University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Midterm Exam
March 7, 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 xx 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 part of 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.

Houses in Canton, New York

Canton is a small town in upstate New York, USA (meaning, not near to New York City). 53 houses were sold there in one year. Some of the data are shown in Figure 2. The variables recorded were the selling `Price` in thousands of dollars, the number of bedrooms (`Beds`), the number of bathrooms (`Baths`), the floor area of the house in square feet (`Size`), and the `Lot` size in acres. We will be interested in predicting selling price from the floor area of the house and the number of bedrooms it has.

(1) (2 points) Scatterplots of `Price` against `Beds` and `Size` are shown in Figure 3. From this Figure, do you think that `Price` depends on either or both of `Beds` and `Size`? Explain briefly.

I think `Price` depends on both of them, because both scatterplots have a moderate upward trend. Other well-supported answers might be possible, but you'll have some work to do to make a convincing argument that there is *no* upward trend on either of those.

This is meant to be a nice gentle warm-up.

(2) (2 points) The realtors analyzing these data decided to predict selling price itself, and not, for example, the log of selling price, from the two explanatory variables. On the basis of Figure 4, why do you think they chose to do that?

The implication in the question is that we are considering a transformation of selling price. The code and output is of Box-Cox, which tells us whether we need to transform selling price. $\lambda = 1$, no transformation, is inside the confidence interval for λ, supporting the decision to predict selling price itself rather than some transformation of selling price.

"1 is inside the confidence interval", by itself, is only one point. For both points, you need to build the link from here to "don't do a transformation of the response, selling price".

Extra: the interval is rather wide, so that square root, $\lambda = 0.5$, and even log, $\lambda = 0$, are also supported by the data. If I had asked you what *you* would have done, I would have guessed that you would prefer a square root transformation, but that's not how I asked the question. You might frame this question as asking "is there any evidence that a transformation is needed", and because $\lambda = 1$ is inside the interval, the answer to that question is "no" (λ is not significantly different from 1). Of course, with a larger data set, that answer probably would change, but with the data we have, that's the answer. (It also makes the other questions here easier to think about, since we are thinking about the regression rather than grappling with a transformation at the same time.)

(3) (2 points) Two regressions are shown in Figure 5 and Figure 6. Why is it that `Beds` is significant in the second one but not the first one?

`Beds` is correlated with `Price` (as shown in the second regression), but when `Size` is in the regression as well, `Beds` does not have anything to add (over and above the effect of `Size`). One point for getting this far (this is really a C32 answer).

The likely reason for this is that the two explanatory variables are correlated with each other. From a practical point of view, you would expect a larger house to have more bedrooms and to sell for more, so that all three variables are probably (positively) correlated. The second point for saying that the two explanatory variables are correlated. (The scatterplot in Figure 8 confirms this. You may not notice that as you are working on this question, but you would do well to keep this question in mind as you are working on the question about confidence interval lengths, so that when you are working on that question, you realize that the information you are using there will also help you with this question.) The second point for asserting this correlation and for giving a good reason for why that would be the case (either a practical one, or referring to the scatterplot in Figure 8).

Another way you might have approached this is via R-squared. The regression in Figure 5 has a higher R-squared than the one in Figure 6, so it is better to use `Size` than `Beds` in predicting selling price. This is true, but doesn't answer the question, and doesn't really go beyond my "C32 answer" above, so no more than 1 point total for this approach.

Extra: you might have noted that Figure 5 seems inconsistent with Figure 3 (that you looked at in the first question). From the scatterplots, you would expect both explanatory variables to be at least weakly significant, but the regression says that `Beds` is actually not. Unexpected non-significance like this is often because of correlation between the explanatory variables (as in the punting example in lecture), which is another clue that the correlation is the reason for the difference between the two regressions.

(4) (2 points) Some predictions are shown in Figure 7, along with confidence limits. What precisely are `conf.low` and `conf.high` limits for? Explain briefly.

This uses `predictions` from `marginaleffects`, so they are confidence limits for the mean selling price of *all* houses (in Canton) (one point) with these values of `Beds` and `Size` (the second point).

If I had used `predict(... , interval = "p")`, which is what we would need to obtain prediction intervals, they would have been prediction intervals for the selling price of individual houses with those values of `Beds` and `Size` (and the intervals themselves would have been longer). But I didn't, so they are not.

(5) (3 points) The second interval in Figure 7 is longer than the first one. Explain briefly why that makes sense. You may find the information in Figure 8 useful.

To look at Figure 8: the first part says that over the whole dataset, the mean house size is 1.68 and the mean number of bedrooms is 3.4. The two predictions in Figure 7 are both for a size of 2.75 and bedrooms 5 and 2. The numbers of bedrooms are about equidistant from the mean (or, the second one is a little closer, but its interval is longer), so this does not explain the difference in interval lengths.

So, look at the scatterplot at the bottom of the Figure. This is of `Beds` vs `Size`, our two explanatory variables, so we can use it to decide whether our predictions are for values typical of the data or not:

- `Size` 2.75 and `Beds` 5: on the scatterplot, this is right next to the two observations above `Beds` of 5, and on the trend of the other points. So there is nearby data to base the prediction on (including also, say, the 4-bedroom houses of similar `Size`), and we should expect the prediction to be reasonably accurate and the interval to be short.
- `Size` 2.75 and `Beds` 2: this is in the top left of the scatterplot. There are no nearby points at all (or, you could say, this combination of values is very *un*typical of our data), and therefore the prediction for these values is likely to be inaccurate and therefore the interval will be longer.

This actually agrees with my intuition about houses: 2 is a small number of bedrooms, but 2750 square feet is a large house. You wouldn't expect to see a large house with such a small number of bedrooms (what is all that square footage being used for?) If you also have this intuition, this might guide you towards an answer, but the best response is going to be based on how this combination of bedrooms and size is unusual *for the data we have*.

Points: 1 for saying that the interval will be shorter if the values being predicted for are "near" the data, or something sufficiently close. In addition, 1 for making the assertion that Size 2.75 and Beds 2 is an unlikely combination, 2 for saying that it is unusual in *this* dataset by reference to the scatterplot. An answer that gets one of these two points is likely to be accompanied by a comment like "how do you know?"

The point of this question is to see whether you can get at the idea of "nearby data equals more accurate prediction". Sometimes, looking at means will help to uncover this, but not always: when you have more than one explanatory variable *and* they are correlated, the actual combination of values is what matters. I gave you the scatterplot so that you could see that `Size` and `Beds` were indeed (positively) correlated.

Turtles

The temperature at which turtle eggs are kept can, it is hypothesized, affect the chance that a turtle that hatches from those eggs turns out to be male or female. To assess this hypothesis, an experiment was run, in which the temperature was controlled at various different values. The data from the experiment are shown in Figure 9. The columns are, in the order shown, the temperature in degrees Celsius, the number of male turtles that hatched from the eggs at that temperature, and the number of female turtles that hatched.

(The experiment was in fact replicated at the same temperatures on three different days. This does not affect our analysis in any way.)

(6) (2 points) Why is logistic regression a sensible technique to use to assess the hypothesis of interest?

The response variable is the sex of the turtle. This is categorical (evidently, with categories male and female). Logistic regression requires a categorical response with two categories, which is exactly what we have here.

One point for categorical response; the second for saying that it has two categories and how you know (eg. by naming the two categories). "The response is categorical with two categories" as an answer is 1.5 because you have not said how you know. A "why" question requires a properly-articulated reason.

Extra: in this case, looking at the *data* will not tell you immediately what the response variable is, because there is more than one individual per row and what is shown there is counts of the two response categories, but you have to be able to see that these are "male" and "female" as categories of the response. It is easier to read the description of the data and work it out from there: temperature is explanatory and is being used to predict male-or-female, that is, sex.

(7) (3 points) Is there one or more than one individual per row of the data in Figure 9? How can you tell? How does this show up in the analysis of Figure 10?

The data in Figure 9 has a count of the *number* of males and females hatched at that temperature, so each row contains *more than one* individual. For example, the first row contains $1 + 9 = 10$ turtles. If there had been only one individual per row, we would have seen a column with a name like `sex` and values `male` or `female`. Two points for a clear enough explanation.

The third point for how it shows up in the analysis: with multiple observations per row, we need to use a two-column response with the two columns of frequencies. This is what the `cbind` in the code at the top of Figure 10 is doing.

(8) (2 points) In Figure 10, how can you tell that the model is predicting the probability that a hatched turtle is *female*?

With a two-column response, the probability being predicted is the one given in the first column of the two-column response, which in this case is `female` rather than `male`. (If my columns in `cbind` had been the other way around, I would have been predicting the probability that a turtle is *male*).

It is true but entirely irrelevant here that `female` is first alphabetically. When you have a two-column response, what counts is what is in the first column, and that can be whatever you put there.

(9) (2 points) Interpret the *sign* (positive or negative) of the number in the `temp` row of the `Estimate` column in Figure 10.

The value -2.21 is negative, which means that as temperature increases, the probability of a turtle being female *decreases*.

You really need all of that to get any of the two points. You can say "odds" or "log-odds" instead of "probability", since they go up or down the same way probability does. No credit for interpreting the number here (that's the next question).

(10) (2 points) Interpret the *value*, including its sign, of the number in the `temp` row of the `Estimate` column in Figure 10.

The Estimate is -2.21. This is a "slope", and the precise interpretation is that as temperature increases by 1 (degree Celsius), the log-odds of the turtle being female *decreases* by 2.21. This is all I need. You *must* talk about log-odds (or odds; see below) here, because that is the scale that the -2.21 is on. Talking about a change in probability here is an error.

If your interpretation of odds is any better than that of log-odds, you can also say that as temperature increases by one degree, the odds of being female changes by a factor of

```
exp(-2.21)
```

```
[1] 0.1097006
```

that is to say, it becomes

```
exp(2.21)
```

```
[1] 9.115716
```

times smaller.[1] On an exam, of course, this interpretation works only if the calculator you brought is a scientific one, so I'm not insisting on you getting this far. If you do this, make sure to show your calculation, otherwise the grader has no idea where the 0.11 or 9 came from.

I suspect people are likewise going to get two marks here, 1.5 for going the odds way and neither showing the calculation nor talking about log-odds, 1 for talking about a change in probability, or nothing.

I split this into two questions to assess the depth of your knowledge. Do you know that a negative Estimate means a decrease in probability (the previous question), and do you in addition know exactly what that number means? I would expect a strong student to be able to answer this question as well as the previous one.

Extra: this is actually a large change in log-odds, relative to an apparently small change in temperature (the range of temperatures in the data is about 3 degrees). If you look back at Figure 9, the turtles go from being almost all female at the lowest temperature to almost all male at the highest, so it is not surprising that the log-odds of being female changes rapidly.

(11) (3 points) A plot is shown in Figure 11. The researchers were interested in estimating the temperature at which 50% of the turtles would be female. What do you think that temperature is? Do you think that temperature has been estimated accurately or inaccurately? Explain briefly in both cases.

This is a plot of the predicted probability of a turtle being female, as it depends on temperature.

Follow the curve of predictions (black line) to where it crosses 0.5 on the vertical axis. The temperature at which this happens is the temperature you want. The small ticks on the x-axis are at 0.5 degrees; the prediction crosses 0.5 (on the y-axis) a bit less than halfway from 27.5 degrees to 28 degrees, say at 27.7 degrees. Two points for a sensible answer and some sort of indication of how you found it. "About halfway between 27.5 and 28 degrees", for an answer of 27.75 degrees, is also good. Misreading the graph in a way that it is obvious what you have done (for example, giving an answer like 27.5 degrees) is likely to be 1 point out of 2, at the grader's discretion.

[1] You can talk about a constant additive change in log-odds, or a constant multiplicative change in odds, because that's what the model is based on, but the change in *probability* depends on what temperature you're looking at, because the relationship between probability and odds is a non-linear one.

To assess the accuracy of this estimate, look at the width of the confidence band, top to bottom. It is enough to say that at any temperature, the probability of being female is being estimated with reasonable accuracy (because the band is narrow all the way along),[2] and therefore the estimate of this critical temperature should also be reasonably accurate. (The third point.) I was only asking you to pick one of "accurate" and "inaccurate" and defend that, but it's also OK to give (good) reasons for both. Citing the small P-value for temperature is only part of the story, because it says only that the confidence interval for the estimated slope does not include zero, but it might still be wide (in which case probabilities would be estimated inaccurately).

Another way to attack this is to say that if you go away from 27.75 degrees, the probability of a turtle being female goes away from 50% very quickly (either by looking at the graph, or the data), so the critical temperature must be very close to 27.75 degrees. The majority of the observed data at 27.7 degrees is actually male, but this temperature is as close to 50-50 as it gets: almost all females for a lower temperature, and a stronger majority of males for a higher one. (There is some uncertainty here, as we might expect, but I would say not too much.)

I was also thinking of reading the confidence band *across* at 0.50; it seems to go from 27.5 to 27.9 degrees. This is strictly not correct, but you could argue that this range of temperatures contains the ones for which the data support a probability of 0.5. If you make this sort of argument, I am happy (that is to say, you have to do something beyond reading across the page: you have to say why there is value in doing so). If you do it this way, you might feel that the interval reading across the page is on the wide side, and therefore the temperature is *not* estimated very precisely.

I need explanations, not assertions: I want to know that you got the answers you got *for a good reason*, otherwise you might have been just guessing. Expect 1 out of 3 if you give plausible answers without explanations.

Extra: estimating this temperature is a sort of "backwards estimation problem": instead of being given a temperature and asked for a probability, we are being given a probability (0.5) and asked what temperature goes with it. In this context, this is known as estimating the "median lethal dose". The name comes from the usual context of treating something with a dose of a poison and finding out what dose of the poison kills 50% of the individuals exposed to it. An accurate estimation of the median lethal dose, including a confidence interval for it, is accomplished using `dose.p` from `MASS`:

```
dose.p(turtle.1, p = 0.5)
```

[2] There are actually something like 135 turtles in the data set, which is reasonably large for a logistic regression, and so we are entitled to expect the predictions to be reasonably accurate.

```
            Dose        SE
p = 0.5: 27.7329 0.1053354
```

This gives a (more accurate[3]) estimate, and also a standard error for it. Assuming that things are reasonably close to normal (temperature is not bounded, so this is a reasonable assumption), you can make a confidence interval by going up and down twice the standard error:

```
27.7329 + c(-2, 2) * 0.1053
```

```
[1] 27.5223 27.9435
```

This seems like a decently accurate estimate of temperature, and reveals also that reading the confidence band "sideways" actually gave us a very sensible answer to this problem.

[3] Given this, I would also be happy with you saying that the temperature should be halfway between 27.5 and 28, ie., 27.75.

Marijuana use

The National Youth Survey collected data on marijuana use among young people aged 11 to 17. This survey was carried out every year from 1976 to 1980, and different young people were sampled each year. Each young person sampled was asked their sex (as they identified), and whether and how often they used marijuana (classified as "never", "once a month or less", "more than once a month"). The responses on marijuana use were abbreviated as `never`, `<1m`, and `>1m` respectively. The data for each individual were summarized into counts of how many individuals fell into each combination of sex, year, and use category (the column `n` contains the counts).

The data are shown in Figure 12, in dataframe `potuse`. There are 30 rows altogether, of which 15 randomly chosen rows are shown in the Figure. A summary is shown in Figure 13.

The survey organization was interested in whether there was a trend over time (`year` is treated as quantitative), and whether males and females used marijuana at a different level within any time trend.

(12) (2 points) A model is fit using the code in Figure 14. Why was it necessary to use `polr` (rather than, say, `glm`)?

`polr` fits "proportional-odds logistic regression", which is used when the response variable, here use category, is:

- categorical with *more than two categories* (here three)
- the categories have a *natural order* (in this case, from no marijuana use up to frequent marijuana use).

The italicized things, with a brief description of how you know in each case, are what I am after. There is no credit for saying that the response is categorical, because that does not distinguish between a regular logistic regression (that you might fit using `glm`) and an ordinal-response logistic regression (that you would use `polr` to fit).

(13) (2 points) How can you tell that the response categories are in a sensible order, based on anything you have seen about these data so far?

Look under the display of data in Figure 12: the `count(use)` displays the categories of `use` *in the order that the model will take them*: from no marijuana use through low to high.

The actual counts are of the number of rows in the dataframe with that value of `use`, which is not relevant to us; the point is that the order in which `count` displays the values of `use` is the same order in which `polr` will use them.

(14) (2 points) Some output is shown in Figure 15. What do you conclude from it?

This is a `drop1` output, showing what, if anything, can be removed from the model. In this case, both `year` and `sex` are significant, so neither of them should be removed from the model: there is both a significant time trend and a significant sex effect. (The P-value for `year`, despite the `+` in its scientific notation, is 0 to the accuracy shown.)

(15) (4 points) Some predictions are shown in Figure 16. Describe the effects of both `year` and `sex`.

A point for each of:

- as year increases, the probability of `never` decreases, and the probabilities of the other two categories increase.
- therefore, the overall level of marijuana use is *increasing* over time (for both males and females).
- Compared to females, males are less likely to have `never` used marijuana, and more likely to have used it at all (both less than and more than once per month).
- therefore, the overall level of marijuana use is *higher* for males than for females (over all years).

If you get the second point for each explanatory variable, I figure you know enough about what's going on to get the first point as well.

For each of the two explanatory variables, the first point is for a general description of the trends, and the second one is a comment about the overall level of marijuana use (low or high) across the values of the explanatory variable in question. With an ordered response, you can expect to see that a change in an explanatory variable will be associated with an *increase* or *decrease* in the response (that is, a higher or lower category becomes more likely), and so that is something you need to comment on. Imagine you were writing an article using this dataset to talk about trends in marijuana use: your reader would want to know whether it is going up or down over time and whether it is higher for males or females, so make sure you tell them that.

Extra: we have only main effects here, so we have an effect of time that is valid for both sexes, and an effect of sex that is valid for all times. When we looked at this in lecture, we hadn't talked about interactions yet, so I didn't put one in here. If I had, this would have allowed the effect of time to be different for males and females (or, equivalently, the effect of sex to be different at different times).

(16) (3 points) A plot is shown in Figure 17. In the `condition =` part of the code, what was the effect of entering those three variables in that order, and why was the order sensible?

The order of the inputs to `condition =` is:

- first: the x-axis variable
- second: colours
- third: facets

`year` is our one quantitative variable, so this makes sense to put on the x-axis. The `group` is the response category (marijuana use value), and we want to display the probabilities of these categories using colours (hence, enter it second). The third input is `sex`; this is a categorical explanatory variable, so it makes sense to use facets to display it.

Points: 0.5 for each of:

- first input: x-axis variable
- second input: colour
- third input: facets
- x-axis: quantitative x (year); "to see time trends" also works.
- colour: response category (use level)
- facets: categorical x (sex)

or an appreciable fraction of each of those, stated or implied. (There are different ways to approach this, but if you say enough about what goes in which position and why, you should be good.)

I'm not asking about an interpretation of the graph (see next paragraph), since you interpreted the predictions in the previous question, and I wouldn't ask you to do the exact same thing twice.

Extra: This is a plot of the predictions, so the conclusion you would draw from this one should be the same as you got from the numerical predictions, and therefore you get a check on your answers. The trend over time is the same, for sure. On this plot, it is not clear that there is much of a sex effect (because the differences between females, on the left, and males, on the right, are not relatively very big and, in fact, the graphs have slightly different y-scales). Because of this, I didn't ask you to interpret this plot specifically.

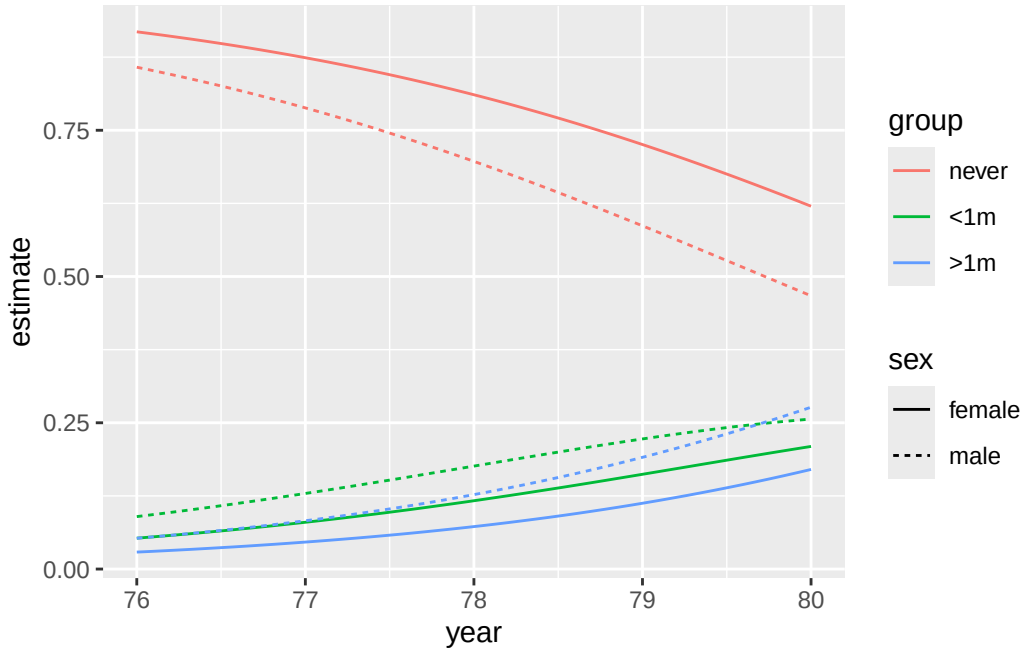
However, if you wanted[4], you could do the trick we used in lecture[5] to plot the males and females on the *same* graph:

[4] That is to say, outside of an exam, and if you had the data.

[5] The one about brand preference, which was actually a multinomial one rather than ordered, but the same graphing trick works.

```
plot_predictions(potuse.1,
                 condition = c("year", "group", "sex"),
                 type = "probs", draw = FALSE)  %>%
  ggplot(aes(x = year, y = estimate, colour = group,
             linetype = sex)) +
  geom_line()
```

```
Re-fitting to get Hessian
```



Now you get the same story as from the numerical predictions: females are more likely to be `never` and less likely to be in either of the "used some" categories, compared to the corresponding males (in the same year).

Treatments for lung cancer

14 patients with lung cancer were randomly allocated to either a new treatment (`newdrug`) or the standard treatment (`control`). The researchers were interested in whether the patients receiving the new treatment lived for longer than the patients who received the standard one.

The data, in dataframe `lungcancer`, are shown in Figure 18. The columns are:

- `time`: time from diagnosis until last observation in days
- `cens`: whether the patient was alive (0) or dead (1) when last observed
- `group`: the treatment received.

(17) (2 points) In Figure 19, some of the output values have plus signs. Why is this?

These correspond to the patients that were still alive when last observed (there were three of them, all in the treatment group).

If you use the word "censored", you need to explain what it means *in this context*. Only one point if you don't (for example, your answer is only "these data are censored"). Likewise, only one point if you say that they correspond to the observations where `cens` is zero (they do, but that doesn't tell your reader why they should be interested in these observations. More interpretation than that is needed).

(18) (2 points) What would have been another way to write the `Surv` code in Figure 19? Explain briefly why your alternative way would have worked.

This also works:

```
with(lungcancer, Surv(time, cens))
```

```
 [1]  257+  476+  355  1779   355+  191   563   242   285    16    16    16
[13]  257    16
```

That is to say, you could also have written `Surv(time, cens)`. One point.

This works because the second input to `Surv` has to be either something that is `TRUE` or `FALSE`, or something that evaluates to 1 (instead of `TRUE`) or 0 (instead of `FALSE`). In this case, `cens` is 1 if the event (death) happens and 0 otherwise, so that `cens` by itself is a perfectly good second input to `Surv`: you don't have to test whether it is equal to 1 because 1 and `TRUE` are equivalent.

The second point for saying that 1 is the event and 1 is equivalent to `TRUE`. (This is the same as in the dancing example in lecture.)

(19) (3 points) A Cox proportional-hazards model is fitted, with output shown in Figure 20. According to this Figure, does the new treatment have (i) a significant effect, (ii) a *positive* effect on survival, compared to the standard one? Explain briefly.

(i) The treatment has a significant effect, because the P-value of 0.0223 on the `groupnewdrug` line is less than 0.05, so that there is a significant difference in survival time between the patients who got the new treatment and the ones that got the standard one (the baseline `control`). One rather easy point.

(ii) To see whether the treatment has a *positive* effect (that is to say, whether the new treatment is better than the standard one), look at the coefficient on the `groupnewdrug` line. This is -2.46, negative, so that the patients who received the new treatment have a lower hazard of death at any time compared to the ones who received the (baseline) standard treatment. Having a lower hazard of death is better because it is less likely that death will happen sooner for these patients.

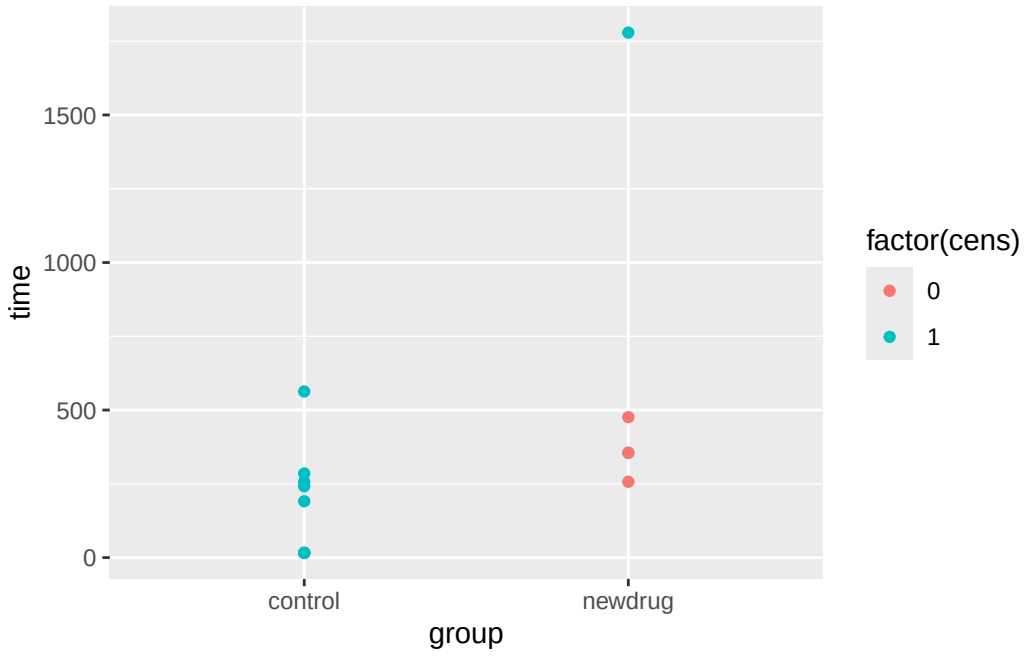
The second and third points are for (ii):

- picking out the coefficient and saying that the hazard of death is lower for the new treatment compared to the standard treatment (the second point, but see below),
- translating that into what it means about time until death (the third point).

There ought to be words like "new treatment compared to standard one", because the `group` variable in the dataset is categorical, and so it needs to be compared to the baseline category (in this case `control`). Saying that the hazard of death is lower without saying what it is lower than is only 0.5 of the second point.

Extra: With only 14 observations, you might suspect that the treatment effect needs to be rather large to be significant. You could draw a graph in an attempt to illustrate this:

```
ggplot(lungcancer, aes(x = group, y = time, colour = factor(cens))) +
  geom_point()
```



(this would normally be a boxplot, but I wanted to show the censoring somehow).

There doesn't appear to be much difference between the two groups in terms of survival time, apart from the one very long survival time in the new-treatment group, but the three patients on the new treatment whose survival times were less than 500 days were all censored, meaning that their actual survival times were greater than shown. If we had been able to observe the actual times until death for these patients, it is entirely possible that all the survival times for the new-treatment patients would have been greater than all the survival times for the patients on the standard treatment, and then the evidence for the effectiveness of the new treatment would have been very clear. (This is, admittedly, not a certainty, and the P-value from the Cox model, being small but not very small, reflects this properly.)

(20) (2 points) A plot is shown in Figure 21. Explain briefly how this plot is consistent with your answer to (ii) of the previous question (or is not consistent, if that's what you think).

In (ii) of the previous question, we concluded that the new treatment was more effective at prolonging life than the standard one. On the plot in Figure 21, the probability of survival until any time is higher with the new treatment, compared with the standard one. These are two ways to say the same thing.

I would like you to get further than "the blue survival curve is further up and to the right than the red one, so it is better" (that is only 1 point).

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

This is a martingale residual plot from the Cox model. The first point. There should be no pattern and the residuals should on average go straight across at zero, but *we should not be concerned with very negative residuals*, as I mentioned in lecture. Hence we can ignore the residual below -2, and say that the residuals are (and hence the model is) basically satisfactory.

The key thing is that martingale residuals (from a model like this) can go very negative, and therefore if they do so it is *not* evidence of fanning-out or an outlier or anything like that. Claiming an outlier or fanning-out will not get you the second point. Nor will claiming there is a problem without telling me what the problem is.

I realized it was too easy to get 2 points from this one by saying there are no problems; it might have been better to ask directly about that very negative residual at the bottom right, like for example "does this indicate a problem with the survival model? Explain briefly". But we have what we have.

Shock

A psychologist designed a experiment to test the effect of electric shock on the number of attempts it took to successfully complete a (difficult) task. They compared three treatments: no shock, medium shock, severe shock. These are labelled respectively as `Group1` through `Group3` in column `group` in the dataset. 27 subjects were randomly assigned to one of the three treatments.

The psychologist wanted to know two things: (i) if there is an effect of any shock vs. no shock, and (ii) how medium shock compared to severe shock. For the response variable, `attempts`, a smaller value is better. The data, in dataframe `Shock`, are shown in Figure 23.

(22) (2 points) Why is it better to use contrasts to analyze these data than the standard one-way ANOVA followed by Tukey?

The psychologist is not interested in comparing all possible pairs of treatments (as Tukey would do), but in making only the two specific comparisons (i) and (ii) given above.

Get this far for the two points.

Contrasts will enable the psychologist to make *only* the comparisons of interest (more powerfully than Tukey, because that makes other comparisons not of interest).

(23) (2 points) What R code will create contrasts `c_any` and `c_med_sev` that we will be able to use to test the comparisons of interest? Note that `Group1` through `Group3` are in that order.

This:

```
c_any <- c(1, -0.5, -0.5)
c_med_sev <- c(0, 1, -1)
```

The first contrast compares the average of any shock (the average of medium and severe) vs. none, and the second one compares medium and severe shock with each other. As usual, each contrast can be multiplied through by anything, so that for example the first contrast could be written as `c(-2, 1, 1)`.

One point for each of these, or for anything equivalent to them. Any errors will get you zero for that contrast, but you get a half point (overall) if I think you're close enough (such as, having two contrasts with three numbers in each, wrong but I can see what you were doing).

You have some checks on your work:

- the three numbers within a contrast have to add up to zero
- the two contrasts should be orthogonal to each other (because those are the only kind of contrasts we deal with). That's the next question.

To check the first of those:

```
sum(c_any)
```

```
[1] 0
```

```
sum(c_med_sev)
```

```
[1] 0
```

(24) (2 points) Verify that your two contrasts are orthogonal. Show your calculation (that is, *not* R code that will do your calculation).

Multiply the first number in your first contrast by the first number in the second one, the second number in your first contrast by the second number in the second one, and so on. Then add up your results and show you get zero. With my numbers (use yours):

$(1)(0) + (-0.5)(1) + (-0.5)(-1) = 0 - 0.5 + 0.5 = 0.$

So my two contrasts are orthogonal.

For your contrasts, do this calculation and say what your calculation tells you about orthogonality. If you end up concluding that they are *not* orthogonal, that's a warning sign to you (to check your work on the previous question), but for this question you can get all the points by doing the calculation based on the contrast coefficients you had, and making the appropriate conclusion about orthogonality. If you couldn't answer the previous question at all, make up two length-three contrasts and show that they are orthogonal (or not). The purpose of this question is to show that you know what orthogonality of contrasts means.

Points: one for doing a calculation like mine for your contrasts correctly, 0.5 if you make a small error (in the grader's estimation). Then one for making the appropriate conclusion from your calculation about orthogonality.

If you don't end with something like "zero, therefore orthogonal", you are not completing the verification that your two contrasts are orthogonal (you are just doing a calculation). Having said that, you can mess up the calculation completely, get zero, assert orthogonality, and still get one point because you have shown that you know why you are doing the calculation (even if you are not able to do it).

(Some of you noticed that the calculation for testing orthogonality is a "dot product": if the dot product is zero, the contrasts are orthogonal; if you think of the contrasts as vectors in 3D, they are at right angles to each other.)

(25) (2 points) What R code will set it up so that running the ANOVA as a regression will test the two contrasts of interest? You may assume that `group` is a `factor`. This question is *not* asking about how to run the ANOVA as a regression; it is asking what you do *before* that, in order to make it work.

This:

```
m <- cbind(c_any, c_med_sev)
contrasts(Shock$group) <- m
```

Put the contrasts into a matrix, and then set that up as the `contrasts` for the categorical variables `group` within the dataframe `Shock`. (This says what comparisons of `group`s you want to make.) It is fine to put the contrasts the other way around when defining `m`, because you will still get the right tests.

This is the code that was run to make Figure 24. I surreptitiously made `group` into a factor there before running this code, but I wanted you to show me you could do this part. There is no problem if you include the code to make `group` into a factor, as long as you also include the two lines above, or something equivalent to them that will work.

One point for each of the two lines. If the grader thinks you made an error but it was only a small one, you might get 0.5 on either line.

(26) (2 points) What do you conclude from Figure 24? Assume that all tests are two-sided. If you use a P-value to draw a conclusion, say which P-value you are using to draw that conclusion from.

Use the P-values on the `groupc_any` and `groupc_med_sev` lines; these are the contrasts you defined earlier, without `group` on the front:

- the `groupc_any` P-value is 2.0×10^{-8}, much less than 0.05, so there is a difference in (mean) number of attempts between subjects who received any shock at all and those that received no shock.
- the `groupc_med_sev` P-value is 9.7×10^{-5}, also much less than 0.05, so there is a difference in (mean) number of attempts between subjects who received a medium shock and those that received a severe shock.

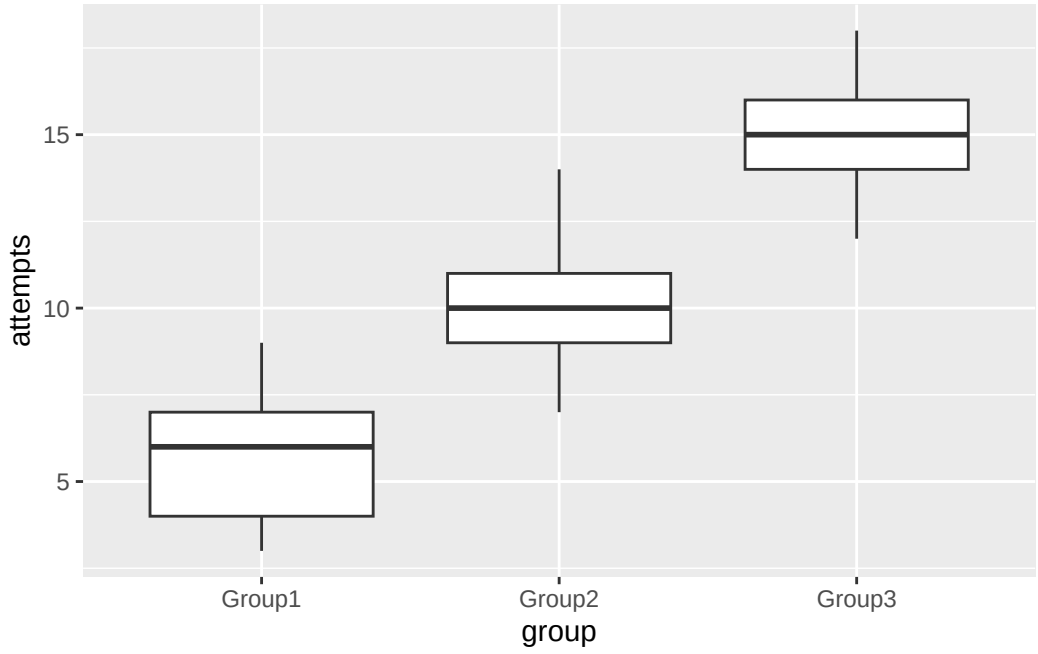
Make sure you use the appropriate P-value in each case to test the appropriate one of the comparisons (i) and (ii) described in the problem's preamble that the psychologist wanted to know about. I tried to give the contrasts memorable names so that you can tell which is which. By "two-sided" I mean that you should say something about the number of attempts being *different*, not about it being greater (or less). See the first Extra below for why I did things this way.

One point for each. Make it clear which P-value you are using for each somehow. For example, you could say "the P-value on the `groupc_any` line is less than 0.05" and infer a difference between any shock and no shock. As long as it is clear which P-value you are using to make which comparison, I am happy.

A reminder that saying "the results are significant" and stopping there will get you less than half the points, because I need to know which P-value(s) you are drawing your conclusions from (there are three in the table plus the one at the bottom), and *you need to state your conclusions in the context of the data*. To get this right, imagine that you are writing a report for someone like your boss, or the principal investigator on this study, who wants to know what *they* should do. That means saying *what* is different (the number of attempts), as well as under what conditions it is different (between any shock and no shock in the first case, and between medium shock and severe shock in the second). It also helps to be able to read R's scientific notation; for example `1.96e-08` means 1.96×10^{-8}, which is actually 0.0000000196 ($8 - 1 = 7$ zeros after the decimal point).

Extra: I didn't give you a graph, because here that makes it too easy to guess what is going on:

```
ggplot(Shock, aes(x = group, y = attempts)) + geom_boxplot()
```



As you see, there is a steady upward trend: the greater the shock, the greater the number of attempts, and the groups barely even overlap. That is why the two contrasts came out so significant. If I had given you this graph, you would (rather easily) have been able to draw a one-sided conclusion like this.

The way I arranged my contrasts, both my `Estimate`s came out negative: attempts are less for no shock vs. any shock, and for medium shock vs. severe shock. But you might have written your contrasts with opposite signs, and if you had done that and actually been able to do the analysis yourself, in that case, your Estimates would have been positive. The Estimate is consistent with the way you write your contrast, so you are not able to draw a one-sided conclusion from *my* Figure 24 without knowing how I wrote my contrasts.

Extra extra: you could do the standard analysis here, and expect to get a similar result:

```
attempts.2 <- aov(attempts ~ group, data = Shock)
summary(attempts.2)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
group        2  364.5  182.26   44.63 8.19e-09 ***
Residuals   24   98.0    4.08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(attempts.2)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = attempts ~ group, data = Shock)

$group
                   diff      lwr       upr     p adj
Group2-Group1 4.555556 2.176691  6.934420 0.0002065
Group3-Group1 9.000000 6.621136 11.378864 0.0000000
Group3-Group2 4.444444 2.065580  6.823309 0.0002771
```

There are differences in number of attempts between all three treatment groups; all three P-values in the Tukey are smaller than 0.05. But:

- you have to do some logical jumping around to get from here to what the psychologist wanted
- the P-values from Tukey, though small, are not all as small as the ones from the contrasts, because there is the extra "overhead" from comparing all pairs of treatments rather than focusing on the specific things we wanted to compare. Where the data are less clear-cut, this might be the difference between being able to demonstrate the effects we care about (contrasts) and not (Tukey, with the additional unwanted comparisons).

Growth of pigs

In the data shown in Figure 25, fifty pigs were randomly allocated to one of five feed treatments, labelled `T1` through `T5` in column `treatment`. Each pig's weight was measured before the study (in `weight1`) and again after the study (in `weight2`); the column `gain` is the difference `weight2` minus `weight1` and reflects how much weight the pig gained over the course of the study. The column `feed` shows how much of their allocated feed the pig consumed during the study. We ignore the column `rep`.

Interest is in whether a pig's weight gain depends on the feed treatment they were on and the amount of feed that they consumed. The dataframe is called `crampton.pig`.

(27) (2 points) Figure 26 shows a graph of `gain`, `feed`, and `treatment`. In predicting `gain`, do you think there is a significant interaction between `feed` and `treatment`? Explain briefly.

I would expect to see a significant interaction because the lines do not all have the same slope. In particular, the trend for treatment `T2` is almost level, while the trends for the other treatments are upward and of similar slopes (given the amount of variability present in the data).

Say that not all the lines are parallel (one point), and say *how you know* as specifically as you can (the second point).

Extra: the `crampton` in the dataframe name is one of the authors of a 1934 paper that used these data.

(28) (2 points) An analysis is shown in Figure 27. What do you conclude from this Figure?

The P-value for the interaction is 0.031, less than 0.05, so the interaction between feed and treatment is significant and should be kept in the model. (That is to say, those lines in Figure 26 really are not all parallel.)

This was meant to be an easy one.

(29) (2 points) Why is it better to use Figure 27 to answer the previous question, rather than the output in Figure 28?

The interaction contains a categorical variable `treatment`, which we want to assess the effect of as a whole, hence the use of `drop1` (which does precisely that). The `summary` output tests each slope against that of the baseline treatment `T1`, which does not assess the overall effect of `treatment` in the interaction.

Something like "Figure 28 is complicated but Figure 27 is simple" doesn't get at any of this.

(30) (2 points) On the graph in Figure 26, what was the most important piece of evidence that you used in answering Question 27? How does this evidence show up in the output from `summary` for this model shown in Figure 28?

On the graph, the most important evidence in favour of the interaction is that the slope for treatment `T2` is very different from the slopes for the other treatments. (You might have said this earlier.)

On Figure 28, the Estimates for the interaction terms say how the slopes for each treatment differ from the slopes for the baseline treatment `T1` (that is to say, the Estimate for `feed:treatmentT1` is zero). The estimated slopes for treatments `T3` through `T5` are close to zero, so the slopes for these treatments are close to the slope for `T1`. The slope for treatment `T2`, however, is very different (much more negative than the others), so its line on Figure 26 should (and does) go up much less steeply than the others. This is the best approach, because it's talking about the same thing seen two different ways, but you might be able to use the P-values instead of the slopes if you are careful (see below).

Two points, one for each of those two things. If you made a different observation from the graph, try to be consistent and say how it shows up in Figure 28. If your different observation is correct and relevant, you can still get the two points this way.

To be more precise, the Estimate for `feed`, 0.24, is the slope for the baseline treatment `T1`, and the Estimates for the interaction terms say how the slope for that treatment compares to 0.24. Thus, for example, the slope for treatment `T4` is $0.24 - 0.05 = 0.19$, close to the slope for treatment `T1`. On the other hand, the slope for treatment `T2` is $0.24 - 0.23 = 0.01$, which makes sense because the line is almost flat for that treatment. (If you work out the slopes for the other treatments, you'll find that they are close to the 0.24 for treatment 1 as well.) I would also consider a well-argued answer that says that only T2 has a significantly different slope than the baseline T1, the implication being that T1, T3, T4, and T5 all have the same slopes but T2 is different.

That is to say, all the evidence is pointing towards the interaction being significant because the slope of the relationship between `gain` and `feed` was different for treatment `T2` than for the other treatments, which might all have the same slopes.

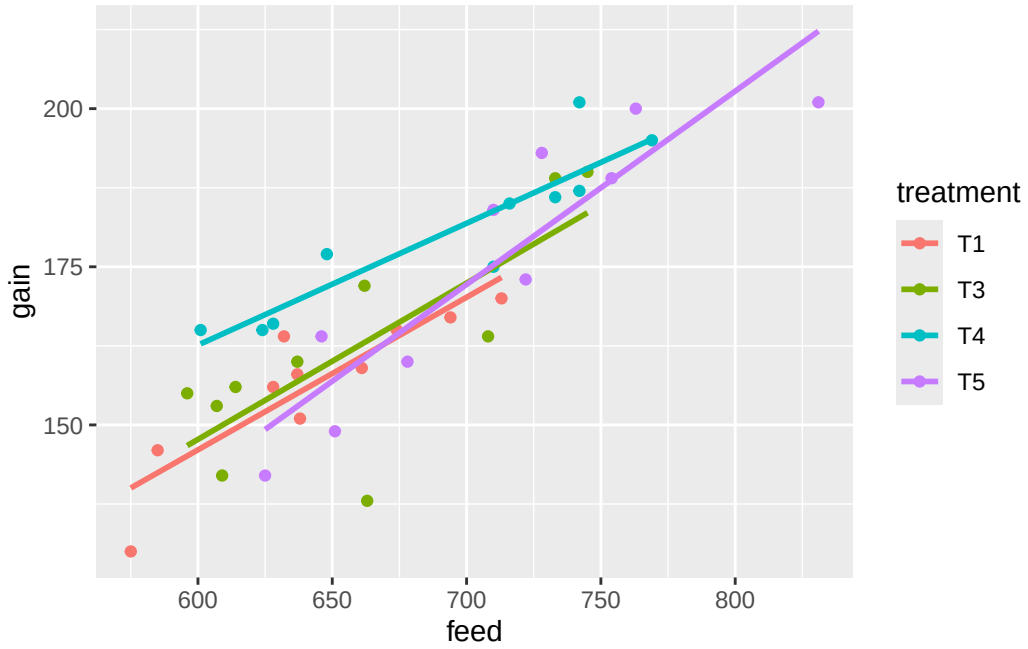
Extra: this is of course something that I can assess. Let's take out `T2` first:

```
crampton.pig %>% filter(treatment != "T2") -> t1345
```

and then repeat what I did in the Figures with these data. The `se = FALSE` below (which I also used in the Figure I gave you) gets rid of the coloured envelopes around the lines, which makes for a less cluttered plot:

```
ggplot(t1345, aes(x = feed, y = gain, colour = treatment)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Those lines look pretty close to parallel, and with the amount of variability there is, there may well no longer be an interaction:

```
t1345.1 <- lm(gain ~ feed * treatment, data = t1345)
drop1(t1345.1, test = "F") %>% knitr::kable()
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	2324.454	178.4944	NA	NA
feed:treatment	3	219.4297	2543.884	176.1027	1.006939	0.4024176

The thing on the end displays the output from `drop1` about the way you are used to seeing it.

Now you see that the interaction is nowhere near significant, and therefore those slopes are not significantly different from each other. You might think that the pale blue line is slightly less steep and the purple one is slightly more steep, but the test says that this is just chance.

To continue with these four treatments, we remove the interaction and fire up `drop1` again:

```
t1345.2 <- lm(gain ~ feed + treatment, data = t1345)
drop1(t1345.2, test = "F") %>% knitr::kable()
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	2543.884	176.1027	NA	NA
feed	1	6961.5163	9505.400	226.8294	95.779958	0.0000000
treatment	3	801.5946	3345.478	181.0594	3.676244	0.0211088

and now we see that there is a significant `feed` effect (that applies for all treatments), and a significant `treatment` effect (that applies for all values of `feed`). To see what it looks like, we look at the `summary`:

```
summary(t1345.2)
```

```
Call:
lm(formula = gain ~ feed + treatment, data = t1345)

Residuals:
    Min      1Q  Median      3Q     Max
-25.295  -4.761   1.562   6.535  13.216

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.70594   16.60032  -0.223  0.82464
feed         0.24904    0.02545   9.787 1.49e-11 ***
treatmentT3  1.88818    3.82858   0.493  0.62497
treatmentT4 11.74578    4.00045   2.936  0.00584 **
treatmentT5  2.18953    4.17755   0.524  0.60350
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.525 on 35 degrees of freedom
Multiple R-squared:  0.8075,    Adjusted R-squared:  0.7855
F-statistic: 36.71 on 4 and 35 DF,  p-value: 4.538e-12
```

The significant Estimate 0.25 for `feed` says that this is the best estimate of the slope of the line for *all* the treatments: as `feed` increases by 1, `gain` increases by about 0.25, regardless of which treatment the pig is on.

The suspicion from looking at the `treatment` terms is that `gain` is larger for treatment `T4` than it is for any of the other remaining treatments, no matter what value of `feed` you are looking at. This is consistent with the pale blue line for treatment `T4` being more or less at the top of the graph all the way across, certainly where most of the values of `feed` are concentrated.

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(survival)
```

Figure 1: Packages

Price	Beds	Baths	Size	Lot
87.0	3	1.5	1.740	0.25
160.0	4	2.0	2.060	0.60
144.0	3	1.5	1.968	0.34
78.0	2	1.0	0.712	1.17
82.7	3	1.0	1.100	2.07
92.5	3	1.0	1.329	0.42
138.5	3	2.0	1.416	0.70
195.0	4	3.0	1.848	1.84
185.0	4	3.5	2.220	0.12
127.0	3	2.0	1.184	0.34
89.0	4	1.0	2.274	1.00
150.0	4	2.0	1.704	0.27
99.0	3	1.0	0.864	1.66
174.0	4	3.0	1.382	0.48
82.0	3	3.0	1.454	2.50

Figure 2: Canton houses data (15 randomly chosen rows)

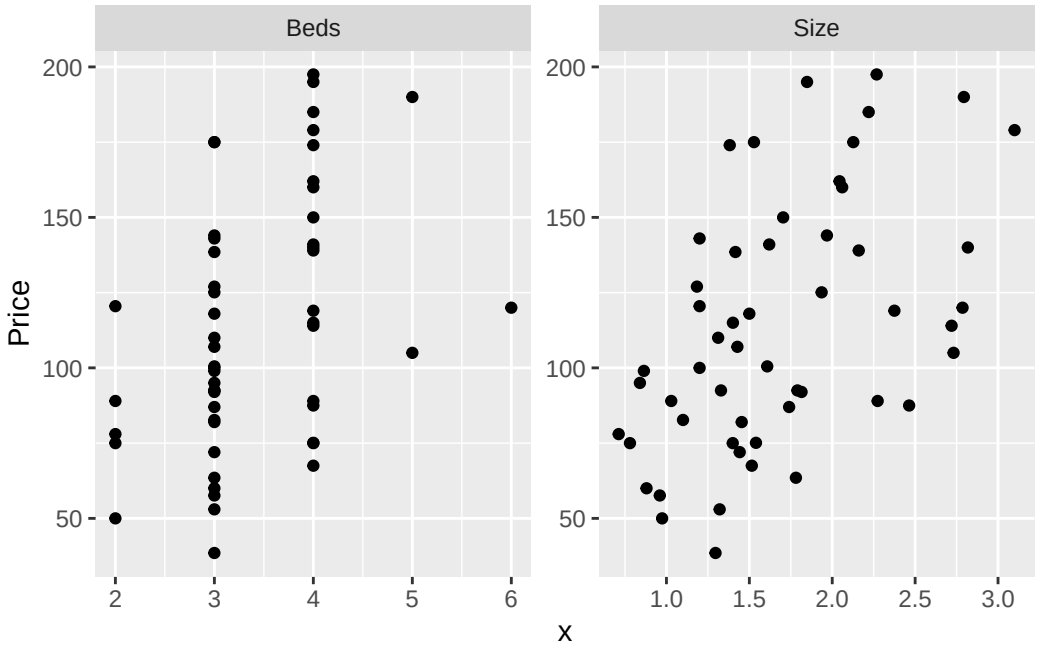


Figure 3: Canton houses scatterplots

```
boxcox(Price ~ Size + Beds, data = houses_ny)
```

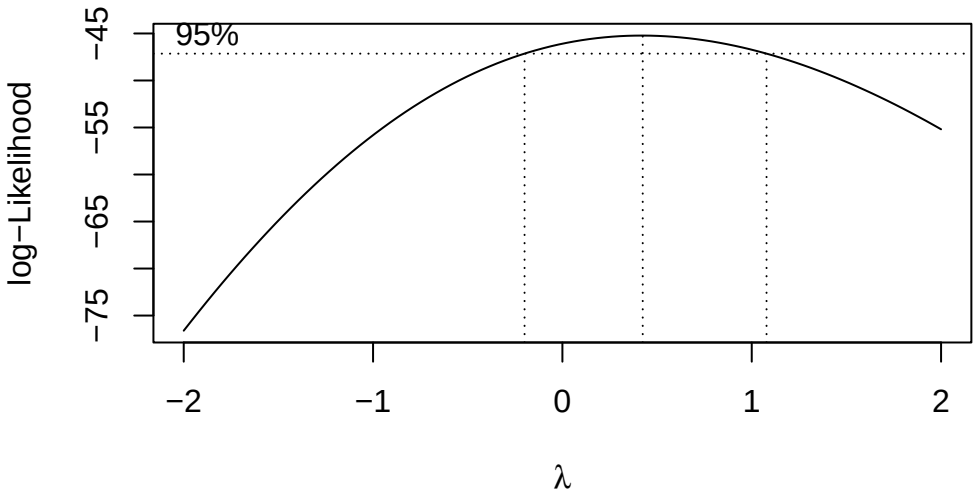


Figure 4: Code and output from houses data

```r
houses.1 <- lm(Price ~ Size + Beds, data = houses_ny)
summary(houses.1)
```

```
Call:
lm(formula = Price ~ Size + Beds, data = houses_ny)

Residuals:
    Min      1Q  Median      3Q     Max
-61.493 -31.920   1.696  27.866  73.436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.498     22.277   2.087    0.042 *
Size          31.169     12.617   2.470    0.017 *
Beds           4.367      9.515   0.459    0.648
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.21 on 50 degrees of freedom
Multiple R-squared:  0.2653,    Adjusted R-squared:  0.236
F-statistic:  9.03 on 2 and 50 DF,  p-value: 0.0004489
```

Figure 5: Regression 1 for Canton houses data

```r
houses.2 <- lm(Price ~ Beds, data = houses_ny)
summary(houses.2)
```

```
Call:
lm(formula = Price ~ Beds, data = houses_ny)

Residuals:
    Min      1Q  Median      3Q     Max
-66.453 -32.953  -5.048  33.142  70.642

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.239     23.161   1.694  0.09632 .
Beds          21.905      6.644   3.297  0.00179 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.98 on 51 degrees of freedom
Multiple R-squared:  0.1757,    Adjusted R-squared:  0.1595
F-statistic: 10.87 on 1 and 51 DF,  p-value: 0.001785
```

Figure 6: Regression 2 for Canton houses data

```
new <- tribble(
  ~Size, ~Beds,
  2.75, 5,
  2.75, 2
)
cbind(predictions(houses.1, new)) %>%
  select(Beds, Size, estimate, conf.low, conf.high) %>%
  mutate(conf.length = conf.high - conf.low)
```

```
# A tibble: 2 x 6
   Beds  Size estimate conf.low conf.high conf.length
  <dbl> <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
1     5  2.75     154.    131.       177.        45.1
2     2  2.75     141.     90.9      191.       100.
```

Figure 7: Canton houses data: predictions (values rounded to 3 significant digits)

```
houses_ny %>%
  summarize(mean_beds = mean(Beds),
            mean_size = mean(Size))
```

```
# A tibble: 1 x 2
  mean_beds mean_size
      <dbl>     <dbl>
1      3.40      1.68
```

```
ggplot(houses_ny, aes(x = Beds, y = Size)) + geom_point()
```

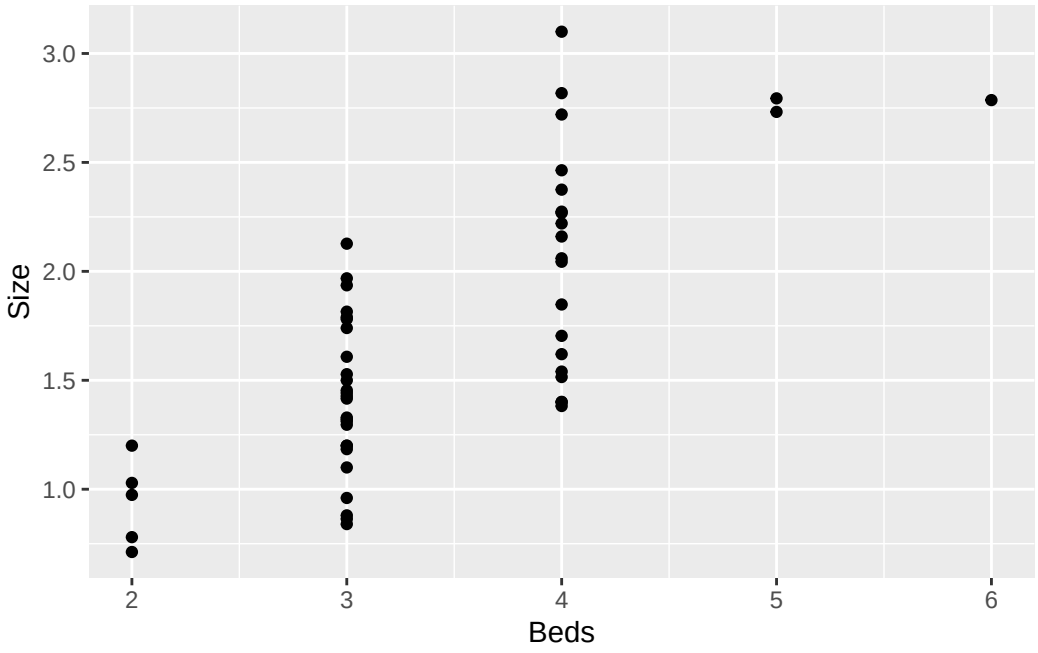


Figure 8: Canton houses data: further analysis

```
# A tibble: 15 x 3
    temp  male female
   <dbl> <int>  <int>
 1  27.2     1      9
 2  27.2     0      8
 3  27.2     1      8
 4  27.7     7      3
 5  27.7     4      2
 6  27.7     6      2
 7  28.3    13      0
 8  28.3     6      3
 9  28.3     7      1
10  28.4     7      3
11  28.4     5      3
12  28.4     7      2
13  29.9    10      1
14  29.9     8      0
15  29.9     9      0
```

Figure 9: Turtle hatch data, in dataframe `turtle` (all)

```r
turtle.1 <- glm(cbind(female, male) ~ temp, data = turtle,
                family = "binomial")
summary(turtle.1)
```

```
Call:
glm(formula = cbind(female, male) ~ temp, family = "binomial",
    data = turtle)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  61.3183    12.0224   5.100 3.39e-07 ***
temp         -2.2110     0.4309  -5.132 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.508  on 14  degrees of freedom
Residual deviance: 24.942  on 13  degrees of freedom
AIC: 53.836

Number of Fisher Scoring iterations: 5
```

Figure 10: Turtle hatch data logistic regression

```
plot_predictions(model = turtle.1, condition = "temp")
```

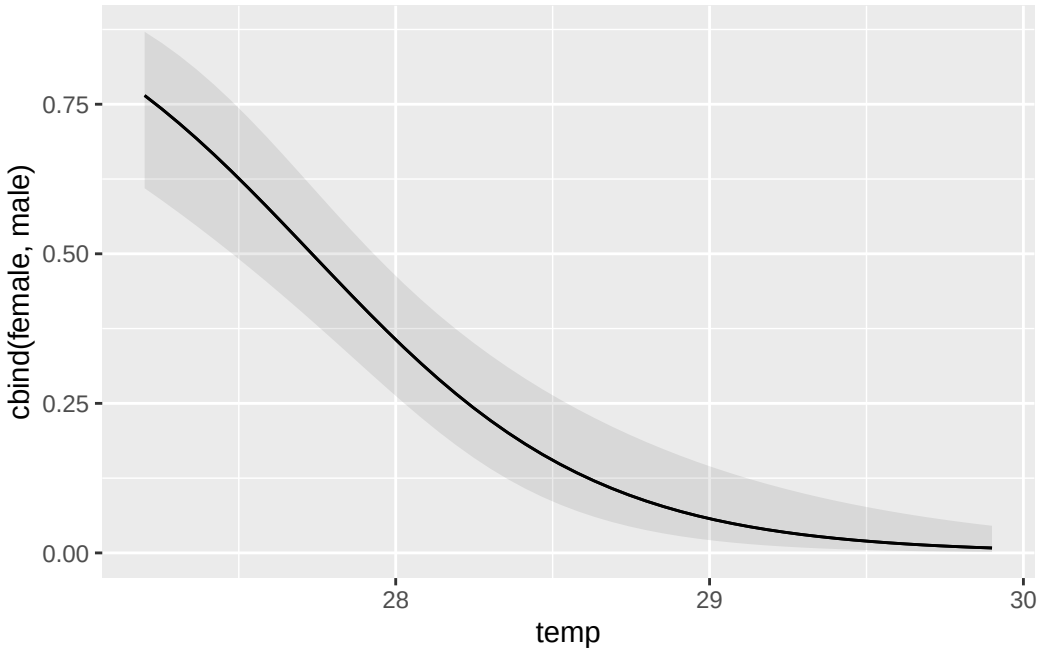


Figure 11: Turtle hatch data plot

```
# A tibble: 15 x 4
   sex     year use       n
   <chr>  <dbl> <fct> <int>
 1 male      76 <1m       9
 2 female    77 >1m       4
 3 male      77 >1m      11
 4 female    79 never    85
 5 male      79 <1m      20
 6 female    77 <1m      10
 7 male      78 >1m      21
 8 female    80 >1m      15
 9 female    76 >1m       1
10 female    78 never    91
11 female    78 >1m       8
12 male      76 never   104
13 male      79 >1m      26
14 male      78 <1m      20
15 female    77 never   106
```

Figure 12: Marijuana use data (15 randomly chosen rows)

```
potuse %>% count(use)
```

```
# A tibble: 3 x 2
  use       n
  <fct> <int>
1 never    10
2 <1m      10
3 >1m      10
```

Figure 13: Marijuana use summary

```
potuse.1 <- polr(use ~ sex + year, data = potuse, weights = n)
```

Figure 14: Marijuana use model

```
drop1(potuse.1, test = "Chisq")
```

	Df	AIC	LRT	Pr(>Chi)
	NA	1675.796	NA	NA
sex	1	1694.766	20.96921	4.7e-06
year	1	1771.661	97.86416	0.0e+00

Figure 15: Marijuana use model output

```
new <- datagrid(model = potuse.1,
                year = c(76, 78, 80),
                sex = c("male", "female"))
cbind(predictions(potuse.1, newdata = new)) %>%
  select(year, sex, group, estimate) %>%
  mutate(estimate = round(estimate, 3)) %>%
  pivot_wider(names_from = group, values_from = estimate)
```

```
Re-fitting to get Hessian

# A tibble: 6 x 5
   year sex    never `<1m` `>1m`
  <dbl> <chr>  <dbl> <dbl> <dbl>
1    76 male   0.858 0.09  0.053
2    76 female 0.918 0.053 0.029
3    78 male   0.697 0.176 0.127
4    78 female 0.811 0.117 0.073
5    80 male   0.467 0.256 0.277
6    80 female 0.62  0.21  0.17
```

Note: the predictions are shown to three decimal places. If only two decimal places are shown, the third one is zero.

Figure 16: Marijuana use: predictions

```
plot_predictions(model = potuse.1, condition = c("year", "group", "sex"))
```

```
Re-fitting to get Hessian
```

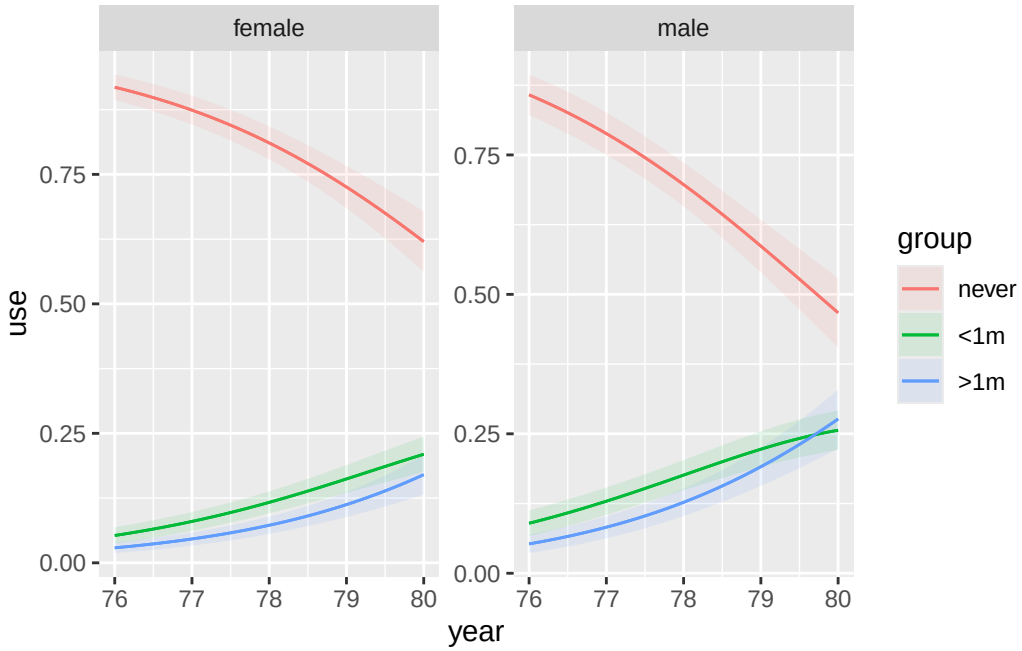


Figure 17: Marijuana use: plot

```
lungcancer
```

```
# A tibble: 14 x 3
    time  cens group
   <dbl> <dbl> <fct>
 1   257     0 newdrug
 2   476     0 newdrug
 3   355     1 newdrug
 4  1779     1 newdrug
 5   355     0 newdrug
 6   191     1 control
 7   563     1 control
 8   242     1 control
 9   285     1 control
10    16     1 control
11    16     1 control
12    16     1 control
13   257     1 control
14    16     1 control
```

Figure 18: Lung cancer data

```
with(lungcancer, Surv(time, cens == 1))
```

```
 [1]  257+  476+  355  1779   355+  191   563   242   285    16    16    16
[13]  257    16
```

Figure 19: Lung cancer code and its output

```r
lungcancer.1 <- coxph(Surv(time, cens == 1) ~ group, data = lungcancer)
summary(lungcancer.1)
```

```
Call:
coxph(formula = Surv(time, cens == 1) ~ group, data = lungcancer)

  n= 14, number of events= 11

                 coef exp(coef) se(coef)      z Pr(>|z|)
groupnewdrug -2.45904   0.08552  1.07581 -2.286   0.0223 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
groupnewdrug   0.08552      11.69   0.01038    0.7043

Concordance= 0.764  (se = 0.064 )
Likelihood ratio test= 8.84  on 1 df,   p=0.003
Wald test            = 5.22  on 1 df,   p=0.02
Score (logrank) test = 7.82  on 1 df,   p=0.005
```

Figure 20: Lung cancer Cox model

```
plot_predictions(lungcancer.1, condition = c("time", "group"),
                 type = "survival")
```

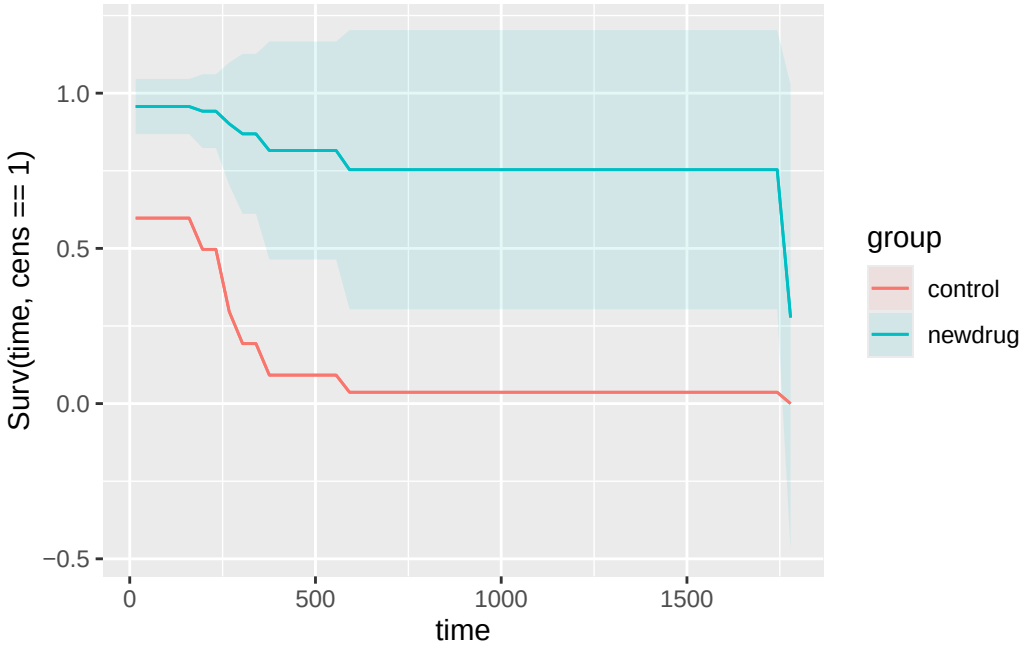


Figure 21: Predictions from lung cancer Cox model

```
lungcancer.1 %>% augment(lungcancer) %>%
  ggplot(aes(x = .fitted, y = .resid)) + geom_point()
```

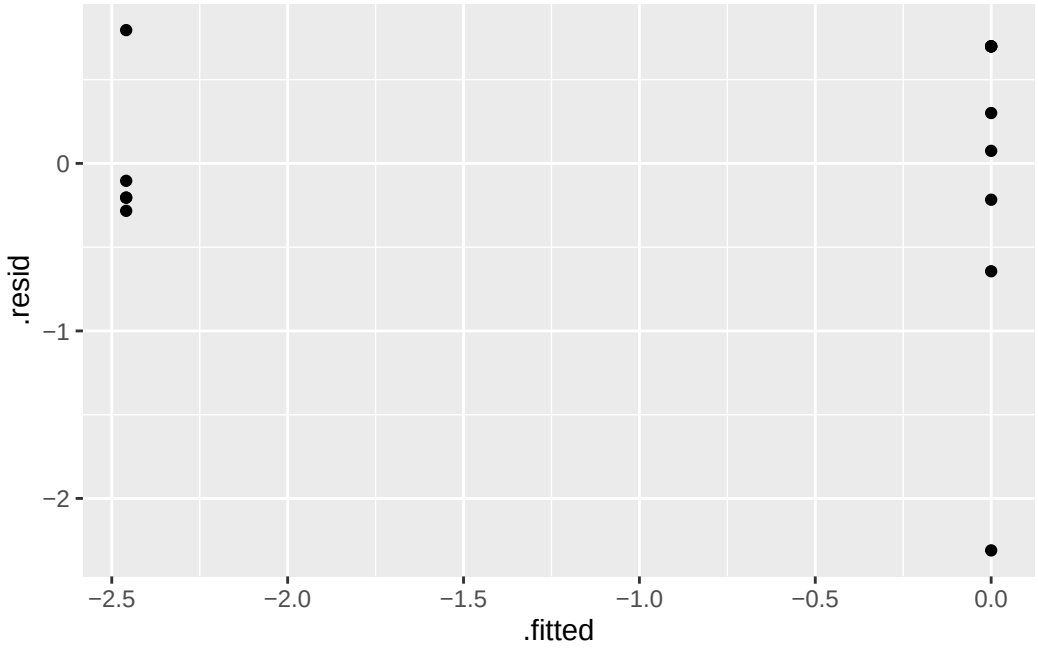


Figure 22: Another plot from lung cancer Cox model

group	attempts
Group2	7
Group2	9
Group2	11
Group2	11
Group2	13
Group1	3
Group3	14
Group1	4
Group1	7
Group2	10

Figure 23: Shock data (10 randomly chosen rows)

```
Call:
lm(formula = attempts ~ group, data = Shock)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4444 -1.4444  0.1111  1.1111  3.5556

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.4074     0.3889  26.762  < 2e-16 ***
groupc_any      -4.5185     0.5500  -8.216 1.96e-08 ***
groupc_med_sev  -2.2222     0.4763  -4.666 9.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.021 on 24 degrees of freedom
Multiple R-squared:  0.7881,    Adjusted R-squared:  0.7705
F-statistic: 44.63 on 2 and 24 DF,  p-value: 8.188e-09
```

Figure 24: Shock data analysis

treatment	rep	weight1	feed	weight2	gain
T3	R1	39	708	203	164
T2	R3	20	668	200	180
T1	R2	21	628	177	156
T2	R7	20	672	191	171
T4	R7	30	742	217	187
T5	R7	30	763	230	200
T2	R9	29	769	208	179
T3	R3	32	733	221	189
T1	R1	30	674	195	165
T3	R7	30	637	190	160

Figure 25: Pigs data (10 randomly chosen rows)

```
ggplot(crampton.pig, aes(x = feed, y = gain, colour = treatment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

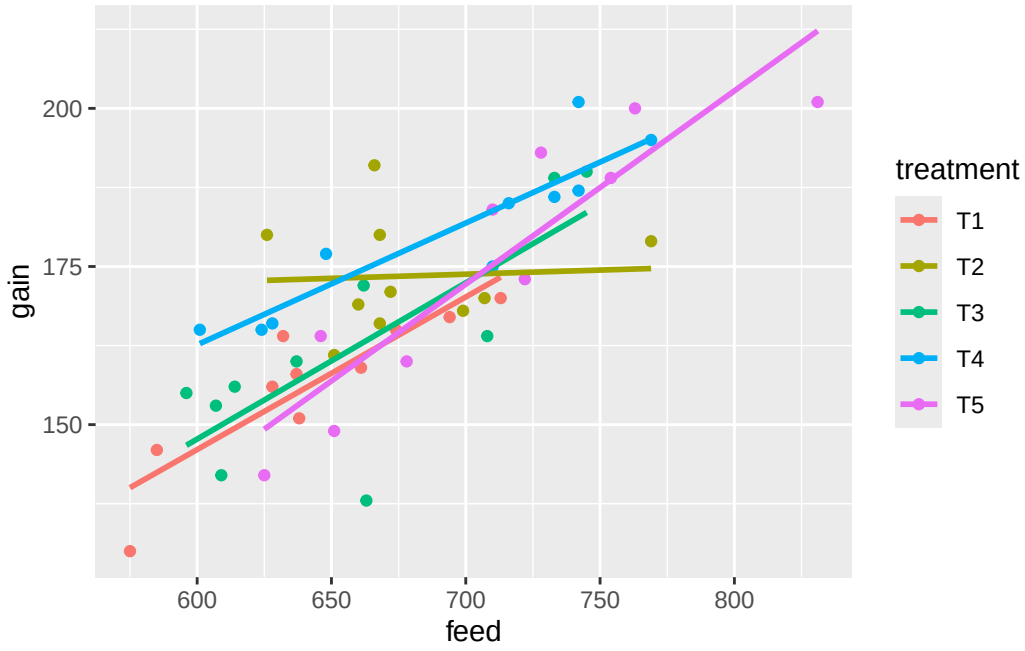


Figure 26: Graph of pigs data

```
pig.1 <- lm(gain ~ feed * treatment, data = crampton.pig)
drop1(pig.1, test = "F")
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	3024.615	225.1258	NA	NA
feed:treatment	4	899.1275	3923.742	230.1389	2.972701	0.030638

Figure 27: ANCOVA of pigs data

```r
summary(pig.1)
```

```
Call:
lm(formula = gain ~ feed * treatment, data = crampton.pig)

Residuals:
     Min       1Q   Median       3Q      Max
-25.2835  -5.1686  -0.0565   6.1270  17.6647

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.342383  42.798105   0.031 0.975134
feed               0.241196   0.066350   3.635 0.000784 ***
treatmentT2      163.288907  66.193398   2.467 0.018011 *
treatmentT3       -1.860747  55.241490  -0.034 0.973297
treatmentT4       45.776065  54.342512   0.842 0.404594
treatmentT5      -43.145509  53.945914  -0.800 0.428556
feed:treatmentT2  -0.228126   0.099615  -2.290 0.027368 *
feed:treatmentT3   0.005866   0.084898   0.069 0.945258
feed:treatmentT4  -0.048686   0.082056  -0.593 0.556298
feed:treatmentT5   0.064521   0.080759   0.799 0.429050
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.696 on 40 degrees of freedom
Multiple R-squared:  0.7857,    Adjusted R-squared:  0.7375
F-statistic:  16.3 on 9 and 40 DF,  p-value: 8.356e-11
```

Figure 28: Summary output from ANCOVA