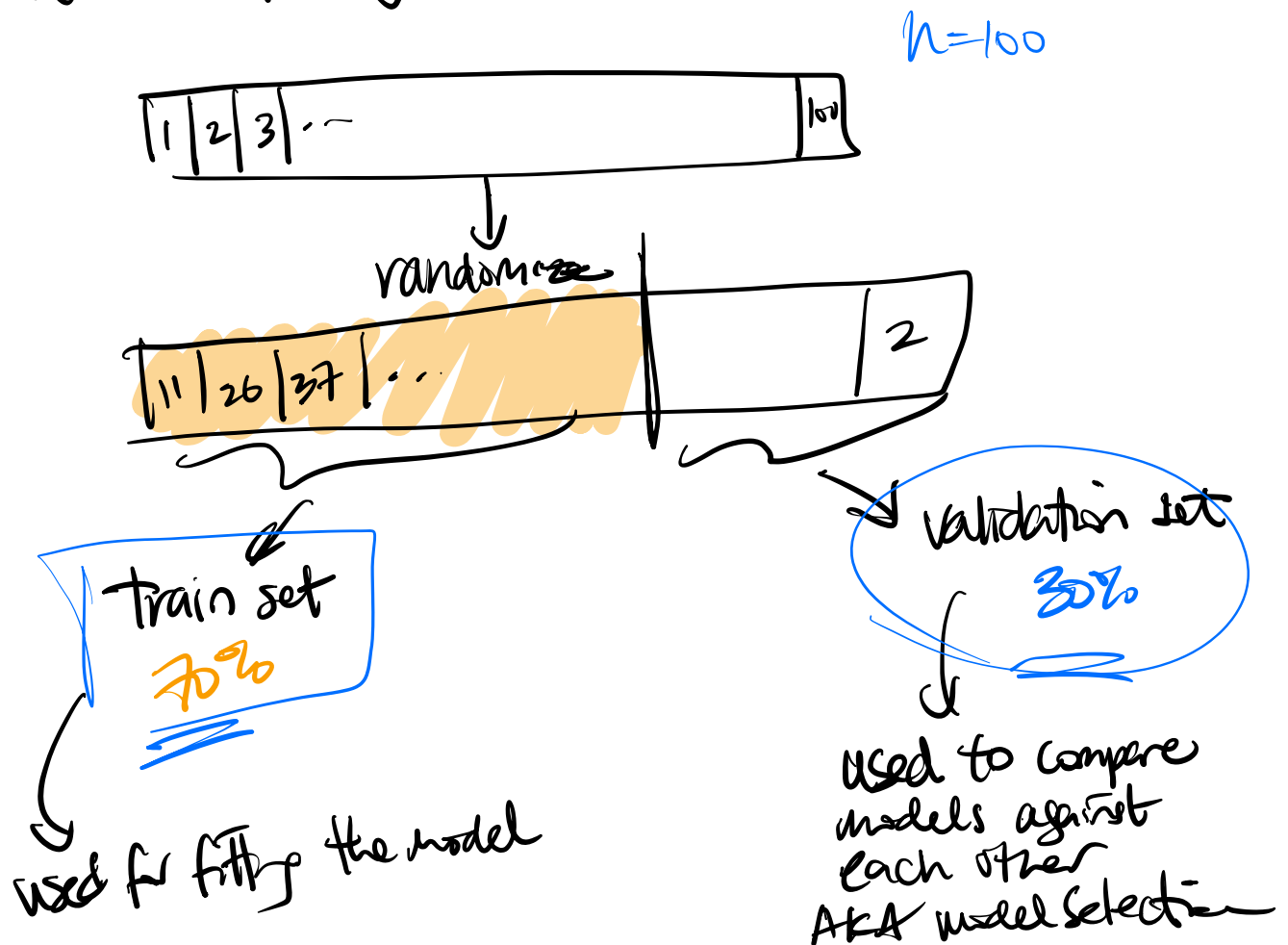
About model selection

- ▶ When we have many predictors (with many possible interactions), it can be difficult to find a good model: Which main effects do we include? Which interactions do we include? Model selection tries to “simplify” this task.
- ▶ This is an “unsolved” problem in statistics: there are no magic procedures to get you the “best model.”
- ▶ Data miners / machine learners often work with very many predictors.
- ▶ To “implement” this, we need
 - ▶ A criterion or benchmark to compare two models. ✱
 - ▶ A search strategy.
- ▶ With a limited number of predictors, it is possible to search all possible models.

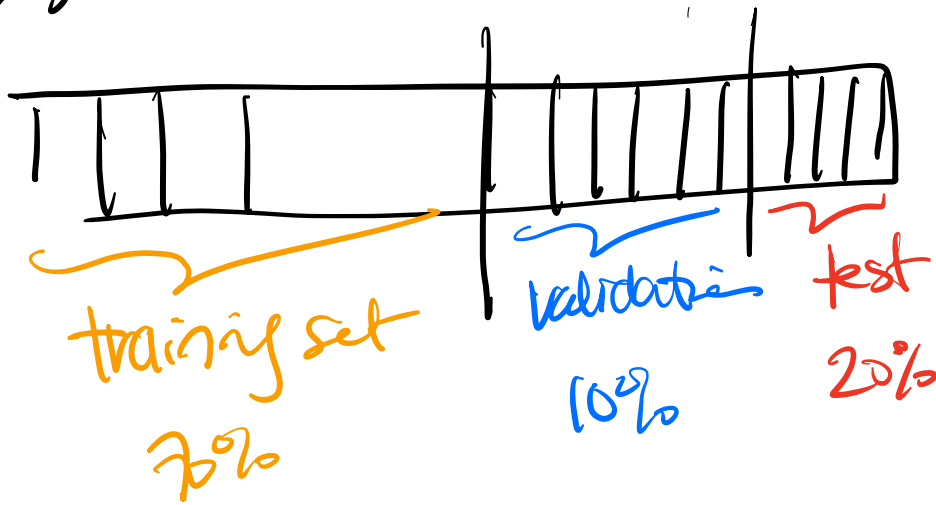
Data Splitting

If we "train" (fit) our model on the same set of data that we then later on evaluate it with, our criterion will be "optimistic".

To prevent this, we need the idea of data splitting.



⊛ Sometimes you see 3 splits:



- ① training: fitting the model (estimating β)
- ② validation: model selection (which model is best?)
- ③ Test: estimate out-of-sample error (generalization error)

Exhaustive search

p potential predictors

of candidate models = 2^p

Best Subset Selection

Stepwise regression ~~X~~

- ▶ The general idea behind the stepwise regression procedure is that we build our regression model from a set of candidate predictor variables by entering and removing predictors — in a stepwise manner — into our model until there is no justifiable reason to enter or remove any more.
- ▶ To avoid underspecified model, a fundamental rule of the stepwise regression procedure is that the list of candidate predictor variables must include all of the variables that actually predict the response.
- ▶ Before the steps, we have to significant levels involved:
 - ▶ α_E -to-enter: significance level for deciding when to enter a predictor into the stepwise model.
 - ▶ α_R -to-remove: significance level for deciding when to remove a predictor from the stepwise model.

$X_1 \quad X_2 \quad X_3$

$Y \sim X_1$

$Y \sim X_1 + X_2$

$Y \sim X_1 + X_2 + X_3$

$2^3 = 8$

$Y \sim X_2$

Forward stepwise regression

1. Fit each of the one-predictor models.
 - ▶ Of those predictors whose $t - test - P - value < \alpha_E$, the first predictor put in the stepwise model is the predictor that has the smallest t-test P-value.
 - ▶ If no predictors has $t - test - P - value < \alpha_E$, stop.
2. Suppose X_1 was deemed the "best" single predictor arising from the the first step, fit each of the two-predictor models that include X_1 as a predictor.
 - ▶ The second predictor to put in the model is the one have smallest t test p value that is $< \alpha_E$. Otherwise stop the model.
 - ▶ Suppose X_2 was deemed the "best" single predictor, check the significance of β_1 now is greater than α_R , remove X_1 .
3. Continue the steps as described above until adding an additional predictor does not yield a t-test P-value below α_E .

Stepwise regression

Cautions:

- ▶ The final model is not guaranteed to be optimal in any specified sense.
- ▶ Stepwise regression does not take into account a researcher's knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
- ▶ One should not over-interpret the order in which predictors are entered into the model.
- ▶ One should not jump to the conclusion that all the important predictor variables for predicting y have been identified, or that all the unimportant predictor variables have been eliminated. It is, of course, possible that we may have committed a Type I or Type II error along the way.
- ▶ The probability is high that we included some unimportant predictors or excluded some important predictors.

- generally outdated

Best subsets regression

The general idea behind **best subsets regression** is that we select the subset of predictors that do the best at meeting some well-defined objective criterion, such as having the largest adjusted R^2 or the smallest MSE.

▶ Recall adjusted R^2 : $R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}}$

▶ Mallows C_p -statistic estimates the size of the bias that is introduced into the predicted responses by having an underspecified model: recall an underspecified model is a model in which important predictors are missing. And, an underspecified model yields biased regression coefficients and biased predictions of the response.

Model Selection Criteria:

- R_a^2
- Mallows's C_p
- AIC
- BIC
- PRESS

Mallows C_p

- Two issues with any regression model:
 - The **bias** in the predicted responses.
 - The **variation** in the predicted responses.
- Bias in predicted responses: $B_i = E(y_i) - E(\hat{y}_i)$.
- Variation in the predicted responses is due to two things: the ever-present random sampling variation $\sigma_{\hat{y}_i}^2$, and the variation associated with the prediction bias B_i^2 .
- The (standardized) measure of the total variation in the predicted responses is denoted by

$$\Gamma_p = \frac{E(\sum_{i=1}^n (\hat{y}_i - \mu_i)^2)}{\sigma^2} = \frac{1}{\sigma^2} \left(\underbrace{\sum_{i=1}^n \sigma_{\hat{y}_i}^2}_{\text{var}} + \underbrace{\sum_{i=1}^n (E(y_i) - E(\hat{y}_i))^2}_{\text{bias}^2} \right)$$

mean squared error

Navigation icons: back, forward, search, etc.

$$\begin{aligned} \sum_i \text{Var}(\hat{y}_i) &= \text{tr}(\text{Var}(\hat{\mathbf{y}})) = \text{tr}(\text{Var}(\mathbf{H}\mathbf{Y})) \\ &= \text{tr}(\mathbf{H}\text{Var}(\mathbf{Y})\mathbf{H}^T) \\ &= \text{tr}(\mathbf{H}\sigma^2\mathbf{I}\mathbf{H}) \\ &= \text{tr}(\sigma^2\mathbf{H}^2) = \text{tr}(\sigma^2\mathbf{H}) \end{aligned}$$

$\Gamma_p = \frac{1}{\sigma^2} \cdot \sigma^2 p = p$ ideal value

$$= \sigma^2 \underline{\text{tr}(H)} = \sigma^2 p$$

exercise

Mallows C_p

- ▶ It can be shown that if there is no bias in the predicted responses, Γ_p achieves its smallest possible value p .
- ▶ The best model is simply the model with the smallest value of Γ_p .
- ▶ Mallows C_p of a model with k predictors (hence, $k+1$ parameters), and $p - 1$ total number of candidate predictors (hence, p parameters in full model) is an estimate of Γ_p given by

$$C_p = p + \frac{(MSE_k - MSE_p)(n - p)}{MSE_p} = \frac{SSE_k}{MSE_p} + 2(k + 1) - n$$

- ▶ Mallows C_p is set up in the way that $C_p = k + 1$ when using the full model. So you shouldn't use C_p to evaluate the full model. When C_p is near $k + 1$, the bias is small; when C_p is much greater than $k + 1$, the bias is substantial.

① Data Splitting

train validation set

② on candidate models

③ Calculate Mallows C_p using validation set

④ Pick model w/ $|C_p - p|$ minimized

Mallows C_p

A reasonable strategy for using C_p to identify "best" models:

- ▶ Identify subsets of predictors for which the C_p value is near $k+1$ (if possible).
- ▶ If all models, except the full model, yield a large C_p not near $k+1$, it suggests some important predictor(s) are missing from the analysis. In this case, we are well-advised to identify the predictors that are missing.
- ▶ If a number of models have C_p near $k+1$, choose the model with the smallest C_p value.
- ▶ When more than one model has a small value of C_p value near $k+1$, in general, choose the simpler model or the model that meets your research needs.

"parsimonious"

Maximum likelihood estimate

Before we go into **AIC** and **BIC**, a quick look at MLE:

- ▶ **Basic idea:** Suppose a random sample X_1, \dots, X_n assumed probability distribution depends on some unknown parameter θ , it seems reasonable that a good estimate of θ would be the value that maximizes the probability, that is, the likelihood... of getting the data we observed.
- ▶ Suppose the p.d.f. of X_i is $f(x_i; \theta)$, then the **joint likelihood function** is defined as

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

- ▶ $\hat{\theta}$ that maximizes the likelihood function is called the **maximum likelihood estimator** of θ

MLE for multiple linear regression

- ▶ Recall under the setup of MLR, $y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = (X\beta)_i$ so $f(y_i; \beta, \sigma) = \frac{1}{\sigma\sqrt{\pi}} \exp[-\frac{(y_i - \mu_{i(\beta)})^2}{2\sigma^2}]$
- ▶ So the likelihood function

$$L(\beta, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{\pi}} \exp[-\frac{(y_i - \mu_{i(\beta)})^2}{2\sigma^2}]$$

- ▶ To maximize the likelihood function, it's equivalent to maximize the *log* of likelihood function

$$\ell(\beta, \sigma) = \log(L) = \sum_{i=1}^n \log[\frac{1}{\sigma\sqrt{\pi}} \exp[-\frac{(y_i - \mu_{i(\beta)})^2}{2\sigma^2}]]$$

$$= -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (X\beta)_i)^2$$

MLE for multiple linear regression

- ▶ MLE for β in this case is the same as OLSE.
- ▶ MLE for σ^2 is $\hat{\sigma}_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - (X\beta)_i)^2 = \frac{1}{n} SSE$
- ▶ Note that MLE for σ^2 is biased.

Akaike's Information Criterion (AIC)

- ▶ AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.
- ▶ In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model. In other words, AIC deals with both the risk of overfitting and the risk of underfitting.
- ▶ Therefore AIC is defined as the ^{negative} log-likelihood term penalized by the number of model parameters. The larger the likelihood, the better the model. The more parameters, the worse the model.

Akaike's Information Criterion (AIC)

- ▶ AIC of a multiple linear model with k predictors (hence, $k+1$ parameters) is defined as (when σ^2 is known)

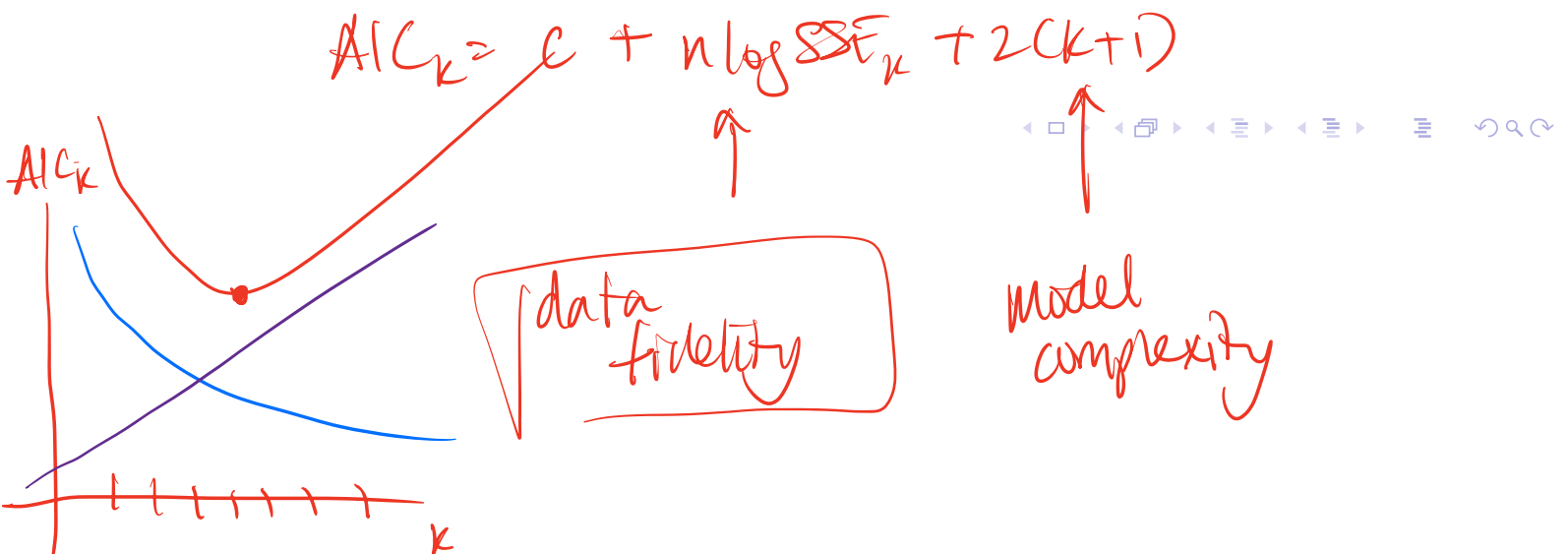
$$AIC_k = 2(k + 1) - 2\ell(\beta)$$

$$= n\log 2\pi + n\log \sigma^2 + \frac{SSE_k}{\sigma^2} + 2(k + 1)$$

when σ^2 is known, otherwise, estimate σ^2 by $\frac{1}{n}SSE_k$ and have

$$AIC_k = n\log 2\pi + \underbrace{n\log SSE_k}_{\text{data fidelity}} - \underbrace{n\log n + 2(k + 1)}_{\text{model complexity}}$$

- ▶ Sometimes, you see people use AIC values without some constants, but the above definition is more precise.



Schwarz's bayesian information criterion (BIC)

- ▶ When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC.
- ▶ The BIC was developed by Gideon E. Schwarz, who gave a Bayesian argument for adopting it. It is closely related to the Akaike information criterion (AIC). In fact, Akaike was so impressed with Schwarz's Bayesian formalism that he developed his own Bayesian formalism, now often referred to as the ABIC for "a Bayesian Information Criterion" or more casually "Akaike's Bayesian Information Criterion".

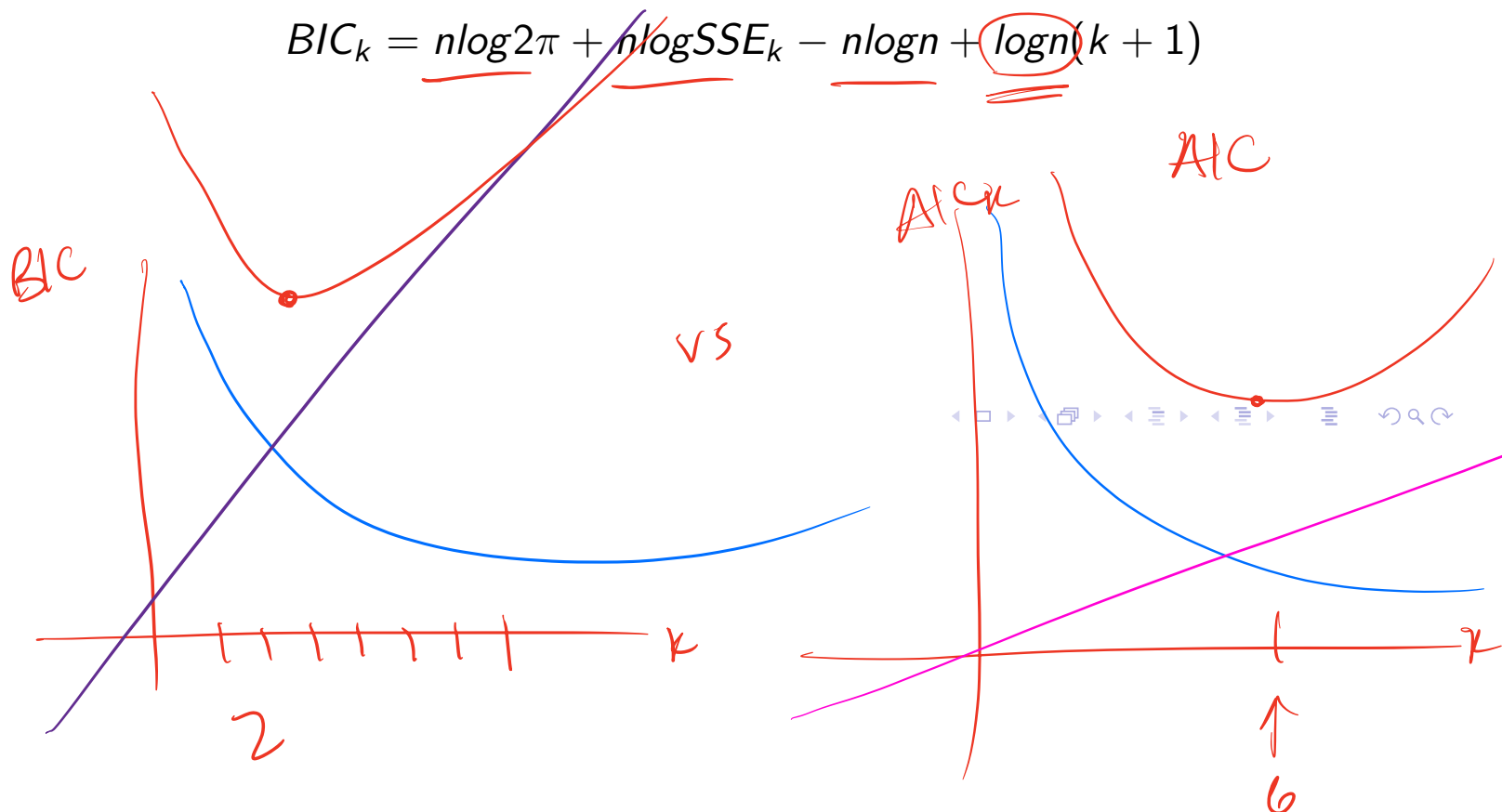
Schwarz's bayesian information criterion (BIC)

BIC of a multiple linear model with k predictors (hence, $k+1$ parameters) is defined as

$$BIC_k = \log n(k+1) - 2\ell(\beta)$$

therefore can be estimated by $\frac{1}{n}SSE_k$ and have

$$BIC_k = \underbrace{n \log 2\pi} + \underbrace{n \log SSE_k} - \underbrace{n \log n} + \underbrace{\log n(k+1)}$$



More about BIC

- ▶ For the BIC there is a Bayesian rationale. Each model has a prior probability and density.
- ▶ It is asymptotically (as $n \rightarrow \infty$) equivalent to choosing the model with highest posterior probability of being the best model, under some not too restrictive conditions.

AIC v.s. BIC

- ▶ AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth.
- ▶ BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model.
- ▶ Both criteria are based on various assumptions and asymptotic approximations.

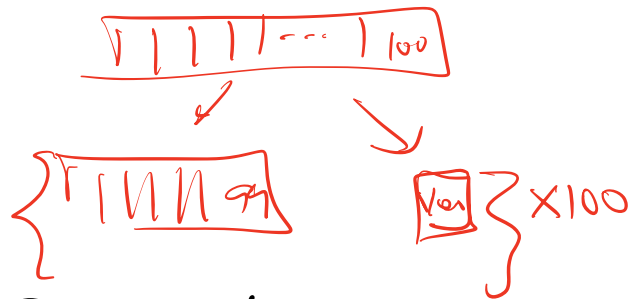
AIC v.s. BIC

- ▶ AIC and BIC are both penalized-likelihood criteria. Their only difference in practice is the size of the penalty: BIC penalizes model complexity more heavily.
- ▶ In general, it might be best to use AIC and BIC together in model selection: Most of the times they will agree on the preferred model, when they don't, just report it.
- ▶ AIC is better in situations when a false negative finding would be considered more misleading than a false positive, and BIC is better in situations where a false positive is as misleading as, or more misleading than, a false negative.
- ▶ More specifically, AIC tries to select the model that most adequately describes an unknown, high dimensional reality. This means that reality is never in the set of candidate models that are being considered. On the contrary, BIC tries to find the TRUE model among the set of candidates.

Prediction Sum of Squared Error (PRESS)

The prediction sum of squares error is a modified version of SSE which uses the predicted value for the i^{th} obs. is obtained from the model fit on the sample of data excluding the i^{th} data pt. Mathematically, we define:

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_{i(i)})^2$$



There is a shortcut for calculating PRESS without having to fit all n models using some clever algebra (of not shown in our class b/c time constraints):

$$\text{PRESS} = \sum_{i=1}^n \frac{e_i^2}{(1-h_{ii})^2} \quad \text{where } e_i = y_i - \hat{y}_i \text{ are the usual model residuals using the model fitted on all } n \text{ datapts.}$$

PRESS can be used as a model selection criteria.