Takeaways:

The goal of a model is not to uncover truth, but to discover a simple approximation that is still useful. So in choosing a model, we must understand its strengths and limitations. When writing diss, should include limitations and reasons as to why you chose the particular model. Alternatives? Recognize that this could be a bias too. Could document model selection in pre-registration!

Points:

- Issue: Flexibility in data collection and analysis. From initial data collection -> publication given great freedom in what we choose to do. On finding that y does not correlate with x we could shift our attention to z. If our treatment has no effect we could test whether that is because its effects are sex-specific. These types of exploratory analyses play an important role in science, but it would be lunacy to subject them to the lenient significance thresholds that we use for predefined tests. In these two simple cases the true positive rate would fall to an abysmal (1/25)*0.5/((1/25)*0.5+(24/25)*(1-0.95^2)) =18%.
- Possible Solution. Author really wants this. Encourage the use of preregistered analysis and data-collection plans. These plans would provide a more honest and more accurate record of intent that readers, reviewers and, perhaps most importantly, authors themselves can trust. Once the embargoes on preregistration plans expire, the underworld of unpublished studies would be exposed and their detrimental effects could be adjusted for.
- Importance of replicating studies: Exact replication, with sufficient power, allows us – as far as is possible – to differentiate false-positives from context dependency.
    1. Iconic studies often have an influence disproportionate to the evidence they provide, and replication can go someway to mitigating this bias.
    2. The discussions of many papers include (semi) plausible biological explanations for why their conclusions differ from previous studies, and only rarely is it suggested that the original study may have been a false positive.
- Adjust reporting standards: Journals are very particular about the minutiae of how we cite published work, and yet can be relatively laissez-faire about how we report the quantitative information on which many of our scientific conclusions are ultimately based. Synthetic analyses, including meta-analyses, are the most objective way we have of drawing evidence from a body of studies and require access to this quantitative information in a useable form.
  Solution: For many common analyses, standards should be fairly straightforward to develop and would place a minimal burden on authors. Unlike citation formats, the opportunity to standardise across journals should make it a relatively painless process.

Models: It's important to understand that a fitted model is just the closest model from a family of models. That implies that you have the "best" model (according to some criteria); it doesn't imply that you have a good model and it certainly doesn't imply that the model is "true".

This chapter has focussed exclusively on the class of linear models, which assume a relationship of the form y = a_1 * x1 + a_2 * x2 + ... + a_n * xn. Linear models additionally assume that the residuals have a normal distribution, which we haven't talked about. There are a large set of model classes that extend the linear model in various interesting ways. Some of them are:

- Generalised linear models, e.g. stats::glm(). Linear models assume that the response is continuous and the error has a normal distribution. Generalised linear models extend linear models to include non-continuous responses (e.g. binary data or counts). They work by defining a distance metric based on the statistical idea of likelihood.
- Generalised additive models, e.g. mgcv::gam(), extend generalised linear models to incorporate arbitrary smooth functions. That means you can write a formula like y ~ s(x) which becomes an equation like y = f(x) and let gam() estimate what that function is (subject to some smoothness constraints to make the problem tractable).
- Penalised linear models, e.g. glmnet::glmnet(), add a penalty term to the distance that penalises complex models (as defined by the distance between the parameter vector and the origin). This tends to make models that generalise better to new datasets from the same population.
- Robust linear models, e.g. MASS::rlm(), tweak the distance to downweight points that are very far away. This makes them less sensitive to the presence of outliers, at the cost of being not quite as good when there are no outliers.
- Trees, e.g. rpart::rpart(), attack the problem in a completely different way than linear models. They fit a piece-wise constant model, splitting the data into progressively smaller and smaller pieces. Trees aren't terribly effective by themselves, but they are very powerful when used in aggregate by models like random forests (e.g. randomForest::randomForest()) or gradient boosting machines (e.g. xgboost::xgboost.)

Tip: Instead of using lm() to fit a straight line, you can use loess() to fit a smooth curve.

- Dealing with categorical variables:

Generating a function from a formula is straight forward when the predictor is continuous, but things get a bit more complicated when the predictor is categorical. Imagine you have a formula like y ~ sex, where sex could either be male or female. It doesn't make sense to convert that to a formula like y = x_0 + x_1 * sex because sex isn't a number - you can't multiply it! Instead what R does is convert it to y = x_0 + x_1 * sex_male where sex_male is one if sex is male and zero otherwise:

```
df <- tribble(
  ~ sex, ~ response,
  "male", 1,
  "female", 2,
  "male", 1
)
model_matrix(df, response ~ sex)
#> # A tibble: 3 x 2
#>   `(Intercept)` sexmale
#>           <dbl>   <dbl>
#> 1             1       1
#> 2             1       0
#> 3             1       1
```