Week 10 Hour 1

~~Shapiro-Wilks Test (from last time)~~

~~Cross-Validation~~

Week 10 Hour 2

Missing Data

Cross-Validation in the Wild

It's often more important to describe the patterns in the data in a way that also applies to new, similar situations.

That's why the standard way of comparing different models from different sources is Cross-Validation.

There are ongoing competitions at kaggle.com where the training set is given out the public in full, and the test set is given out, but with the response variable blanked out.

The Netflix Prize

In 2007-2009, Netflix held a competition with a million dollar prize to find a better model to predict how users would rate movies.

They weren't interested in a model that simply fitted the ratings that were already given. They wanted one that could be applied to predict ratings for movies their users had not yet seen.

For example: Assume user ‘John’ gave both movies ‘Braveheart’ and ‘Mad Max’ a high rating.

User ‘Marsha’ gave ‘Braveheart’ a low rating. We can use the above information predict that she won’t like ‘Mad Max’ either.

Netflix wants to present recommendations that their users will like, so that they stay.

Since the problem of interest was to predict new scores, not existing ones, cross-validation was key.

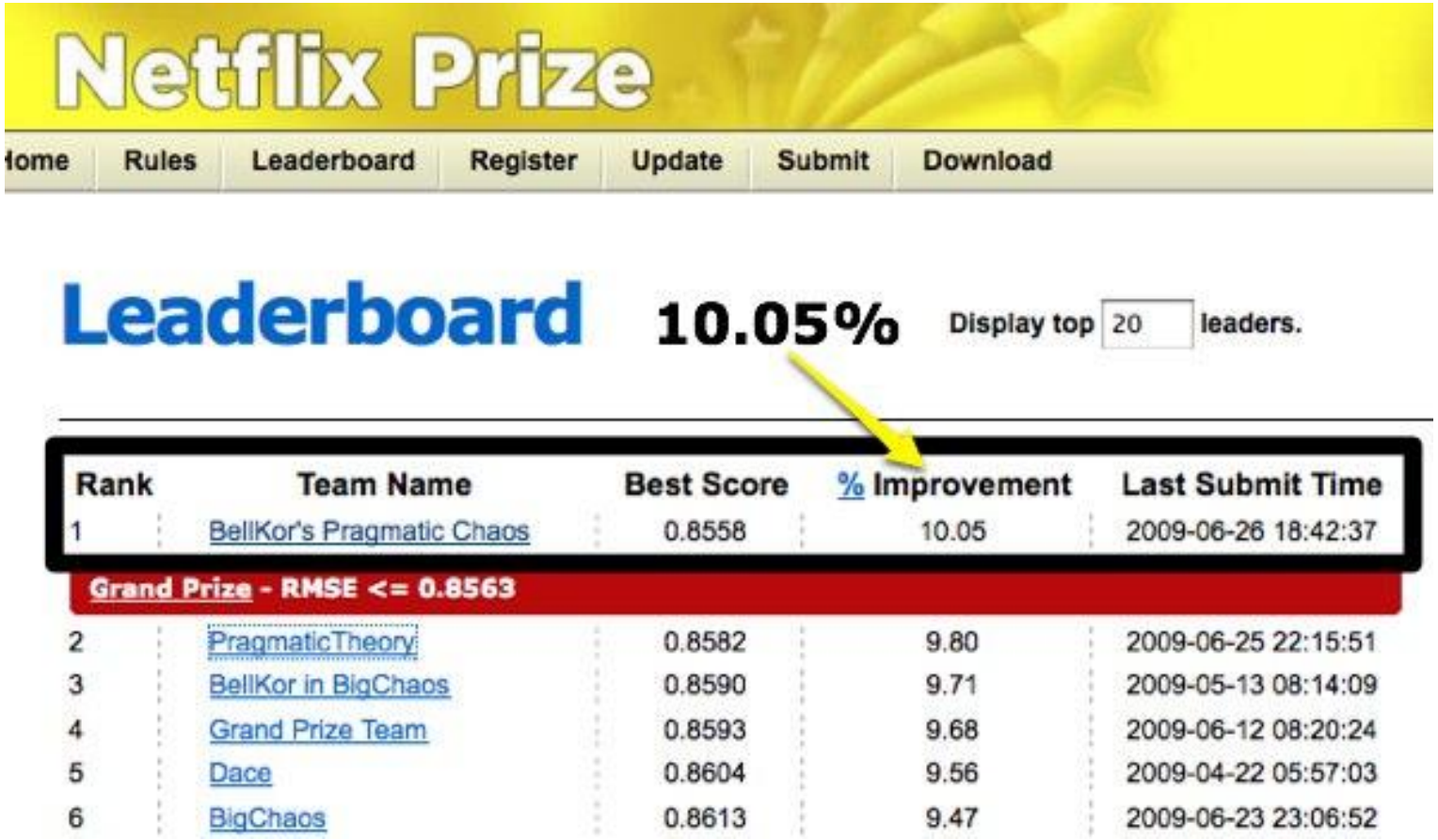
The training data set had 100,480,507 ratings on 17,770 movies from 480,159 different users;
the test set had only 2,817,131 ratings.

The data looked like this:

type	user	movie	rating
Train	116	Princess Bride	5
Train	122	Mad Max	5
Train	129	Braveheart	4
Train	128	Mad Max	4
Train	121	Princess Bride	2
Test	125	Princess Bride	NA
Test	115	Braveheart	NA
Test	112	Princess Bride	NA
Test	115	Braveheart	NA
Test	128	Mad Max	NA

The winning team improved the prediction error by 10% over Netflix's original system.

(The square root of the PRESS statistic)



…solve one problem, and another takes its place.

Missing Data

Missing, or ***incomplete,*** data has several definitions. For the sake of this class, we will treat it to be:

Any occurrence where data for a variable has not been recorded for some observation is considered missing from that observation.

In the Netflix example in cross-validation, the responses in the test set were missing data.
Missing data shows in R as ‘NA’ for ‘Not Applicable’.

Why do we care about missing data?

The default approach is to ignore any observation in a model that has any model variable missing.

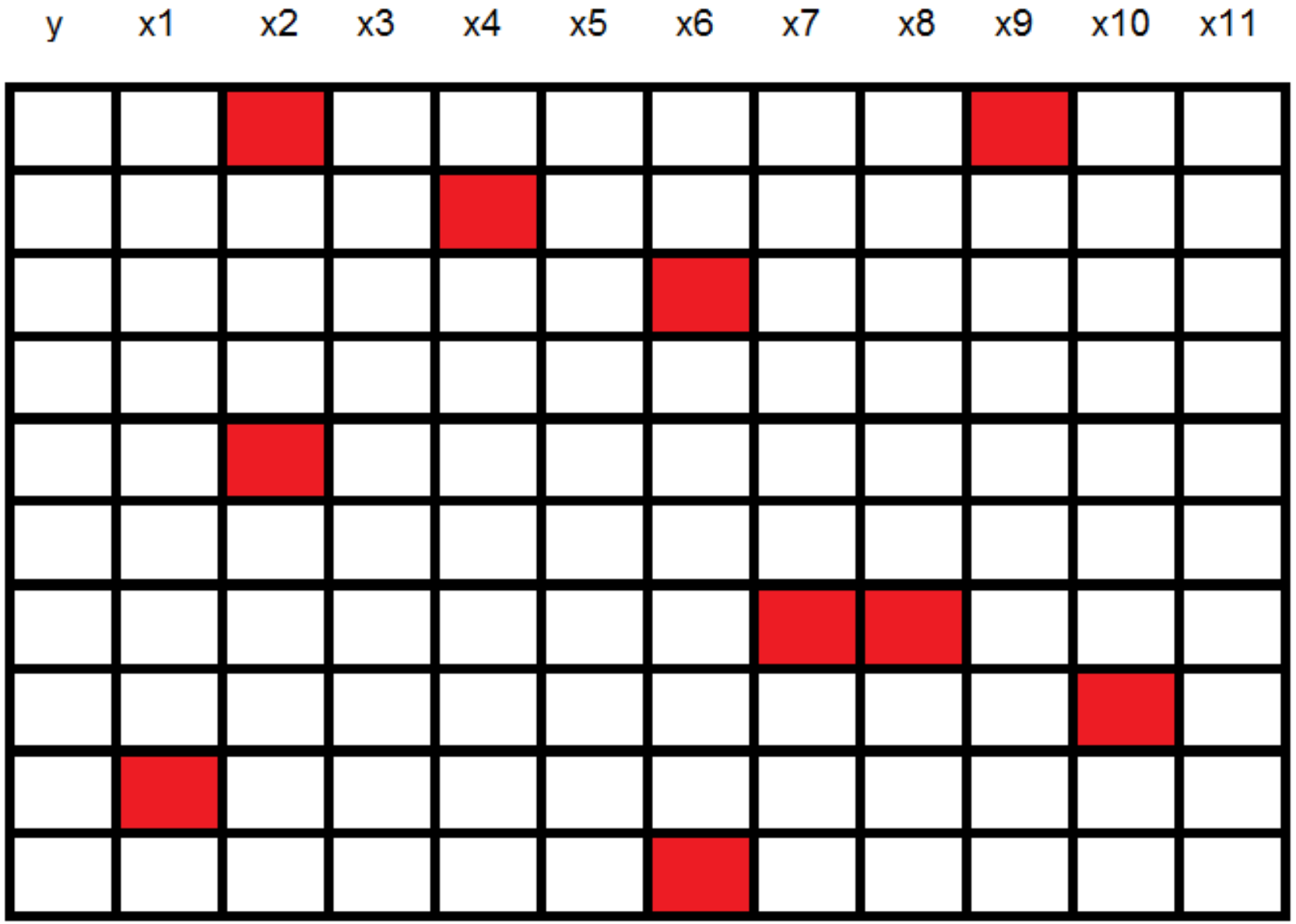
For example, in the birth rate model using the gapminder data, any country that didn't have a recorded value for GDPpercap is not included in any model that uses GDPpercap.

Ignoring observations with missing data is bad for 3 reasons.

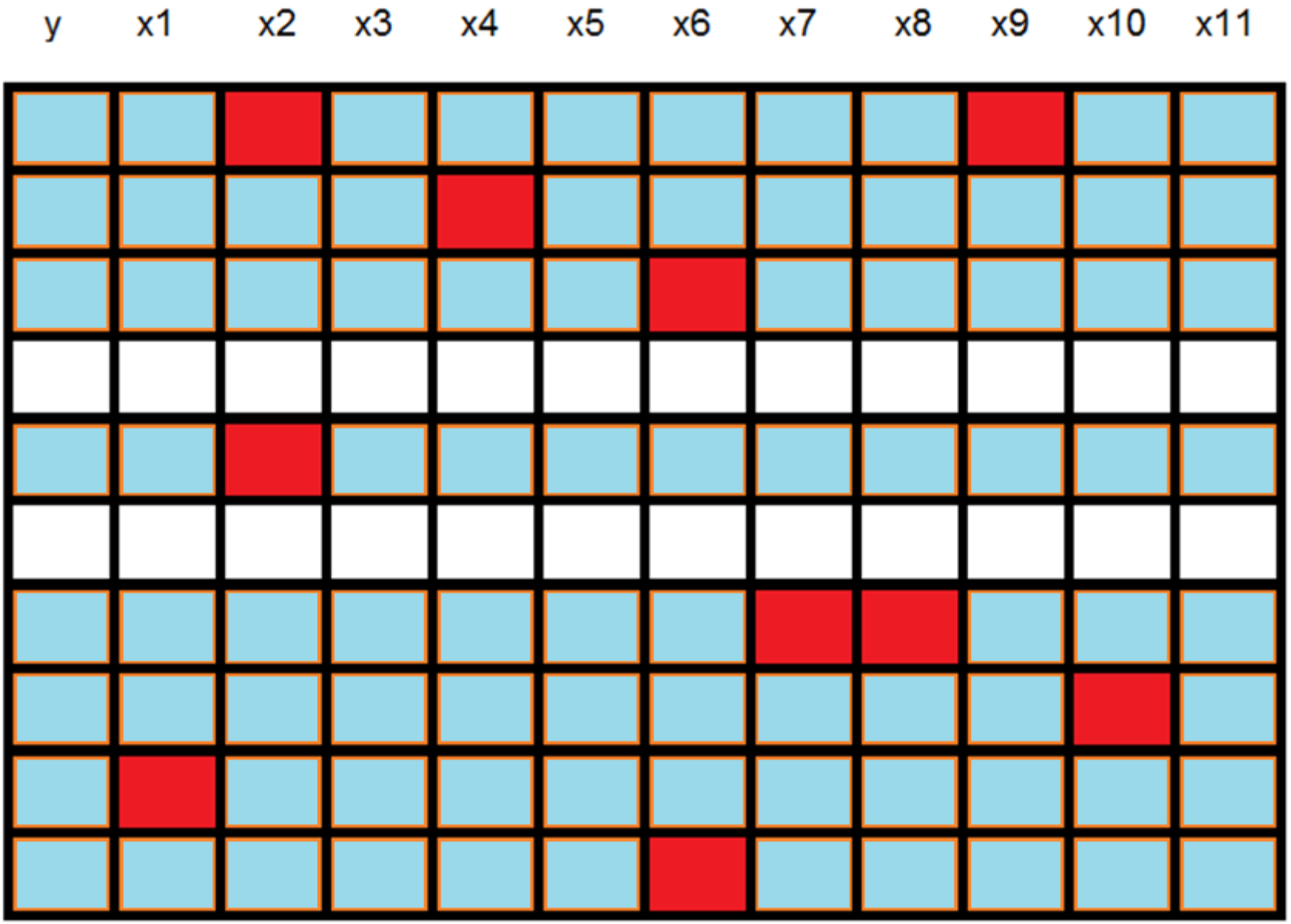
1. **It's wasteful.** If every other relevant variable is recorded for a country, none of that information is used in the model.

In the following images, red squares represent missing data and blue squares unused data. A small portion (about 8%) of the data is missing, but a model using all 11 explanatory variables is to use 8 of the 10 (80%) of the cases.

A relatively small amount of missing data can have a big impact on your sample sizes, especially when many variables are used.



Red squares are missing data. White squares are filled data.



Blue squares are unused data.

Ignoring observations with missing data is bad for 3 reasons.

2. **It creates inconsistency.** Countries are only excluded from the model if one or more of the model's variables are missing for that country.

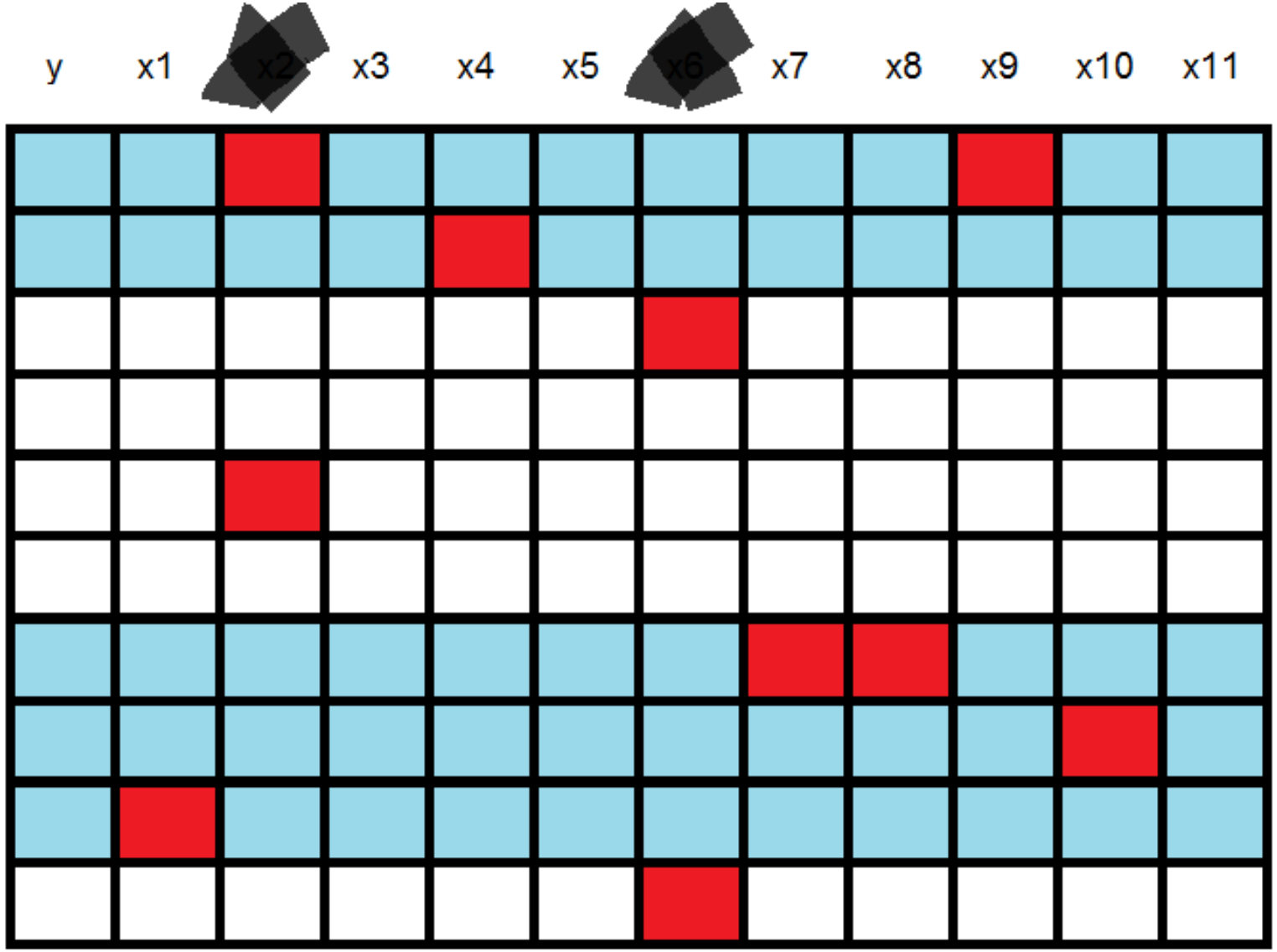
If one model uses GDPpercap, and another doesn't, but some countries have missing GDP data, then how are the two supposed to be compared?

2. **It creates inconsistency. (Continued)**

Methods based on likelihood, such as the AIC and BIC, need the same set of observations between every model in order to make a viable comparison.

In some cases where the AIC is used in an automated process, that process will fail to run when there is missing data.

In the following image, blue is unused data, but only when the explanatory variables x2 and x6 aren't used.



Blue squares are unused data when only x2 and x6 are unused.

Ignoring observations with missing data is bad for 3 reasons.

3. **It creates bias.**

What countries are most likely to have missing GDP per capita data, or missing data for most other variables?

Countries with low income and a poorly developed vital statistics department? Countries with high corruption that are prone to misreporting vital data?

The trends we see may only apply to developed countries.

Missing data is a problem that can’t be solved by ignoring it.

Missing data comes in three classes*.

1. MCAR: Missing ***Completely At Random***
2. MAR: Missing At Random
3. NMAR: Not Missing At Random

These classes represent both the inherit reasons for data to be missing, and the range of things we can do about it to preserve validity.

*The ‘MAR’ class of data is further split in your reading for Assignment 4.

MCAR – Missing Completely At Random

If every value for a variable has the same chance of not being recorded, missing data for this variable is considered MCAR.

Example:
Imagine tracking the number of cars at an intersection over time using a webcam. But the Wi-Fi on your laptop fails occasionally, and you cannot record cars during the outage.

The missing car counts are MCAR.

The fact that they are missing has nothing to do with the cars.

Their missingness also has nothing to do with the weather, time of day, or any other variable related to the cars.

In short, not having that data just means you have a smaller sample.

MCAR data is also called data with ***ignorable*** missingness because the parameter estimates you get from a model should be the same whether the data is imputed or the observations removed.

MAR: Missing At Random

If the chance that a value is missing can be determined entirely by other variables in the dataset, then the data is missing at random.

Returning to the car counting example.

Say the webcam is known to shut down every night from 1am to 5am to save power, or because it's too dark.

These missing car counts are MAR.

A pollster can assume that phone calls made between 9am and 5pm have a small chance of being answered unless the respondent is a stay-at-home parent.

In this case, the chance that the responses for anything in the content of the call being missing can be explained by the time of call, and the demographics of the group or person receiving this call. This information would be given by the address, etc.

In other words, if you know WHY a certain value is missing or not, then you can assume that the missingness in your data is either MAR or MCAR.

There are lots of mechanisms that can cause data to be missing at random.

Some could even be intentional. When pollsters survey people, they often skip questions that they know are irrelevant based on previous answers.

Someone who reports their biological sex as male isn't going to have relevant answers to questions about their pregnancies.

Someone who has never heard of a politician will not have relevant information about their opinion of them.

NMAR: Not Missing At Random

Of the three types of missingness, this is by far the worst.

If data is NMAR, the chance that any value for the given variable is missing depends on data which is itself missing.

The Census encounters NMAR data with it uses households to measure key population statistics. People who do not live in permanent homes are much more likely to have missing data in a census (or most order surveys) because they less likely to be found by pollsters.

This sort of non-random missingness can lead to biases such as underestimating mental or physical disease prevalence in a community.

The worst part is that the biases can't be corrected for, because we don't what data is missing, and we can't explain why any given value is missing from the

We also can't reliably fill in values in the missing data with imputation because we cannot estimate what reasonable replacements should look like.

Imputation

Imputation is the act of filling in missing data.

Missing data be filled with predefined values (e.g. 0)

It can be filled with predictions of what the values should be.

It can even be filled with more than one value per missing data point.

Details of imputation will be given next day.

Next day:

Imputation. (end of Midterm 2 material)

‘Pepsi Challenge’ Source:

Malcolm Gladwell – The Power of Thinking without Thinking, cited in

http://www.trustmeimascientist.com/2012/01/01/the-zen-art-of-choosing-speakers/

Optional discussion: The circular argument and cross-validation.

A hypothesis test asks the question 'how likely is this observed pattern/difference by random chance alone.

This works great when we have a specific relationship of difference to example from the very start.

Example: A basic clinical trial – Does this drug show a difference in the intended outcome, compared to a placebo, in ***the general*** *population of sufferers?*

Answer: ***T-Test***

Sometimes the question isn't restrictive to fit into a single hypothesis test.

Example: An open clinical trial - Does this drug show a difference in the intended outcome, compared to a placebo, in ***a subgroup of*** *the population of sufferers?*

To answer an open question like this, we may start with a t-test using entire sample. After that, it is typical to examine the data and look for patterns (e.g. by doing a multiple regression, or a model selection).

*When we look through a dataset for patterns, we are conducting **exploratory** data analysis.*

After exploratory data analysis is done, hypothesis tests follow.

…and here we find a circular argument.

In a sufficiently complex dataset, there could be hundreds of possible hypotheses to test. When exploring the data, you are going to find some of the most promising ones.

Then you test to see if those ones are random noise.

Are the ones you decide to test, (e.g. with an multi-way ANOVA, or a multiple regression) going to be a fair representation of the possibilities? Of course not.

We have already addressed a similar problem with ***multiple comparisons****.*

Could it work here too?

Our multiple comparison options:

1. P-value adjustments (e.g. Bonferroni correction)

2. Post-hoc tests (e.g. Tukey's Honestly Significant Difference)

For the Bonferroni and for the Tukey HSD (and any other p-value adjustment or post-hoc test), we need to know the number of hypotheses are being tested.

How many hypotheses are there when exploring a dataset?

Do terms not included in the multiple regression count?

Do modifications of terms (transforms, polynomials, interactions) that weren't considered count?

It's unclear, and any answer would be arbitrary.