University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 12, 2025

Aids allowed (on paper, no computers):

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Non-programmable, non-communicating calculator

This exam has 13 numbered pages of questions, including this cover page.

In addition, you have an additional booklet of Figures to refer to during the exam.

The maximum marks available for each question are shown next to the question number.

If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.

You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.

The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Tom Izzo

Each year, 64 teams are selected to compete for the national championship in a US college sport. The teams are divided into four regions, and the sixteen teams within a region are seeded from 1 (believed to be best) to 16 (believed to be worst). The teams play a knockout tournament, with the winning team in each region advancing to the Final Four. (These winning four teams then play for the national championship). A total of 1664 teams were selected for the championship between 1985 and 2010. In the dataset shown in Figure 2, each row represents one of those teams; the year of participation of that team is shown, together with their seed number that year, and whether or not they advanced to the Final Four (1 is yes, 0 is no). The dataframe is called `FinalFourIzzo`.

Tom Izzo is a well-known college coach in this sport. Some experts think that his teams do better in the national championship than you would otherwise expect. In the dataframe, the column `Izzo` is 1 if the team was coached by Tom Izzo in that year, and 0 if coached by someone else. (Izzo only coached one team in any given year.)

(1) (2 points) A logistic regression is shown in Figure 3. Why is this logistic regression appropriate?

The response variable `Final4`, whether or not the team advanced to the Final Four, is categorical (one point) with two categories, 1 (yes) and 0 (no) (the second point).

The word "this" in the question implies that you should distinguish from other kinds of logistic regression such as ordinal response (eg. the coal miners) and multinomial response (eg. the brand preferences), which you can most easily do by saying there are only two response categories: once you have done that, you have made the case for `glm` because with two response categories it doesn't matter whether they are ordered or not.

This being the first question on the exam, for the second point I am happy if you assert that the response has two categories without naming them, or if you name them (1 and 0) without explicitly saying that there are two of them. The best answer, of course, says both.

Extra: Even though the values of `Final4` are actually numbers, they are only labels for categories; the numbers themselves don't have any meaning. (See the next question for some more discussion of this.)

(2) (2 points) How do you know that the logistic regression in Figure 3 is predicting the probability that a team *does* advance to the Final Four, as opposed to the probability that the team does not do so?

The strictly correct answer is that in a logistic regression, if the response variable is coded as 0 and 1, the model will predict the probability of the category labelled 1.

The answer I expect you will give, and which I am perfectly happy with, is: the categories are arranged in "alphabetical" (here actually numerical) order, so that 0 is the baseline, and we predict the probability of the other one, 1, which corresponds to the team actually advancing to the Final Four.

Extra: Somebody asked on the course discussion board why sometimes we needed to use `factor` to create categorical variables, and the answer I gave there was that if you have a variable where the category levels are numbers, then you need to use `factor` to get it treated as categorical rather than quantitative (if it looks like a number, R will treat it as a number). In that light, you might be wondering why I didn't need a `factor(Final4)` on the left side of my logistic regression code. The answer to that is that `glm` with `family = "binomial"` is special: it will accept *either* a categorical variable with two levels (in which case the first one alphabetically is the baseline), *or* a numerical variable with values 0 and 1, in which case it predicts the probability of 1. Here, 1 corresponds to the event I would like to model the probability of, so I am happy to accept this default.

Medical people love to use 0 and 1 to indicate whether something of interest happened (1) or didn't (0), which is probably where this alternative came from. (You might recall that the same thing comes up in survival analysis, where 1 by default means "the event happened" in `Surv`.)

Extra extra: Sometimes you get this odd error from `glm`:

```
d <- tribble(
  ~x, ~y,
  0, "a",
  1, "b",
  2, "a",
  3, "c"
)
d.1 <- glm(y ~ x, family = "binomial", data = d)
```

```
Error in eval(family$initialize): y values must be 0 <= y <= 1
```

It is not actually true that the values of my `y` need to be between 0 and 1, but that was how `glm` was originally coded, so that's the error message you get (before you were able to use a two-level factor as the response). The meaning of the message is that there is something wrong with the response, which in this case is that it has three levels instead of

two. (`glm` checks to see that the response is the right kind of thing, and if it is not, it gives you this "canned" and not really appropriate error message.)

(3) (2 points) In Figure 3, why does the negative sign of the Estimate for `Seed` make practical sense?

The negative sign on the Estimate for `Seed` (-0.588) means that as `Seed` increases, the probability of advancing to the Final Four decreases (one point). A higher value of `Seed` corresponds to a weaker team, so it should be and is less likely that a weaker team will make it to the Final Four. (The second point.)

Make sure you "connect the dots": say what the negative sign means in terms of the model, and connect that to what it means in practice (from the words in the question).

(4) (2 points) Some predictions are shown in Figure 4. Describe the effect of the seed number as shown in that Figure.

As the seed number increases, holding `Izzo` fixed, the probability of advancing to the Final Four decreases. You said something like this in the previous question, so add some support to it. For example, when Izzo is not coaching, a number 1 seed has probability 0.363 of reaching the Final Four, but a number 5 seed has only probability 0.052, and teams with a higher seed number have an even smaller chance.

Compare at least two of the predicted probabilities shown in Figure 4, for different seed numbers holding `Izzo` fixed.

(5) (3 points) Assess the belief of the "experts" mentioned in the description of the data, using *both* Figure 3 and Figure 4. What do you conclude?

There are at least two things to talk about: whether there is a significant effect, and whether that effect is positive rather than negative.

In Figure 3, the `Izzo` term is significant (P-value 0.0017). This indicates that being coached by Tom Izzo has a significant effect on the probability of advancing to the Final Four. One point.

So far, this might be a positive or a negative effect. For the second point, note that the Estimate for `Izzo` is positive, indicating that a team coached by Izzo is *more* likely to reach the Final Four than a team coached by someone else, all else equal (that is, if the `Seed` is the same).

For the third point, look at the predictions in Figure 4 and quantify the effect of a team being coached by Izzo. For example, a number 5 seed has probability only 0.052 of reaching the Final Four when coached by someone else, but a much bigger probability 0.357 when coached by Izzo. (This is actually a *very* large effect.)

Use something from both Figures. You don't need to mention the positive Estimate if you use Figure 4 well enough to illustrate that the effect of being coached by Tom Izzo is a positive one.

Extra: this effect could be because Izzo actually is a better coach than people think, or because his teams look worse during the regular season than they actually are, or because they do play worse during the regular season and are somehow more highly motivated during the national championship.

(6) (2 points) A graph of the predictions is shown in Figure 5. Why does it make sense that the envelope around the blue curve is bigger than the envelope around the red curve?

The size of the envelope around the predictions indicates how accurately the probabilities are being estimated, and therefore the probability of reaching the Final Four is less accurately estimated for teams coached by Tom Izzo than for teams with other coaches. One point for describing what is happening.

As to why: Tom Izzo can only coach one team in a year, and there are 64 teams in the championship every year, so the other 63 teams are all not coached by Tom Izzo. Hence there is a lot more data from teams not coached by Tom Izzo than from teams that *are*. Therefore the probabilities for teams coached by other people are estimated (a lot) more accurately than for teams coached by Izzo. The second point for a convincing argument about why.

Extra: there's nothing stopping you from using Figure 5 to check your answers to the other questions: as seed number goes up, the probability of advancing to the Final Four goes down (and is almost zero for the "lowest" seeds with the highest seed numbers). The effect of being coached by Tom Izzo doesn't just increase the probability of reaching the Final Four, it makes it *much* bigger.

If you know the name Tom Izzo, you will realize that he is a basketball coach, and the sport we are talking about here is basketball. (The fact that the national championship is a 64-team knockout tournament might also have been a clue to you.) I didn't mention in the question that this is a basketball tournament, because for the question, it doesn't matter what sport it is. (The exam has another question that is also about basketball, and there, what sport it is actually matters.)

Anyway, according to Wikipedia, "his teams are often recognized for their rebounding prowess and defensive tenacity", which are the sort of qualities that will help a team win high-pressure tournament (basketball) games.

Using both sides of the brain

Eighty subjects were randomly assigned to one of two kinds of task (Visual and Verbal in column `Task` in our dataset), and they were instructed to report on the result in one of two randomly-chosen ways (also Visual and Verbal in column `Report`). The total Time needed to complete both the Task and the Report is recorded in the column `Time`, in seconds. A smaller time is better. The two types of task and the two types of report were designed to take the same amount of time to complete when done in isolation. Some randomly chosen rows of the data are shown in Figure 6.

According to psychological theory, visual and verbal activities are carried out by opposite sides of the brain. Thus, when the task and the report are of different kinds, the subject can use both sides of the brain to complete them at the same time, but when the task and the report are of the same kind, the subject's brain has to do them one after the other, taking a longer time. The data in Figure 6 were collected to investigate whether this psychological theory is correct.

(7) (2 points) In a two-way analysis of variance predicting `Time` from `Task`, `Report`, and their interaction, would you expect to see a significant interaction, according to the psychological theory? Explain briefly.

The fastest times should come when `Task` and `Report` are different, and the slowest times should come when they are the same. So what determines whether the time will be fast or slow is the *combination* of the two values, not either one of them by itself. This would imply a significant interaction.

Alternatively, if there is an interaction, the effect of `Task` will depend on the level of `Report`. According to the theory, if `Task` is `Visual`, the faster time should come when `Report` is `Verbal`, but if Task is `Verbal`, the faster time is when `Report` is `Visual`. Thus, if the theory is correct, the effect of `Task` *does* depend on the level of `Report`, and therefore we would expect to see an interaction.

Say something about what a significant interaction entails, and how the psychological theory implies that it will happen. This question is not asking about the boxplot in Figure 7, so you cannot refer to that.

(8) (2 points) Some boxplots are shown in Figure 7. How do these plots suggest that it would be better to use the log of `Time` in the analysis rather than `Time` itself?

All four of the distributions are skewed to the right or have upper-end outliers (or both: mention at least one of these two things; one point), and so would benefit from a transformation like log to bring the higher values down (the second point). I would accept something like "a log transformation is often useful for right-skewed data" (or "to make

right-skewed data more normal"), but mentioning what it will do to the high outliers or long upper tails is more insightful.

Extra: You may not use this evidence in answering the previous question, but it gives you a check on the likely interaction: when the task is verbal, there is not much difference in times between the two kinds of report, but when the task is visual, the times are a lot quicker if the report is verbal. If you see this now, this will give you a hint for later (when you discuss the support for the psychological theory), and at that time, you can investigate whether the apparent impression created by the boxplots is supported by P-values.

(9) (2 points) Some analysis is shown in Figure 8. What do you conclude from it, in the context of the data?

The interaction is (strongly) significant, with a P-value of 3.5×10^{-5}. One point. This means that the time taken depends on the combination of Task and Report, or that the effect of Report depends on the kind of Task. The second point.

(10) (4 points) Some further analysis is shown in Figure 9 and Figure 10. What do you conclude, in the context of the data? (You might find Figure 7 helpful.)

These are simple effects analyses, holding `Task` fixed, and investigating the effect of `Report` for the chosen level of `Task`. Hence:

- (from Figure 9) when the task is verbal, it makes no difference (P-value 0.40) to the time whether the report is verbal or visual.
- (from Figure 10) when the task is visual, there is a clear difference in time between the two types of report (P-value 4.3×10^{-8}). Look back to the boxplot in Figure 7, to see that the time is less if the report is verbal.

Two points for each of these. For the second one, the first point is for a conclusion based on the P-value, and the second point for saying which way the difference goes (which will help you answer the next question).

(11) (2 points) Does this data support the psychological theory given in the description of the data? Discuss briefly.

The best answer here is "yes and no".

The theory says that the time should be lowest when the Task and the Report are different.

That is certainly true when the task is Visual: the time was significantly less when the Report was Verbal.

However, when the task is Verbal, the theory predicts that the time should be lower when the Report is Visual, and that is not the case: there is no significant difference between the times for the two types of report, and in fact the median time when the report is Visual is slightly *larger*, in contradiction to the theory.

One point each for discussing each type of task and whether the time for each type of Report within that Task agrees with the theory or not.

Barley yields

Five varieties of barley (a grain) were grown in each of six locations in each of 1931 and 1932. The yield of barley (the total amount grown) was measured in each year. The scientists were mainly interested in whether the variety affected yield, allowing for any effect of location. The data, in dataframe `immer`, are shown in Figure 11. `Y1` contains the yields in 1931 and `Y2` contains the yields in 1932. The highest yield is best. The locations, in `Loc`, and the varieties, in `Var`, are indicated by the initial one or two letters of their names. For the varieties, their full names are: Manchuria (M), Svansota (S), Velvet (V), Trebi (T), and Peatland (P). I don't know what the full names of the locations are. There are 30 rows of data altogether.

(12) (2 points) Why will it *not* be possible to estimate an interaction effect between location and variety for these data?

In short, because there is no replication: there is only one observation per variety-location combination, and you know from your second statistics course that you need replicated observations in order to estimate an interaction. This is because, if you try, you end up with 0 degrees of freedom for error, and you then cannot test anything. You can tell here that there is no replication either by noting that there are 5 varieties, 6 locations, and 30 observations, and $5 \times 6 = 30$, so there is exactly one observation per variety-location combination, or by looking at the data in Figure 11 and seeing that each combination of variety and location only appears once. (I listed the whole data set to allow you to check this.)

The flip side of this argument was seen in the rats and vitamin B data in lecture. There, there were two sizes of rat and two diets with 28 observations altogether; you can check there that $28/(2\times 2) = 7$ rats were in each size-diet combination. The number of individuals per combination doesn't have to be always the same, though it makes things easier if it is, in a so-called "balanced" design. In a balanced design, you need at least two observations per treatment combination to be able to estimate an interaction, so in the rats and vitamin B data we had no trouble estimating an interaction.

(13) (2 points) Why is it appropriate to analyze these data using a MANOVA?

We have two response variables, the yield in 1931 (`Y1`) and the yield in 1932 (`Y2`). Hence we need a multivariate analysis of variance or MANOVA. If we only had one year's yield, it would be a regular ANOVA (in this case, a randomized block design).

Extra: you might also consider a repeated measures (or matched pairs) analysis, because we have the same thing (yield) measured at two different times. For this question, I chose to consider the yields as two different but possibly correlated response variables. (I think, for two time points, it doesn't actually make any difference.)

(14) (3 points) A MANOVA is shown in Figure 12. What do you conclude from this analysis?

The best answer looks at what the scientists wanted to know and says something like "there is an effect of variety on yield in one or both of 1931 or 1932 (or their combination), allowing for any differences between locations". Three points for that.

The locations are a "block" or "nuisance factor", like the engine sizes in the Octel filter example in lecture; we are taking for granted that there may be differences among them that we don't really care about. The reason for including locations in the analysis here is to get a better picture of differences among varieties *allowing for* any differences due to locations.

The second-best answer simply explains the results of the two F-tests in Figure 12, as a two-parter:

- the yields in 1931 or 1932 (or a combination of them) are not all the same among varieties
- the yields in 1931 or 1932 (or a combination of them) are not all the same among locations.

Or you can phrase these as "there are differences among the yields according to variety" and then "according to location". Two points for this kind of answer.

Be careful that your wording doesn't imply that *all* the varieties are different or *all* the locations are different; we don't know that yet. Or at all, even if we do something like a discriminant analysis.

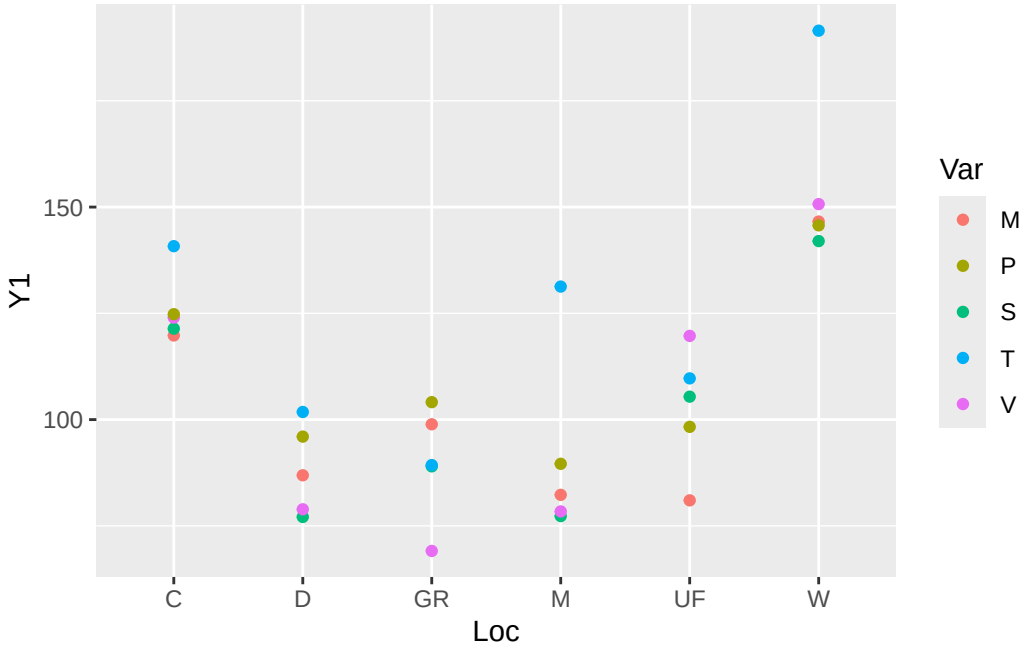
(15) (2 points) Figure Figure 13 shows graphs of yields `Y1` and `Y2`, separately, against location and variety. There are too many varieties for colour to distinguish them clearly, so I used a different technique (which I explain in the next question, so you don't need to ask about it here). Why did I put location on the x-axis rather than variety?

The best answer again considers what is of interest to the scientists: they want to compare varieties, and that is easiest if you put the blocking factor (location) on the x-axis and have the factor of interest (variety) all gathered together above each location, so that you can compare varieties (here indicated by letters) within each location. ("Within" locations or "for each" location are good words to use here.) Two points for this.

The second-best answer says that there are more locations (six) than varieties (five), so put locations on the x-axis because it is easier to distinguish five colours than six. One point for this less insightful answer.

As it turns out, the five colours are not that easy to distinguish either, especially under exam conditions:

```
ggplot(immer, aes(x = Loc, y = Y1, colour = Var)) + geom_point()
```



(this is the `Y1` plot), and so I decided to do something different for my plot (that I explain below). As a side note, these are points rather than grouped boxplots because there is only one observation of `Y1` per variety-location combination, and thus the boxplots, based on one observation each, look rather silly.

(16) (2 points) The package `ggrepel` contains a function `geom_text_repel` that places text as close to a point as possible. The text that appears on the plot comes from

the variable in `label`. Some of the points are close together on the plot, but you may assume that if a letter is further up the page, its corresponding point on Figure 13 is further up the page as well. Why do you think the significance (or not) of the MANOVA for variety in Figure 12 makes sense according to Figure 13? Explain briefly.

The MANOVA for variety was significant (P-value 0.035), as we saw earlier. The significance of the MANOVA implies that there is some more or less consistent pattern in how the varieties compare. Observe something that happens most of the time, such as that variety T (Trebi) is usually best (at the top, high yield) for both years, and variety S (Svansota) is usually near the worst (at the bottom, low yield) for both years. (Observing one of these things, or something equivalent, is enough.)

The pattern is not completely consistent, which shows up in the P-value being small but not *very* small. (For example, Trebi in 1932 at location GR was actually second *worst*, but this is more than counterbalanced by Trebi being clearly the best at location M in 1931 and at location W in both years.)

If you thought that there was no significant effect of variety, you need to make the case that there is *no* consistent pattern among the varieties in Figure 13, which you might try to do by picking a variety or two like `V` (Velvet) that is sometimes near the top and sometimes near the bottom.

Cross-country ski grip

In cross-country skiing, the fastest skiers are the ones who can generate the most upper-body power. This might be influenced by how they grip the ski poles. There are three standard ways in which a cross-country skier might grip the poles, called Classic, Integrated, and Modern. These are shown in Figure 14, in column `grip.type`. 12 skiers (labelled in `id`) were randomly assigned to use one of the grip types. The researchers were concerned about a possible effect of practice, so they measured the upper body power (`UBP`) generated by each skier on three separate occasions, labelled 1, 2, and 3 in column `replicate`. The dataframe is called `grip`.

(17) (1 point) How, specifically, do you know that a repeated measures analysis will be necessary here?

For each skier, the upper-body power was measured for the same individual (skier) on *three separate occasions* (the three different time points in `replicate`). Enough of that for the one point.

The three different grip types are not relevant here, because each skier only uses one of them all the way through; this is a between-subjects factor in the jargon.

Extra: you might be wondering why the skiers didn't get to try out all three of the grip types. In actual fact, they did (making this a doubly-repeated measures or a crossover design, since the same thing was actually measured for all combinations of replicate and grip type), but we haven't analyzed that kind of thing in class, and therefore I am not going to give you that on an exam. So I simplified things by having each skier only try one of the grip methods, and then we have one between-subjects factor (grip type which is the same for all times for each skier, but which differs from one skier to another), and one within-subjects factor (replicate), the same way that we are used to seeing it.

(18) (3 points) An interaction plot is shown in Figure 15, and a spaghetti plot is shown in Figure 16. Using both Figures, do you think there will be a significant interaction between grip type and time, or not? Explain briefly.

- From the interaction plot, you can say either that the lines appear to be not parallel (pointing to an interaction) or that they are close to parallel (no interaction).
- From the spaghetti plot, it is clear that there is a lot of variability, even within the same grip type.
- Hence, there is too much variability to be able to conclude convincingly that there will be a significant interaction between grip type and time. Alternatively, there is not a consistent pattern of the different-coloured traces in Figure 16 having different patterns over time.

One point for making a reasonable conclusion from the interaction plot (either one of the two things in the first bullet point), and the other two points for getting to the third bullet point (either via the second one or not). The best answer uses the piece in the Alternatively sentence, since that is precisely the point, but I will accept a "there is too much variability in the spaghetti plot to be able to demonstrate a significant interaction" since that shows the right kind of thought process.

It is perhaps better to be more circumspect about asserting an interaction from Figure 15, at least until you have looked at the spaghetti plot, since if you think there actually will be a significant interaction from the interaction plot, you should be ready to change your mind when you see the spaghetti plot.

Extra: "circumspect" means literally "looking around"; in this sense, it means to be aware of all the issues before you come to a decision (about a significant interaction in this case).

Extra extra: you might remember those science TV shows you watched in school. A British comedian made a series of satires on these, called Look Around You, which you can find on YouTube. My favourite is the Maths one.

(19) (2 points) Some code is shown in Figure 17. It was necessary to run this code before running the analysis in Figure 18. Why, specifically?

The original data in `grip` is "long", with the three measurements for each skier on separate lines. The analysis in Figure 18 requires all the measurements for one skier on the *same* line, and therefore we need to pivot the original long data wider.

"We have to make the long data wider" is true as far as it goes, but not specific enough. One point for this kind of answer. To get the second point, say something about why you need "wider", specifically all the measurements for the same skier on the same line.

(20) (4 points) The analysis in Figure 18 created `grip.1`, and output from `summary(grip.1)` is shown in Figure 19. The people who collected the data wanted to know whether there was any difference in upper body power between the types of grip, whether there was a practice effect, and whether the upper body power depended on the combination of grip type and replicate. What does Figure 19 tell you about the answers to these questions? Describe your process clearly.

This is a repeated measures analysis, so first test for sphericity. Both P-values are 0.68, so there is no problem with sphericity, and therefore we can ignore the adjusted P-values below that. One point.

Then test the interaction between grip type and time, from the top table. This is also not significant (P-value 0.918). Therefore there is no effect of the combination of grip type and replicate (beyond the main effects of grip type and replicate, which we test next.) The second point. You need to test the interaction *before* testing the main effects, because if the interaction had been significant, that would have been the finding (you cannot make conclusions about main effects in the presence of a significant interaction; you would need to do something like simple effects). Minus one if you test the interaction other than first after testing the sphericity. (There is no guarantee that the questions of interest to the people who collected the data will be answered by you in the same order.)

Normally, we would take out the interaction and re-fit. In a repeated-measures analysis, though, we cannot take out time because of the way it pervades the whole analysis, so we proceed to test the main effects from the same table we tested the interaction from (the top one).

There is no effect of time (replicate), P-value 0.970. Thus there is no practice effect (which would have shown up as something like an increase over time). The third point. (The best answer here links a practice effect to an effect of time.)

There is, however, a (strongly) significant effect of grip type (P-value 0.00012). The upper body power *is* different among at least some of the grip types. (This analysis does not tell us how.) The fourth point.

If you don't test the main effects, you can expect to earn only two (total) for the question, but if you say *why* you are not testing them (for example, something along the lines of "we need to take out the interaction"), I was more sympathetic, and you might earn 3 instead.

I undoubtedly needed to give you more space for your answer on this one, but then, that's why there was page 13.

(21) (2 points) From Figure 15 and Figure 16, what seems to be the reason for any significant results you found in the previous question? Explain briefly.

In the previous question, I only found one significant effect, that of grip type. (You might infer for yourself that there are not many significant effects, because this question is only two points.)

In the interaction plot in Figure 15, the Integrated grip was associated with clearly higher upper body power than the other two grips (which did not seem to differ from each other). To confirm this, look at the spaghetti plot in Figure 16: the green traces for Integrated are mostly at the top, even given the considerable variability, while the red and blue traces for the other two grip types are all mixed up, mainly at the bottom of the plot.

So I think the reason for the significant effect of grip type is that Integrated was associated with higher upper body power than the other grip types (which were not different from each other). For the two points, say this, along with some kind of justification for it. I don't think you can infer anything more than "Integrated is best", because on the spaghetti plot the other two grip types are too mixed up. Your justification ought to use a word like "usually" to describe how the UBP for the Integrated grip is mostly but not always at the top of the spaghetti plot, or have some other way of saying that there is variability shown on the spaghetti plot but that Integrated tends to be at the top.

You need to be clear that you have referred to *both* figures, which you can most easily do by talking about means (the interaction plot) and variability (the spaghetti plot).

If your answer says something about other effects being significant, expect me to check back to the previous question for consistency; if you can make a good argument for your answer here, even though your previous conclusions were wrong, you can get credit here.

Extra: You don't need to talk about the non-significant effects, but the spaghetti plot confirms them as well:

- the traces on average go straight across for all grip types (there is no tendency for the traces to have a consistent pattern over time like going uphill, which you would expect if there was a practice effect). That is to say, some of the traces go up, some go down, some go straight across, and they average out to straight across.

- the non-pattern over time is the same for all grip types, so we would not expect to see any kind of interaction effect. If there had been a pattern like an increasing trend over time for some of the grip types but not others, that would have been indicative of an interaction effect between grip type and time.

NBA players 2015

Basketball players play in one of three positions: Centre, Guard, or Forward. The NBA basketball league compiles statistics on its players. Do players in different positions tend to have different statistics? In basketball, players can score in one of three ways: a field goal (2 points), a 3-point field goal from outside of a line on the court (3 points), and a free throw (1 point), awarded after a foul (illegal play). After a field goal attempt is missed, the player who catches the ball is awarded a rebound. A player who intercepts a pass by a member of the opposing team is credited with a steal.

In the data shown in Figure 20, each player's name and their position is shown. Next are their success rate (number made divided by attempts) at field goals, three-point field goals, and free throws, and finally the number of rebounds, steals, and fouls committed per game. (Questions continue on the next page)

(22) (2 points) Some analysis is shown in Figure 21. What do you conclude from this analysis?

The null hypothesis is that all the statistics have the same mean for each of the positions. With a P-value of less than 2.2×10^{-16}, this null hypothesis is rejected, and therefore there are some statistics (or combinations of statistics) that differ among the positions.

(23) (2 points) Some further analysis is shown in Figure 22. Why is this a helpful analysis, given what we know so far?

We concluded in the previous question that there are some differences on at least some of the variables to find among the positions. The discriminant analysis will help us find where those differences are (and how well the player positions actually can be distinguished).

(24) (2 points) In Figure 22, what are the two most important of the original variables in LD1? Explain briefly.

Field goal percentage (`FGPct`) and Steals. They have the largest coefficients in size (the first is negative and the second is positive).

(25) (2 points) Based on your answer to the previous question, what would make a player have a very *negative* score on LD1? Explain briefly.

A high field goal percentage and a low number of steals.

(We will try to give the points based on your answer to the previous question.)

(26) (3 points) Figure 23 shows the results of some further analysis. What do you learn from this table specifically? (You have some choices here; for full credit, find **two** distinct insightful choices.)

This is a table showing the actual and predicted positions of all the players. (One point, which you can also get without saying this if it is clear that you understand it.)

For the second and third points, say something that shows that you have learned something about the quality of classification from this table, for example:

- the majority of the players are correctly classified in terms of position (the numbers down the top-left to bottom-right diagonal are mostly large, or the numbers off this diagonal are mostly small)
- most of the actual centres are correctly predicted to be centres (18 out of 24)
- almost all the actual guards are correctly predicted to be guards (83 out of 88)
- half of the actual forwards are misclassified (31 out of 62)
- a lot of the actual forwards are misclassified as guards (22 out of 62)
- no centres got mistaken for guards, and only one guard got mistaken for a centre

or another possibility that offers some insight about how good the classification is.

If you offer more than two insights and one of them is wrong, you risk losing points. The key thing is knowing when to stop.

There were a lot of ways to get three points here, unless you made any mistakes (such as getting the observed and predicted the wrong way around).

(27) (2 points) A plot is shown in Figure 24. Using your conclusions from the previous question, describe whether this plot leads you to the same conclusions as in that question. Explain briefly. (If you are having trouble distinguishing the three colours, an invigilator can tell you the colour of a point you indicate. I have tried to make the graph colourblind-friendly.)

I'm calling the colours green, red, and blue (in the same order as the legend).

To take my suggestions in order (use yours):

- the players (coloured dots) are somewhat distinguishable, but only somewhat (there is a lot of overlap between the forwards and the guards, red and blue, especially)
- the centres (green, on the left) are mostly distinguishable from the others, but there are some red dots mixed in with the other colours

- the guards (blue) are almost all on the right
- the forwards (red) are all over the plot, a lot over on the right with the centres
- the centres (green) and guards (blue) are almost all distinct on the plot: centres on the left, guards on the right.

It is likely that if you say something sensible here, I won't check back to your question 26 for consistency (that is to say, if you make two sensible observations here, you'll get the two points). But if you say something odd here, I will check back to see if you were consistent with what you said before.

Extra: I surreptitiously drew this graph with different colours than usual, to make the colours easier to distinguish and more colourblind-friendly. On the default graph, the forwards would have been green and the guards blue, and the green would have been darker and less distinguishable from the blue.

(28) (2 points) In Figure 24, what is the most important *one* thing you can say about values on the original variables for the players at the bottom of the plot? Explain briefly.

These players have a low score on LD2. Going back to Figure 22, the most important contributor to LD2 is field goal percentage (largest coefficient in size), with a positive coefficient, so the players at the bottom of the plot have a *low* field goal percentage.

I asked for the most important one thing to make the question easier to mark fairly. In practice, you might also mention that LD2 depends on three-point field goal percentage (negatively), and maybe free-throw percentage (positively), which you could then investigate by identifying the players at the bottom of the plot and checking out their original data. (I didn't ask you to do this.) On the exam, you can name a second thing if you want, but if the second thing is not three-point field goal percentage, you risk losing marks.

Extra: of course, if I had given you a biplot, you could have judged this from the way the arrows were pointing, but I didn't, so you have to work it out yourself.

(29) (2 points) What are *two* pieces of evidence you have seen so far that the second linear discriminant is in fact not worth considering?

- In Figure 22, LD2 has a low proportion of trace
- In Figure 24, the groups are mostly distinguished left and right (in the direction of LD1), and not at all up and down (in the direction of LD2).

If you can come up with something else, go for it, but I think those are the key two things.

(30) (2 points) Look at Figure 25. For Kawhi Leonard, was his position correctly predicted? Was the prediction a close decision or a clear decision? Explain briefly in each case.

A point each:

- Kawhi Leonard is actually a forward (in `position`), but he was predicted to be a guard (in `class`), which is incorrect. It is not enough to say "not correctly predicted": you need to say how you know.
- According to the posterior probabilities, he is certainly not a centre (correct), but has posterior probability 0.349 of being a forward and 0.651 of being a guard. I would call this a clear (wrong) decision, because these probabilities are not close,[1] but you can also say that it is not that unlikely that he is a forward, so the decision is not badly wrong. (As ever, the point for the reasoning, not for the decision.)

Extra: on Figure 24, Leonard is presumably one of those forwards (red dots) over on the right, in with the guards (blue dots), which would explain why there is a fair bit of uncertainty over which of those two positions he actually plays. This suggests (according to Figure 22) that he has a low field goal percentage, a high 3-point field goal percentage and/or a high number of steals. You could (if you had the original data) check to see whether that was correct. If it is, it might be because he tends to shoot from further away from the basket than is typical for a forward (a low FG percentage), or be better defensively than is typical for a forward (a high number of steals).

Hearing test

One hundred males of age 39 with no history of hearing disorders did a hearing test. Each individual is exposed to signals of varying frequencies with an increasing loudness until the individual indicates that they can hear the signal. Each individual was exposed to signals of 500, 1000, 2000, and 4000 hertz (frequency) in each ear. The loudness values were recorded in columns with names starting with L or R (left or right ear) followed by the frequency. The loudness was measured on a log scale, so some of the values are negative. A larger loudness value indicates that the individual had more difficulty hearing the signal. Each row of the data shown in Figure 26, in dataframe `hearing`, is for a different individual, identified in `Sl_No`. The dataframe `hearing0` contains all the columns of `hearing` except for `Sl_No`.

Interest was in what distinguished the individuals' ability to hear amongst the eight variables recorded.

[1] I was prepared to tolerate a call that these probabilities are "close", not because I really agree with it, but because you are at least making a comparison of the right things, which was the goal of the question.

(31) (2 points) A screeplot is shown in Figure 28. Someone suggests to you that four principal components is a good number. Do you agree or disagree? Explain briefly.

Look for an elbow that is far enough down the mountain. I see elbows at 2, 4, and 5 (the last one is a small one, but it seems that everything beyond it is scree). I think the elbow at 2 is too far up the mountain. So you have two relevant options, remembering to *subtract one* to get the number of principal components:

- elbow at 4: 3 principal components, and therefore you disagree
- elbow at 5: 4 principal components, and therefore you agree.

Make sure you get the logic right. Only one point if you fail to subtract one.

Extra: you can check your answer by looking back at the standard deviations at the top of Figure 27. This supports both answers above (these things are rarely clear-cut): the third standard deviation is clearly bigger than the fourth (supporting three components), and even though the fourth one is not that much bigger than the fifth, the standard deviations beyond that don't decrease very fast (suggesting random variation or "scree"), supporting stopping at four components.

(32) (1 point) A principal components analysis is carried out as shown in Figure 27. The first component is often a measure of "overall size". How do you know that this is the case here?

The loadings of the first component are all about the same size, so that the scores on the first component are more or less an average or total of all the variables. Another way to approach this is to say that an individual will have a large score on component 1 if they have large values on *all* the variables (with the implication that you need to consider all of them, not just some).

Nothing more dramatic needed for one point.

This question has nothing to do with standard deviations, and in particular nothing to do with the first component having the largest one (it always will, regardless of whether the first component contains all the variables or just some).

(33) (2 points) Looking at Figure 27, which two of the original variables are the most important in component 2? What, if anything, do they have in common? Explain briefly.

Look for the two largest loadings in size (that is, ignoring the sign). These are `R4000` (-0.514) and `L4000` (-0.474). One point. What these have in common is that they are the same frequency (4000 Hz), or that they are both high frequencies. The second point.

Not the two most positive loadings `L500` and `R500`. That shows a lack of understanding of what principal component loadings are for. I suppose, if you pick those, and you say "the same frequency", you get the second point without getting the first one. If this happens to you, count yourself lucky.

(34) (4 points) A graph of scores on the first two principal components is shown in Figure 29. Find individual 66 on this graph. Using Figure 26 and Figure 30, explain briefly why it is not surprising that this individual appears on Figure 29 where they do. (Or "is surprising", if that's what you think.) You only need to consider the one most relevant component in this question.

Individual 66 is near the bottom of the graph, with a component 2 score of about -3. One point for enough of that. (You don't need to mention that component 1 is close to zero, though no harm if you do here. The last sentence of the question tells you that you don't need to consider component 1 the rest of the way.)

Hence, focus on component 2. We said earlier that component 2 mainly[2] depends (negatively) on `L4000` and `R4000`, so an individual with a negative score on component 2 is *high* on these two variables. The second point.

Use the data in Figure 26 to see that individual 66 has an `L4000` value of 60 and an `R4000` value of 65 (they are in the fourth row). Then use Figure 30 to see where those two values stand relative to the other values of these two variables. In both cases, they are (well) above Q3 and close to the largest. The third point.

The fourth point is for saying that these high values for these variables are therefore not surprising. Connect the dots.

If you said earlier that component 2 depends positively on `L500` and `R500`, and then follow through thus:

- component 2 will be low if `L500` and `R500` are both *low*
- individual 66 has values -10 on both those variables, which are (tied for) lowest (see the Extra)
- these values are therefore not surprising

then you can get the four points here even though you didn't get any on the previous question. Expect the grader to check back for consistency if you give an answer like this here.

Extra: I have to admit that I cheated a bit to make this question work for you. When you look at the component 2 loadings, there are others that are almost as big in size as `L4000` and `R4000`, such as `R500` and `L500` (positive). So, the component score 2 could also

[2] But see the Extra.

have been negative because R500 or L500 were unusually *small* (as it turns out, individual 66 had tied-for-smallest values of −10 on these two variables as well). So, in this broader view, component 2 is a contrast between values on the highest frequencies and values on the lowest. But I wanted to make something where the logic was clean for you. In the real world, things get confusing fast, but on an exam, if you can follow the logic, I want to make sure you go in the right direction.

In the display of the data in Figure 26, the values are actually shown in descending order by L4000, and I could be pretty sure that individual 66 would be on there somewhere.

This is actually the kind of data where factor analysis (that I didn't talk about much in class) comes out cleaner, because of the rotation designed to make the factors easier to interpret. I'll use four factors as suggested by my scree plot:

```
hearing.2 <- factanal(hearing0, 4)
hearing.2$loadings
```

```
Loadings:
      Factor1 Factor2 Factor3 Factor4
L500  0.659   0.126   0.103   0.502
L1000 0.933   0.116   0.220   0.251
L2000 0.370   0.308   0.613
L4000 0.126   0.980   0.114
R500  0.322                   0.941
R1000 0.537   0.126   0.299   0.501
R2000 0.112   0.189   0.970
R4000         0.682   0.239

                Factor1 Factor2 Factor3 Factor4
SS loadings       1.867   1.606   1.537   1.474
Proportion Var    0.233   0.201   0.192   0.184
Cumulative Var    0.233   0.434   0.626   0.810
```

Here, factor 2 is rather more clearly the highest frequencies, and factor 1 is dependent on the lower ones.

Extra extra: what happens with hearing is that as you get older, the highest frequencies are the ones you have most difficulty hearing (and thus the required loudness values are greater). Individual 66 had no problem hearing the lowest frequencies, but had a lot more difficulty with the highest frequency, in that the signal needed to be pretty loud for them

to hear it. In terms of hearing, component 1 could be described as "good hearing overall" (low values) vs "bad hearing overall" (high values), while component 2 measures how well the individual can hear high frequencies relative to how well they can hear low ones (badly, in individual 66's case).

In music, a "concert A", 440 Hz, is the note that musicians tune to before a concert (the A above middle C). If you arrive early for a performance (or rehearsal) by our concert band, you will hear the "concertmaster"[3] play this A and everybody makes sure their instrument is tuned to it, so that the whole band will be in tune through the concert. (These days, presumably the concertmaster uses an app to tune their instrument to a concert A beforehand. In the old days, the concertmaster would use a metal tuning fork, which, when tapped, would produce a pure sound of the correct frequency, and then they would tune their instrument to it by ear.)

Going up by an octave doubles the frequency, so the A an octave above concert A is 880 Hz. There are 12 semitones in an octave, and going up a semitone increases the frequency by a constant multiple,[4] so that if you are if you are n semitones above concert A, the frequency is $440 \times 2^{n/12}$. (n can be negative, if you are that many semitones *below* concert A.) Hence the B just above concert A, two semitones up, has a frequency of

```
440 * 2^(2/12)
```

```
[1] 493.8833
```

and the E below it, five semitones down, has a frequency of

```
440 * 2^(-5/12)
```

```
[1] 329.6276
```

Source. The highest frequency in our data, 4000 Hz, corresponds to somewhere between B and C, just over three octaves up from concert A.

[3] In a symphony orchestra, this is the player in the first violins that sits nearest the conductor, but I forget who it is in the concert band.

[4] This is what musicians mean by "equal temperament".

Chest pain

Each of 10186 New Zealand adults was asked their age and whether they experienced any pain or discomfort in their chest over the last six months. If yes, they indicated whether it was on their left and/or right side of their chest. The data are shown in Figure 31. For each combination of age group and whether or not pain was experienced on the left side or the right side, the frequency `n` is shown. The researchers were interested in whether age had any influence on the presence of any sort of chest pain.

In the data, `(30,40]` means "strictly greater than 30 and less than or equal to 40", so that a person aged exactly 40 would be in this age group and not the `(40,50]` age group.

(35) (3 points) Some log-linear modelling is shown in Figure 32. Describe my process, and say why I stopped where I did.

- I fitted a model containing all the associations (up to the three-way association). In this model, the three-way association is not significant (P-value 0.15) so I removed it, resulting in the model `chest.2`.
- In `chest.2`, the `age:right_side` association is not significant (P-value 0.91), so I removed that as well, resulting in model `chest.3`.
- In `chest.3`, both associations are significant (P-values both less than 0.001), so we stop there.

Marking guideline: a point for enough of each of those bullets to show that you understand what I am doing. It is good to use the word "interaction" or "association" to show that you know what is going on.

(36) (2 points) Why do the researchers care more about `age:left_side` than `left_side:right_side`?

In a log-linear model, interactions between effects correspond to associations between them. In this scenario, age is an explanatory variable, and `left_side` and `right_side` are responses. The researchers want to know whether age is associated with chest pain, and so the significant `age:left_side` association is something they care about. The `left_side:right_side` association has nothing to do with age and so is of less interest (but not no interest; we interpret this below.)

An answer like "the researchers care about associations between age and chest pain, and `age:left_side` is the only one of those here" is about the minimum I would accept for two points. If you write less than that, one point for "want associations between age and one or other chest pain column" and one point for saying "`age:left_side` has `age` in it" and/or "`left_side:right_side` does not have `age` in it". Copying the last sentence of

the first paragraph of the data description (above question 35) does not demonstrate much understanding; the point is to show me you know that the `age:left_side` term tells us about the association between age and left-side pain; the significance of that term means that there *is* an association between left-side pain and age.

Another approach you might take is to say that age is explanatory (in the context of this study) and left- and right-side pain are responses. Then you can say that we mostly care about associations between explanatory and response variables, not between two things that are both responses.

Some people pointed out that the heart is on the left side, and so left-side chest pain might indicate heart problems. This is interesting (and not something I had thought of) but only tangentially relevant; the important thing is the association between age and whatever indication of chest pain is still in the model.

(37) (2 points) A graph is shown in Figure 33. In the code above the graph, why did I use `x` and `fill` as shown, rather than having the variables the other way around?

Because age is explanatory, and whether or not there was pain in the left side is a response (outcome). Doing it this way, I can see what fraction of people reported left-side pain in each age group. (If I had done it the other way around, I would have had to interpret it as "out of the people who had left-side pain, how many were in each age group" which makes no sense logically.)

One point each for "age is explanatory" and "left-side pain is response", or something that clearly implies that (like "effect of age on incidence of left-side pain").

A generous one point for "there are more categories of age than of left-side pain"; this would be relevant if the two variables had the same role (eg. were both explanatory), but logically left-side pain being a response and age being explanatory is more important than that, and so should be mentioned first. If there had been five categories of pain (like, "none, low, moderate" etc.) and only two age groups, it would still have been correct to use `x` and `fill` as shown, and we would have had to look at the five colours to assess whether the overall pain level was better or worse for the older age group. Presumably in that case the colours for pain level would have been ordered from highest to lowest (or would have been reordered so that they came out that way on the graph) so the comparison would have been easier than if they were in "random" (eg. alphabetical) order.

(38) (2 points) Interpret the graph in Figure 33, in the context of the data.

Say at least two things out of:

- the presence of left-side chest pain is greatest for those aged over 70
- the presence of left-side chest pain is least for those aged between 50 and 70

- the age group with the second-most left-side chest pain is 30–40
- I would expect the presence of left-side chest pain to increase with age, but it actually does not (it increases, decreases, then increases again).

or anything else I see as relevant.

Part of the purpose of this question is for you to judge what the most relevant things are and how many of them to talk about. You don't need to go overboard, and if you say so many things that some of them are wrong, you can expect to lose points. (There is a small hint that two things might be enough: the question is out of 2 points.)

There *is* a (strongly) significant effect of age on left-side chest-pain, so that there are differences to talk about, but what we are testing is a null hypothesis that the incidence of pain is the same for all age groups vs. an alternative that the null is not true somehow. The `age:left_side` association will be significant if there is *any* difference between the age groups that is bigger than chance, which includes an irregular pattern like this one. If we had had the actual numerical ages, we would have had to fit some kind of logistic regression (perhaps having a combined `pain` response that is none, left side only, right side only, both sides and then using `multinom`), and then, by treating `age` as quantitative, we *would* be testing for a change in probability that is consistent with an effect of increasing age. But that is not the case here: like ANOVA, we are testing for *any* difference among the age groups.

Extra: truncating the y-axis (only displaying it from 0 to 0.2 rather than 1) makes it easier to see the effect of age. Like the severe cases in the Israel COVID data on worksheet 12, the number of people experiencing left-side chest pain was never very large (around 10%), so truncating the y-axis made the effect of age easier to see. I didn't draw your attention to this, so I didn't penalize answers that said "about 50-50 for left-side chest pain vs. not", since that's what it looks like on first glance.

(39) (2 points) Interpret the graph in Figure 34, in the context of the data.

People who have experienced chest pain in the left side are more likely to have experienced it in the right side as well. Those who have *not* experienced chest pain in the left side are very unlikely to have experienced it in the right side either.

A decent fraction of both of those for the points.

Or, chest pain in the right side is never very common, but it is more common in those who have experienced chest pain in the left side. (The issue is to *compare* pain in the right side between those who do and do not have pain in the left side.)

Extra: this interpretation is mildly interesting in its own right (chest pain on the two sides of the chest seems to go together), but it's not what the researchers really cared about, which was the effect of age.

You might be wondering why `age:right_side` turned out to be very far from significant. I suspect it's not so much that age is worthless in predicting right-side pain, but because left-side and right-side pain are also associated, you don't really need to think about the association between age and right-side pain as well in that it has nothing to add (in the same way that a non-significant explanatory variable in multiple regression might be useful by itself, but it has nothing to add over what's already there). Here, I think you could use age to assess whether a person will have left-side pain, and then use the association between left-side and right-side pain to assess whether they will have right-side pain.

Added later: this may also be the place where the heart being on the left side is relevant; chest pain on the right side may not have anything to do with age, but on the left side it may indicate heart problems (which would then make you wonder why it doesn't increase with age).

If you need any more space, use this page, labelling each answer with the question number it belongs to.

Figures

```
library(tidyverse)
library(MASS, exclude = "select")
library(marginaleffects)
library(broom)
library(car)
library(survival)
```

Figure 1: Packages

	Year	Seed	Final4	Izzo
1	2001	6	0	0
2	1998	1	1	0
3	1988	11	0	0
4	1991	16	0	0
5	2006	5	0	0
6	2001	1	1	1
7	2001	12	0	0
8	1990	8	0	0
9	2007	14	0	0
10	1985	13	0	0
11	1998	8	0	0
12	1995	14	0	0
13	2001	11	0	0
14	1988	9	0	0
15	2001	8	0	0
16	2005	15	0	0
17	2006	4	0	0
18	1996	7	0	0
19	1995	13	0	0
20	1993	11	0	0

Figure 2: Izzo data (20 randomly chosen rows)

```r
izzo.1 <- glm(Final4 ~ Seed + Izzo, data = FinalFourIzzo, family = "binomial")
summary(izzo.1)
```

```
Call:
glm(formula = Final4 ~ Seed + Izzo, family = "binomial", data = FinalFourIzzo)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.02776    0.19967   0.139  0.88942
Seed        -0.58809    0.05909  -9.953  < 2e-16 ***
Izzo         2.32441    0.73971   3.142  0.00168 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 778.06  on 1663  degrees of freedom
Residual deviance: 519.56  on 1661  degrees of freedom
AIC: 525.56

Number of Fisher Scoring iterations: 8
```

Figure 3: Izzo data logistic regression

```
new <- datagrid(model = izzo.1, Seed = c(1, 5, 9, 13), Izzo = c(0, 1))
cbind(predictions(izzo.1, new)) %>%
  select(Seed, Izzo, estimate, conf.low, conf.high)
```

```
  Seed Izzo     estimate     conf.low   conf.high
1    1    0 0.3634724918 0.2963916353 0.436320736
2    1    1 0.8537203338 0.5735880449 0.962008512
3    5    0 0.0515307373 0.0370854692 0.071186657
4    5    1 0.3570333767 0.1174754651 0.698471533
5    9    0 0.0051427167 0.0024183140 0.010902808
6    9    1 0.0501821692 0.0112579158 0.196888042
7   13    0 0.0004915945 0.0001470338 0.001642276
8   13    1 0.0050017290 0.0008634866 0.028408518
```

Figure 4: Izzo data predictions

```
plot_predictions(izzo.1, condition = c("Seed", "Izzo"))
```

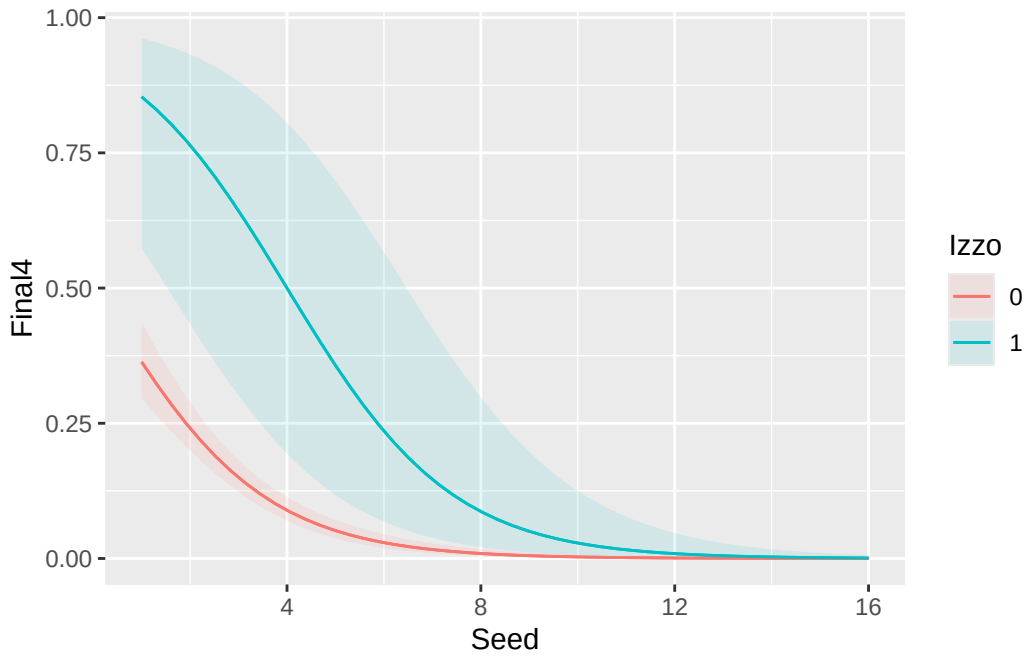


Figure 5: Graph of predictions for Izzo data

```
     Task Report  Time
1  Verbal Visual 17.23
2  Verbal Verbal 10.90
3  Verbal Visual 13.85
4  Visual Visual 11.60
5  Verbal Verbal 12.31
6  Verbal Visual 18.45
7  Verbal Visual  8.44
8  Verbal Visual 15.48
9  Verbal Verbal 15.85
10 Visual Verbal  7.77
```

Figure 6: Brain side data (10 randomly chosen rows)

```
ggplot(VisualVerbal, aes(x = Task, y = Time, fill = Report)) + geom_boxplot()
```

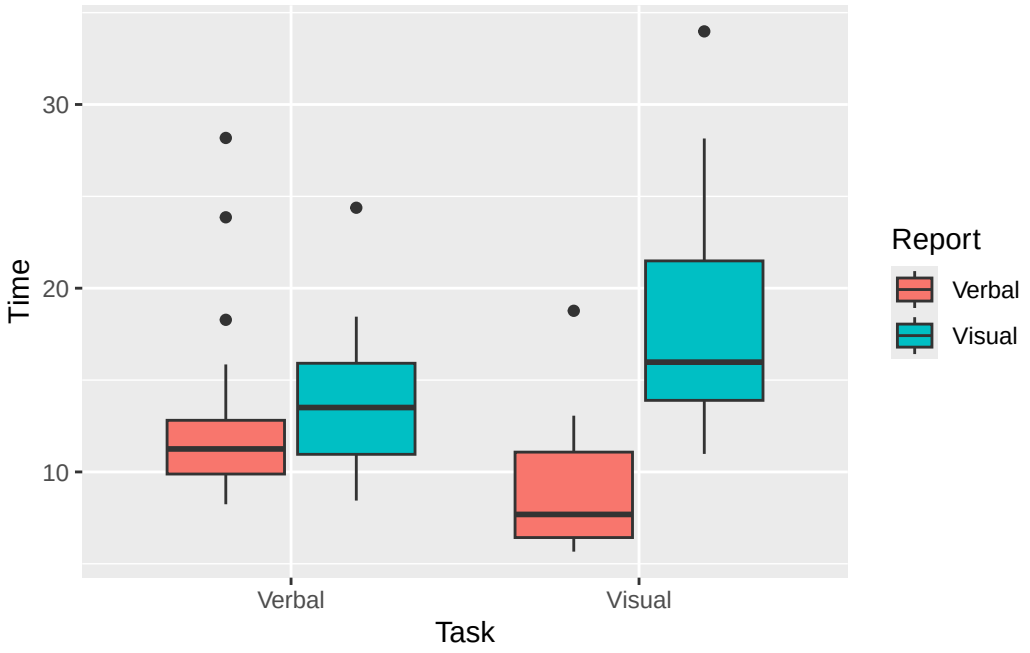


Figure 7: Brain side boxplots

```r
vis.1 <- aov(log(Time) ~ Task * Report, data = VisualVerbal)
summary(vis.1)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Task          1  0.033  0.0335    0.33    0.567
Report        1  3.136  3.1360   30.92 3.83e-07 ***
Task:Report   1  1.963  1.9630   19.36 3.49e-05 ***
Residuals    76  7.708  0.1014
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8: Brain side ANOVA

```r
VisualVerbal %>% filter(Task == "Verbal") -> verbals
verbals.1 <- aov(log(Time) ~ Report, data = verbals)
summary(verbals.1)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Report        1  0.068 0.06839   0.722  0.401
Residuals    38  3.599 0.09470
```

Figure 9: Brain side further analysis part 1

```r
VisualVerbal %>% filter(Task == "Visual") -> visuals
visuals.1 <- aov(log(Time) ~ Report, data = visuals)
summary(visuals.1)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
Report        1  5.031   5.031   46.52 4.3e-08 ***
Residuals    38  4.109   0.108
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10: Brain side further analysis part 2

	Loc	Var	Y1	Y2
1	UF	M	81.0	80.7
2	UF	S	105.4	82.3
3	UF	V	119.7	80.4
4	UF	T	109.7	87.2
5	UF	P	98.3	84.2
6	W	M	146.6	100.4
7	W	S	142.0	115.5
8	W	V	150.7	112.2
9	W	T	191.5	147.7
10	W	P	145.7	108.1
11	M	M	82.3	103.1
12	M	S	77.3	105.1
13	M	V	78.4	116.5
14	M	T	131.3	139.9
15	M	P	89.6	129.6
16	C	M	119.8	98.9
17	C	S	121.4	61.9
18	C	V	124.0	96.2
19	C	T	140.8	125.5
20	C	P	124.8	75.7
21	GR	M	98.9	66.4
22	GR	S	89.0	49.9
23	GR	V	69.1	96.7
24	GR	T	89.3	61.9
25	GR	P	104.1	80.3
26	D	M	86.9	67.7
27	D	S	77.1	66.7
28	D	V	78.9	67.4
29	D	T	101.8	91.8
30	D	P	96.0	94.1

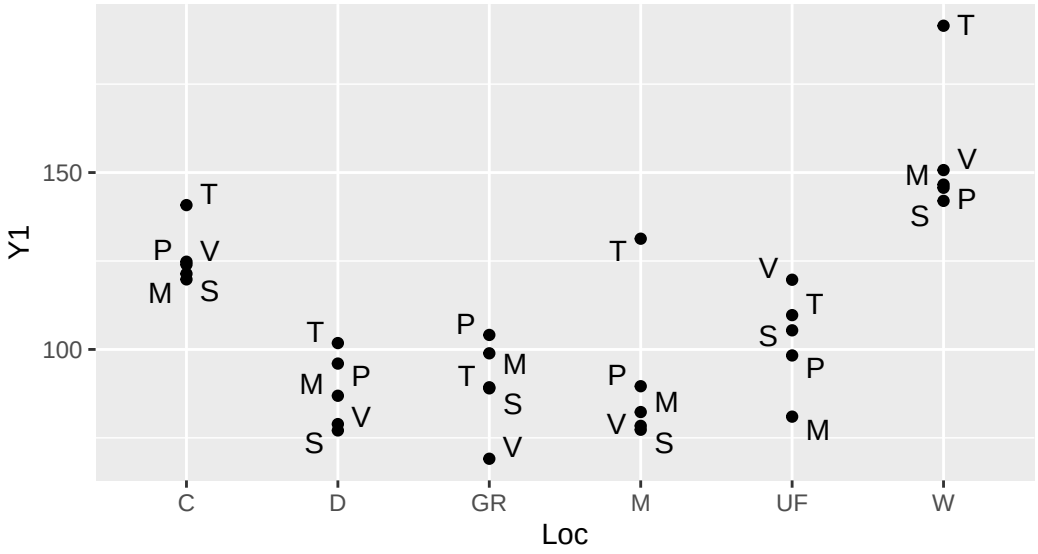
Figure 11: Barley yield data (all)

```
immer %>% select(starts_with("Y")) %>% as.matrix() -> y
immer.1 <- manova(y ~ Var + Loc, data = immer)
summary(immer.1)
```

```
          Df  Pillai approx F num Df den Df    Pr(>F)
Var        4 0.64205    2.364      8     40   0.03469 *
Loc        5 1.50658   12.213     10     40 2.543e-09 ***
Residuals 20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 12: Barley yield MANOVA

```
ggplot(immer, aes(x = Loc, y = Y1, label = Var)) + geom_point() +
  geom_text_repel()
```




```
ggplot(immer, aes(x = Loc, y = Y2, label = Var)) + geom_point() +
  geom_text_repel()
```

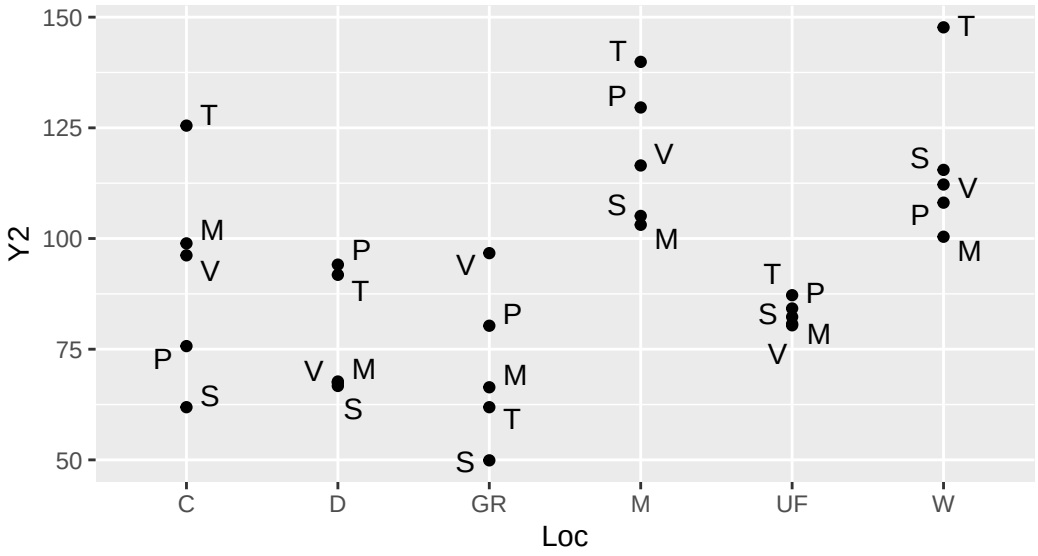


Figure 13: Barley yield plots

	id	grip.type	replicate	UBP
1	c_1	classic	1	168.2084
2	c_1	classic	2	161.4141
3	c_1	classic	3	163.2345
4	c_2	classic	1	155.9429
5	c_2	classic	2	168.5388
6	c_2	classic	3	166.3163
7	c_3	classic	1	162.6191
8	c_3	classic	2	157.8030
9	c_3	classic	3	171.6529
10	c_4	classic	1	165.1400
11	c_4	classic	2	164.9525
12	c_4	classic	3	158.2008
13	m_1	modern	1	160.0739
14	m_1	modern	2	161.2383
15	m_1	modern	3	166.7635
16	m_2	modern	1	161.8334
17	m_2	modern	2	162.7900
18	m_2	modern	3	157.5793
19	m_3	modern	1	165.2248
20	m_3	modern	2	162.7804
21	m_3	modern	3	159.7632
22	m_4	modern	1	160.3049
23	m_4	modern	2	168.5381
24	m_4	modern	3	164.4688
25	i_1	integrated	1	166.7134
26	i_1	integrated	2	173.0319
27	i_1	integrated	3	173.2537
28	i_2	integrated	1	165.4825
29	i_2	integrated	2	166.0498
30	i_2	integrated	3	170.5794
31	i_3	integrated	1	174.8182
32	i_3	integrated	2	166.8222
33	i_3	integrated	3	165.2776
34	i_4	integrated	1	174.8661
35	i_4	integrated	2	173.0058
36	i_4	integrated	3	165.1532

Figure 14: Ski grip data (all)

```
grip %>%
  group_by(grip.type, replicate) %>%
  summarize(mean_ubp = mean(UBP)) %>%
  ggplot(aes(x = replicate, y = mean_ubp,
             colour = grip.type, group = grip.type)) +
    geom_point() + geom_line()
```

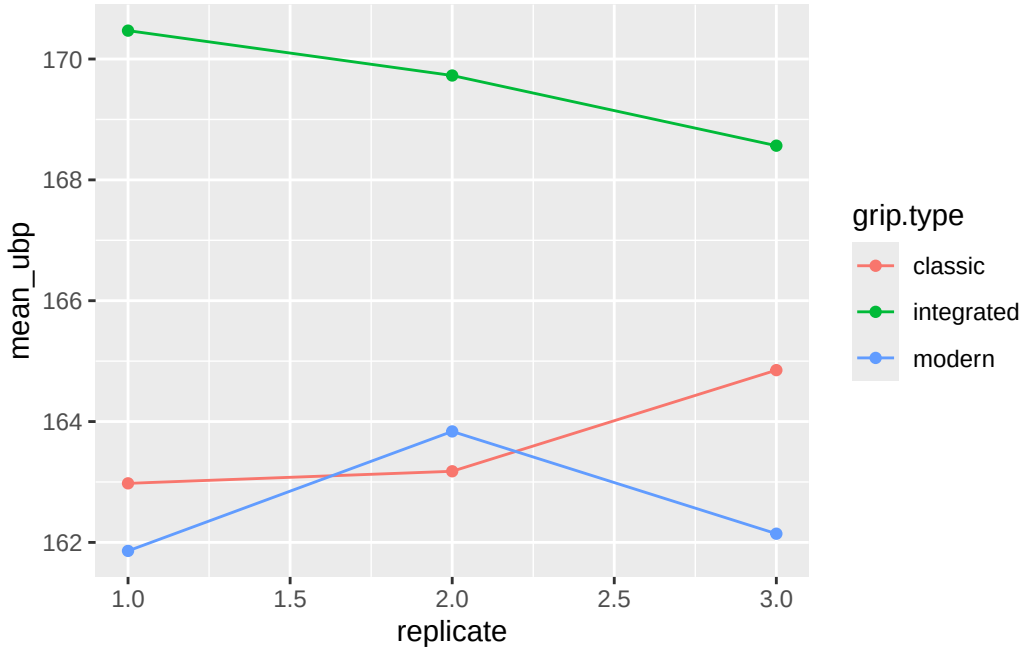


Figure 15: Ski grip interaction plot

```
ggplot(grip, aes(x = replicate, y = UBP, colour = grip.type, group = id)) +
  geom_point() + geom_line()
```

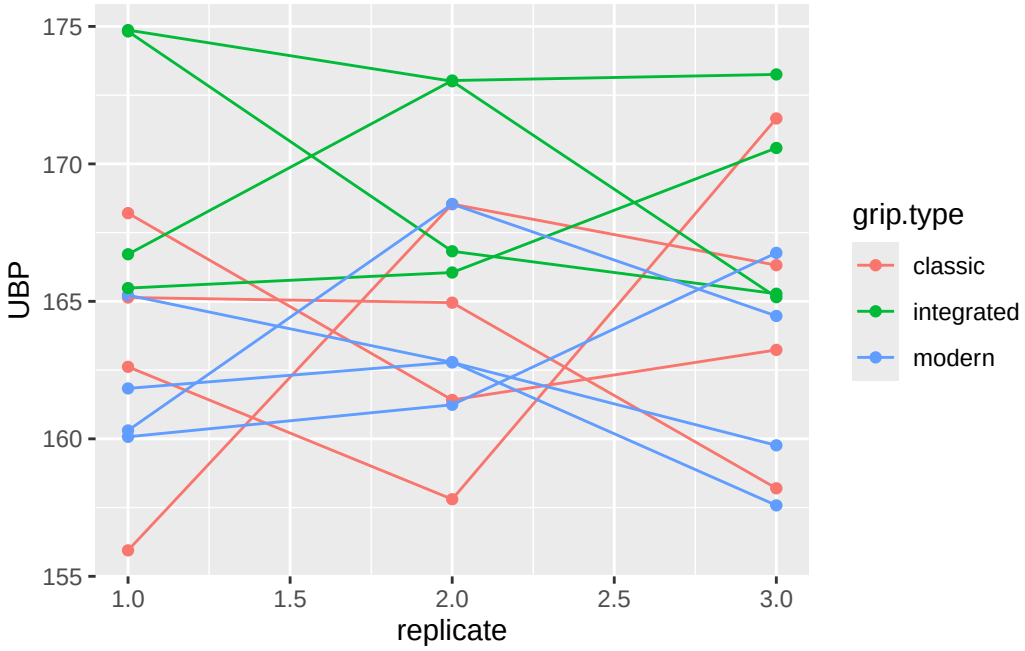


Figure 16: Ski grip spaghetti plot

```
grip %>% pivot_wider(names_from = replicate, values_from = UBP) -> grip_wide
```

Figure 17: Ski grip code

```
grip_wide %>%
  select(`1`:`3`) %>%
  as.matrix() -> y
grip.1a <- lm(y ~ grip.type, data = grip_wide)
times <- colnames(y)
times.df <- data.frame(times = factor(times))
grip.1 <- Manova(grip.1a, idata = times.df, idesign = ~ times)
```

Figure 18: Ski grip analysis code

```
Univariate Type II Repeated-Measures ANOVA Assuming Sphericity

                Sum Sq num Df Error SS den Df    F value    Pr(>F)
(Intercept)     983547      1    52.50      9 1.6861e+05 < 2.2e-16 ***
grip.type          339      2    52.50      9 2.9073e+01 0.0001182 ***
times                2      2   458.88     18 3.0600e-02 0.9698719
grip.type:times     23      4   458.88     18 2.2970e-01 0.9181327
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

                Test statistic p-value
times                   0.9088 0.68214
grip.type:times         0.9088 0.68214


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

                 GG eps Pr(>F[GG])
times           0.91642     0.9616
grip.type:times 0.91642     0.9057

                  HF eps Pr(>F[HF])
times           1.139118  0.9698719
grip.type:times 1.139118  0.9181327
```

Figure 19: Ski grip analysis output

Player	position	FGPct	FG3Pct	FTPct	Rebounds	Steals	Fouls
Dante Cunningham	forward	0.457	0.100	0.617	3.878788	0.6969697	1.530303
DeAndre Jordan	centre	0.710	0.250	0.397	14.951220	0.9878049	2.987805
Jabari Brown	guard	0.412	0.371	0.753	1.894737	0.6315789	1.736842
J.R. Smith	guard	0.417	0.383	0.750	3.114286	1.1714286	2.328571
J.J. Redick	guard	0.477	0.437	0.901	2.141026	0.5000000	1.717949
Chris Bosh	centre	0.460	0.375	0.772	7.045454	0.9318182	1.568182
Tim Hardaway	guard	0.389	0.342	0.801	2.228571	0.2857143	1.700000
Jamal Crawford	guard	0.396	0.327	0.901	1.937500	0.9218750	1.687500
Anthony Davis	forward	0.535	0.083	0.805	10.235294	1.4705882	2.073529
Marc Gasol	centre	0.494	0.176	0.795	7.777778	0.8641975	2.567901
John Wall	guard	0.445	0.300	0.785	4.632911	1.7468354	2.278481
Avery Bradley	guard	0.429	0.352	0.790	3.129870	1.0649351	2.298701
Gerald Henderson	guard	0.437	0.331	0.848	3.412500	0.6375000	1.687500
Eric Gordon	guard	0.411	0.448	0.805	2.606557	0.8196721	2.377049
Paul Pierce	forward	0.447	0.389	0.781	4.027397	0.6301370	2.191781

Figure 20: NBA 2015 data (15 randomly selected rows)

```
y <- with(nba, cbind(FGPct, FG3Pct, FTPct, Rebounds, Steals, Fouls))
nba.2 <- manova(y ~ position, data = nba)
summary(nba.2)
```

```
           Df  Pillai approx F num Df den Df    Pr(>F)
position    2 0.72575   15.852     12    334 < 2.2e-16 ***
Residuals 171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 21: NBA 2015 MANOVA

```
nba.1 <- lda(position ~ FGPct + FG3Pct + FTPct + Rebounds + Steals +
               Fouls, data = nba)
nba.1
```

```
Call:
lda(position ~ FGPct + FG3Pct + FTPct + Rebounds + Steals + Fouls,
    data = nba)

Prior probabilities of groups:
   centre   forward     guard
0.1379310 0.3563218 0.5057471

Group means:
            FGPct    FG3Pct     FTPct Rebounds    Steals    Fouls
centre  0.5214167 0.2160417 0.6998750 8.642295 0.6699533 2.614796
forward 0.4553065 0.3293065 0.7515161 5.941030 0.9615395 2.228547
guard   0.4277500 0.3467841 0.7935682 3.518978 1.1223494 2.085363

Coefficients of linear discriminants:
                LD1        LD2
FGPct    -6.7149807 17.4822905
FG3Pct    1.1377356 -5.7597904
FTPct     0.3022703  3.2938611
Rebounds -0.4652831 -0.4747002
Steals    1.4300714 -0.1587456
Fouls    -0.1288587  0.9834488

Proportion of trace:
   LD1    LD2
0.9537 0.0463
```

Figure 22: NBA 2015 discriminant analysis

```
p <- predict(nba.1)
d <- cbind(nba, p)
with(d, table(obs = position, pred = class))
```

```
          pred
obs        centre forward guard
  centre       18       6     0
  forward       9      31    22
  guard         1       4    83
```

Figure 23: NBA 2015 further discriminant analysis

```
ggplot(d, aes(x = x.LD1, y = x.LD2, colour = position)) + geom_point()
```

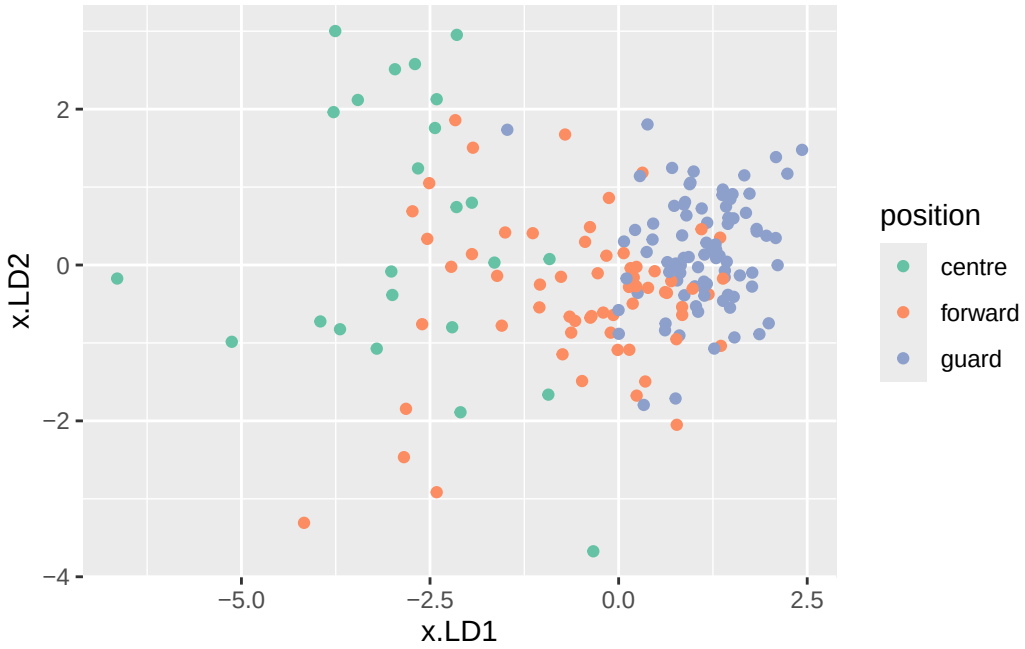


Figure 24: NBA 2015 discriminant analysis plot

	Player	position	class	p.centre	p.forward	p.guard
25	Jose Calderon	guard	guard	0.000	0.171	0.829
173	Thaddeus Young	forward	guard	0.001	0.329	0.670
103	Kawhi Leonard	forward	guard	0.000	0.349	0.651
57	Danilo Gallinari	forward	guard	0.000	0.275	0.724
105	Jeremy Lin	guard	guard	0.000	0.061	0.939
13	Nicolas Batum	forward	forward	0.001	0.647	0.353
167	Andrew Wiggins	guard	guard	0.001	0.354	0.645
140	Iman Shumpert	guard	guard	0.000	0.133	0.867
168	Deron Williams	guard	guard	0.000	0.169	0.831
141	Marcus Smart	guard	guard	0.000	0.064	0.936

Note: The columns with names starting with `p.` originally started with `posterior.` The column `p.centre`, for example, was originally called `posterior.centre`. I changed this to fit the table on the page.

Figure 25: NBA 2015 posterior probabilities (selected)

	Sl_No	L500	L1000	L2000	L4000	R500	R1000	R2000	R4000
1	47	5	0	10	70	-5	5	15	40
2	14	5	15	5	60	5	5	0	50
3	55	15	20	10	60	20	20	0	25
4	66	-10	0	5	60	-10	-5	0	65
5	71	0	10	40	60	-5	0	25	50
6	75	0	-10	0	60	15	0	5	50
7	28	-5	-5	-5	55	-5	5	10	70
8	50	-5	0	10	55	-10	0	5	50
9	67	5	10	40	55	0	5	30	40
10	98	10	10	15	55	0	0	5	75
11	18	5	0	0	50	10	10	5	65
12	27	0	0	5	50	5	0	5	40
13	73	0	5	45	50	0	10	15	50
14	34	-10	-10	-10	45	-10	-10	5	45
15	35	-5	10	20	45	-5	10	35	60

Figure 26: Hearing data (15 selected rows)

```
hearing %>% select(-Sl_No) -> hearing0
hearing.1 <- princomp(hearing0, cor = TRUE)
hearing.1
```

```
Call:
princomp(x = hearing0, cor = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
1.9821719 1.2721328 0.9875853 0.6832146 0.5831723 0.5620420 0.4473378 0.3930313

 8  variables and  100 observations.
```

```
hearing.1$loadings
```

```
Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
L500   0.401  0.317  0.158  0.328         0.446  0.329  0.546
L1000  0.421  0.225         0.482 -0.379                -0.623
L2000  0.366 -0.239 -0.470  0.282  0.439        -0.526  0.186
L4000  0.281 -0.474  0.430  0.161  0.350 -0.417  0.427
R500   0.343  0.386  0.259 -0.488  0.498  0.195 -0.159 -0.343
R1000  0.411  0.232        -0.372 -0.351 -0.614         0.361
R2000  0.312 -0.317 -0.563 -0.391 -0.111  0.265  0.478 -0.147
R4000  0.254 -0.514  0.426 -0.159 -0.396  0.366 -0.414

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

Figure 27: Hearing data principal components

```
ggscreeplot(hearing.1)
```

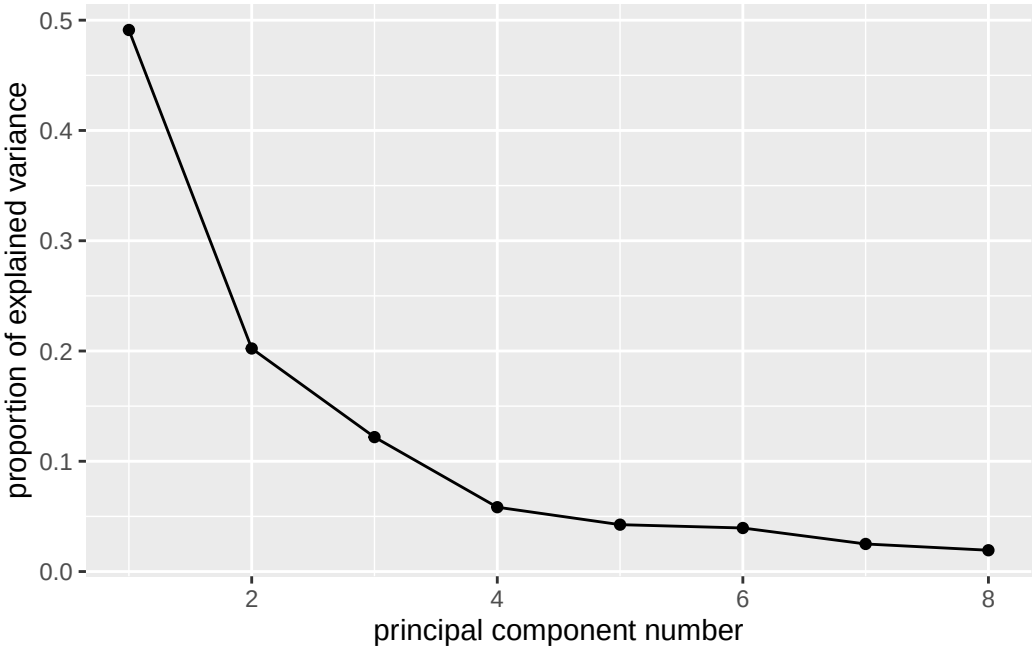


Figure 28: Hearing data screeplot

```
cbind(subject = hearing$Sl_No, hearing.1$scores) %>%
  as_tibble() -> hearing_scores
ggplot(hearing_scores, aes(x = Comp.1, y = Comp.2, label = subject)) +
  geom_text()
```

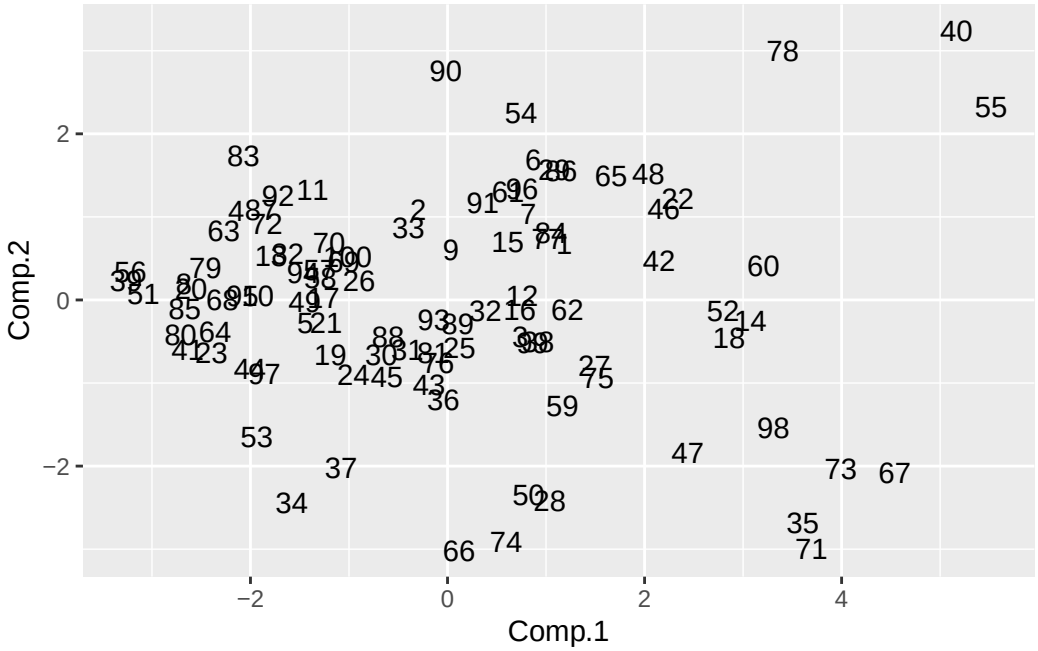


Figure 29: Hearing data principal component scores graph

```
# A tibble: 5 x 9
  percentile  L500 L1000  L2000  L4000  R500 R1000 R2000 R4000
  <chr>      <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 0%           -10   -10 -10    -10      -10   -10   -10   -10
2 25%          -10    -5  -5      8.75   -10    -5    -5     5
3 50%           -5     0   0     20       -5     0     0    15
4 75%            5     5   6.25  35        0     5     5    30
5 100%          15    20  45     70       25    20    35    75
```

Figure 30: Hearing data summary, showing the 0th, 25th, 50th, 75th, and 100th percentiles (min, Q1, median, Q3, max) of each variable

```
# A tibble: 20 x 4
   age     left_side right_side     n
   <fct>   <chr>     <chr>      <int>
 1 (15,30] no        no          1913
 2 (15,30] no        yes           41
 3 (15,30] yes       no           149
 4 (15,30] yes       yes           63
 5 (30,40] no        no          2226
 6 (30,40] no        yes           48
 7 (30,40] yes       no           190
 8 (30,40] yes       yes           84
 9 (40,50] no        no          2262
10 (40,50] no        yes           40
11 (40,50] yes       no           148
12 (40,50] yes       yes           70
13 (50,70] no        no          1974
14 (50,70] no        yes           31
15 (50,70] yes       no           113
16 (50,70] yes       yes           60
17 (70,90] no        no           671
18 (70,90] no        yes           10
19 (70,90] yes       no            55
20 (70,90] yes       yes           38
```

Figure 31: Chest pain data

```r
chest.1 <- glm(n ~ age * left_side * right_side, data = chest, family = "poisson")
drop1(chest.1, test = "Chisq")
```

```
Single term deletions

Model:
n ~ age * left_side * right_side
                         Df Deviance    AIC    LRT Pr(>Chi)
<none>                        0.0000 175.45
age:left_side:right_side  4   6.6877 174.13 6.6877   0.1533
```

```r
chest.2 <- update(chest.1, . ~ . - age:left_side:right_side)
drop1(chest.2, test = "Chisq")
```

```
Single term deletions

Model:
n ~ age + left_side + right_side + age:left_side + age:right_side +
    left_side:right_side
                     Df Deviance     AIC    LRT Pr(>Chi)
<none>                     6.69  174.13
age:left_side         4    19.52  178.97  12.83  0.01211 *
age:right_side        4     7.71  167.15   1.02  0.90674
left_side:right_side  1   983.32 1148.76 976.63  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

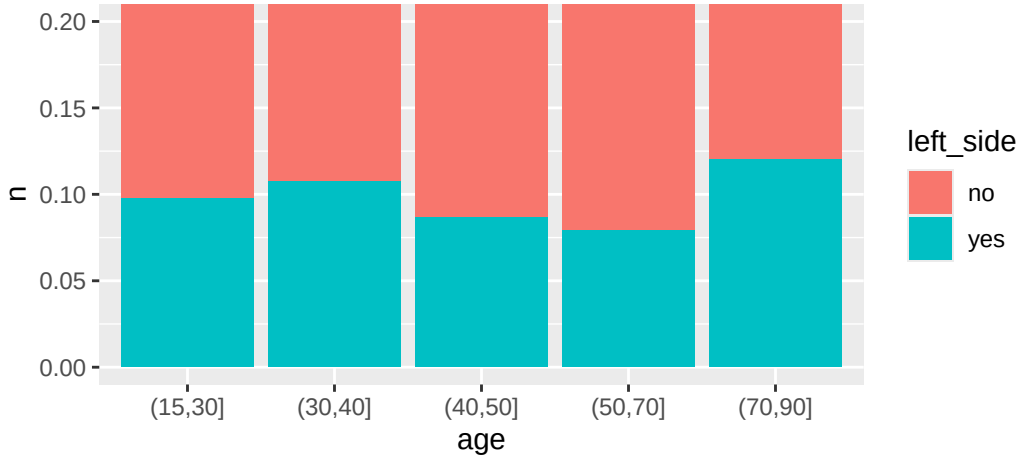
```r
chest.3 <- update(chest.2, . ~ . - age:right_side)
drop1(chest.3, test = "Chisq")
```

```
Single term deletions

Model:
n ~ age + left_side + right_side + age:left_side + left_side:right_side
                     Df Deviance     AIC    LRT  Pr(>Chi)
<none>                     7.71  167.15
age:left_side         4    26.33  177.78  18.63 0.0009308 ***
left_side:right_side  1   990.13 1147.58 982.42 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 32: Chest pain model fitting

```
ggplot(chest, aes(x = age, y = n, fill = left_side)) +
  geom_col(position = "fill") +
  coord_cartesian(ylim = c(0, 0.2))
```



Note: the `coord_cartesian` is used to truncate the y-scale, as on Worksheet 12.

Figure 33: Chest pain graph 1

```
ggplot(chest, aes(x = left_side, y = n, fill = right_side)) +
  geom_col(position = "fill")
```

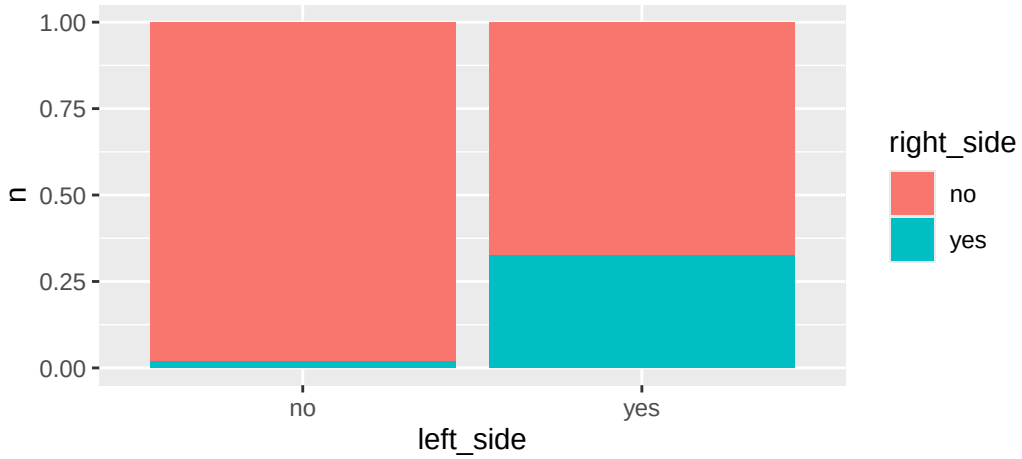


Figure 34: Chest pain graph 2